Comparative Genomics of the Skin Staphylococci

Rosanna Coates-Brown

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

September 2015

Abstract

The human skin is a complex ecosystem which supports a diverse population of bacteria. Comparative genomic analyses are increasingly being used to explore the functional potential of this bacterial population . The ubiquity of *Staphylococcus* on human skin means this genus represents the most well-studied of the microbial skin residents, however most analysis has focussed on the significant clinical pathogenic species *S. epidermidis* and *S. aureus*.

To investigate the biology of *S. hominis*, the second most frequent *Staphylococcus* species isolated from human skin after *S. epidermidis*, seven isolates were sequenced using Illumina and PacBio technologies. An intraspecies comparative genomic analysis was performed with these and several publically available *S. hominis* genomes to identify core and accessory genes. The complement of encoded cell wall-anchored proteins was studied using bioinformatics to describe the range of surface-attached proteins and revealed a unique species set. Investigation also revealed the presence of *S. hominis* genes described as virulence factors in *S. aureus* and *S. epidermidis*. This further highlights non-pathogenic staphylococci as a reservoir of genes, which can be exchanged with pathogenic *S. aureus*, and the potential for recruitment of these genes into virulence pathways.

Interspecies comparative analysis of twenty *Staphylococcus* species, based on clusters of orthologous genes, confirmed the designation of staphylococcal species groups previously established by DNA-DNA hybridisation and single gene analysis methods. The bioinformatic algorithm randomForest was used to identify drivers forming species groups based on the orthologous gene cluster analysis leading to a subset of orthologous clusters defined as being contributory. This interspecies analysis also revealed diversity between the staphylococcal species groups with respect to their response mechanisms for antimicrobial peptide (AMP) resistance. Specifically, the presence or absence of the BraRS two-component system (TCS) was identified to be one of the important drivers differentiating a nine species member group that included *S. aureus*, *S. hominis* and *S. epidermidis*.. Experimental evolution in the presence of the lantibiotic nisin was used to dissect differences in the global response of the BraRS-positive species *S. hominis* and *S. aureus,* from the BraRS-negative species *S. saprophyticus*, . Identified SNPs from the resistance evolution revealed complex relationships between the regulons of staphylococcal TCSs and identified that YurK should be investigated for a potential role in AMP resistance of *S. aureus* and *S. hominis*.

Acknowledgements

First and foremost I would like to thank my primary supervisor, Dr Malcolm Horsburgh, for all of his help and support over the last four years. It's impossible to say in a couple of sentences how his open door policy, pep talks and sarcasm have given perspective to both the difficult moments and the high points along my PhD journey.

I would also like to thank my secondary supervisor Professor Neil Hall, my Unilever supervisory team Sally Grimshaw and the late Dr David Taylor, as well as my assessors Dr Heather Allison and Dr Alistair Derby, for their interest, advice and everything I have learned from them over my time at the University of Liverpool.

The members of Lab H past and present have provided a wealth of knowledge and an excellent sounding board for the "stupid questions". I have also learned so much from all of you- not least how to pull off the perfect prank!! Dr Jennifer Kelly and Dr Josephine Moran; you have both been there with the time to spare when I just didn't know where to start. Thanks to you both for the regular and frequent troubleshooting and debugging. Thanks also to Pisut Pongchaikul with all of his bioinformatics advice, and for patiently talking me through so many models and algorithms.

The Lunch group, there are too many people to name but you know who you are. All of your viva celebrations have reminded me that there is light at the end of the tunnel. Thanks for letting me share the good times. Of course a PhD has its share of bad times too, so thank you all for understanding the frustration of errant semi colons and contaminated universals.

Daniel, what can I say? We both know I would have lived on toast for the last six months without you. Thank you for being there for me and making it easy for me to reach for my goals. Finally I would like to thank the rest of my family. I feel like you have lived the trials and tribulations of my PhD with me. You can all breathe a sigh of relief- it's finished now!

Table of Contents

List of figures

List of tables

Chapter 1

General introduction

1.1 The human skin microbiome

The human skin is a rich and complex ecosystem supporting multiple communities of microorganisms, living in diverse physical and topographic niches. For example, despite the spatial proximity of areas such as the back and the base of the neck, the conditions of the two niches, and therefore the abundances of particular bacterial species which inhabit them, are distinct. The back is dominated by *Propionibacteriaceae* whereas staphylococcal species are most abundant at the base of the neck (Grice & Segre 2011). The human microbiome consists of a diverse mixture of bacteria which range from obligate aerobes to facultative anaerobes. Characterisation of the full diversity of the human microbiome has been greatly enhanced by the application of molecular-based approaches, since culture based methods reveal only those species which are cultivable in a laboratory environment (Fierer *et al.* 2008).

Better characterisation of the diversity of skin residents by molecular means has revealed multiple aspects of variation in skin microbial populations. The most obvious driver for variation is the previously mentioned differences in niche environment. Sebaceous niches, such as the forehead, represent the most phylotype poor niches, whereas dry environments support the greatest diversity of phylotypes (Costello *et al.* 2009).

The population of bacteria present in certain niches varies over time, with occluded sites such as the inner ear and nostril showing the most stability in terms of population and structure. Exposed sites with a greater diversity of inhabitants, such as the forearm and finger web spaces, were found to be the most volatile over time (Grice *et al.* 2009).

Finally, the microbiome populations between individuals varies greatly. This can depend on multiple host genetic factors and others such as disease and age. The skin microbiome of infants, for example, consists predominantly of staphylococcal species and diversifies throughout the early years of life (Capone *et al.* 2011). Different disease states can favour colonisation by particular bacterial populations. Atopic dermatitis favours colonisation by staphylococcal

species, in particular the non-resident species *S. aureus*. Following phases of intervention in the disease, however, populations of *Corynebacterium, Propionibacterium* and *Streptococcus* were increased (Kong *et al.* 2012). As well as variation between individuals of different age groups and disease states, there are also certain differences in the microbiomes of healthy adults. This variation is predominantly observed in exposed niches such as the hands, which share only 13% of phylotypes across individuals. In contrast, sebaceous niches such as the back and forehead exhibit the least variation (Grice & Segre 2011). At the species level, it is the low-abundance community members which represent the greatest variation between individuals. Strain variation is also present between healthy adults for some species such as *P. acnes,* whereas for strains of other species like *S. epidermidis* variation occurs predominantly on a site-specific basis, rather than between individuals (Oh *et al.* 2014).

The functional potential of the human microbiome is beginning to be understood through metagenomic sequencing efforts that elucidate the genetic complement of microbial communities. This has been achieved successfully for the gut microbiome and controversially one study revealed the potential for an obesity-associated microbiome to harvest increased amounts of energy from food (Turnbaugh *et al.* 2006). However a recent study by Walters *et al* found only a low correlation between the gut microbiota and clinical presentation of obesity (Walters *et al.* 2014). One of the issues that has hindered similar studies on the skin metagenome is the lack of reference genomes for the individual residents. Oh *et al* however exploited reference-free methodologies to investigate the functional potential of the skin microbiome which underlined variation both between individuals and between body sites (Oh *et al.* 2014).

Colonisation of the skin by resident skin flora inhibits invasion by pathogenic bacteria through competition for nutrients and the production of antimicrobial peptides (AMPs) and is therefore stated to fulfil a barrier function (Cogen *et al*, 2008; Segre *et al*, 2006). The importance of the human microbiome to both health and disease is increasingly being recognised. Despite this, there is an interesting dichotomy between efforts to protect and enhance the microbiota of

the gut through nutrition and probiotics, and efforts to reduce and remove the microbiota of the skin through hand and foot sanitising products and other personal hygiene cosmetics (Schloss 2014).

1.2 The skin staphylococci

The *Staphylococcus* genus comprises species of bacteria found in diverse niches from fermented meat products to healthy mammalian skin. As human skin residents the staphylococci are associated with a wide spectrum skin disorders such as impetigo, and the more serious staphylococcal scalded skin syndrome (Cho *et al*, 2010). Furthermore, members of the genus also have implications in an agricultural setting through their association with animal skin and are a frequent cause of bovine mastitis and infections in chickens. Species of staphylococci are used in the food industry as starter cultures for the fermentation of meat products (Dordet-Frisoni *et al*, 2007).

Due to their ubiquity on human skin, the staphylococci represent the most well-studied microbial skin residents. *S. epidermidis* is the most abundantly isolated species from human skin, particularly from the axilla and nares (Otto 2009). As previously mentioned skin resident microbes inhibit pathogenic bacteria, preventing them from colonising the skin and causing disease. *S. epidermidis* prevents *S. aureus* colonisation and subsequent biofilm formation in the nares. Furthermore, studies have shown that AMP production from keratinocytes is upregulated by *S. epidermidis* (Lai *et al.* 2010). Human skin isolates most commonly segregate into the clonal complex CC2. This is also true of the prolific ST2 isolate which is the most common isolate with a high potential for invasiveness, due to the presence of IS256 and the *ica* genes that facilitate biofilm formation in this sequence type (M. Li *et al.* 2009).

In addition to the well-characterised *S. epidermidis* several other species are abundant on human skin. Kloos *et al.* found *S. hominis* to be the most frequently isolated after *S. epidermidis*, closely followed by *S. haemolyticus.* Other species of coagulase negative staphylococci are also isolated less frequently from the skin including; *S. capitis, S. cohnii, S. saprophyticus, S. simulans, S. warneri and S.*

xylosus (Kloos 1980). Less is known about the role of these species in the microbiome. Although *S. aureus* is isolated from human skin it is not considered to be a healthy skin resident but a transient skin coloniser most likely present due to carriage in the anterior nares. This long held view of *S. aureus* as a non-resident of skin was recently challenged by the identification of particular clones of USA300 that were found to be residents on the skin of non-carriers (Schechter-Perkins *et al.* 2011). This is contradictory to previous evidence that showed decolonisation of the nares eradicates skin colonisation by *S. aureus.* For example Reagan *et al* found that nasal decolonisation with mupirocin resulted in decolonisation of the hand 72 hr after treatment in patients colonised by *S. aureus* (Reagan *et al.* 1991). It is to be noted that the study by Schechter-Perkins *et al* was an isolated incidence of sampling, and so no longitudinal data is available for extranasally-colonised patients (Schechter-Perkins *et al.* 2011). Further study is needed to determine the basis for this skin residence. Certain disease states such as atopic dermatitis increase the prevalence of *S. aureus* on the skin and increase the prevalence of anterior nares colonisation of people in the surrounding social community (Chiu *et al.* 2010).

1.3 The structure of the human epidermis

To better understand the skin as a bacterial niche it is necessary to first describe its structure and function, and the role the skin plays as the largest organ of the human body. The human epidermis primarily acts as a barrier preventing the entry of microbes into the sterile areas such as deep tissues, and prevents the egress of the large amounts of water required for the function of these tissues and organs.

A single hair follicle, sebaceous gland and the accompanying interfollicular epidermis comprises a pilo-sebaceous unit. The epidermis as a whole is made up of multiple repeated units which allow the skin to fulfil thermoregulatory and protective functions. Each unit also harbours a pool of stem cells integral to its protective function and to skin homeostasis (Segre 2006).

The epidermis itself is comprised of multiple layers (Figure 1.1). The cells in these layers increase in their differentiation as they move from the stratum basale as basal keratinocytes, through the stratum spinosum where the basal keratinocytes become spinous keratinocytes, to the stratum granulosum as granular cells. Finally in the stratum corneum the terminally differentiated cell become corneocytes (Blanpain & Fuchs 2009). Corneocytes are lost from the stratum corneum following the breakdown of the structures anchoring the cells to one another: corneodesmosomes. This process of proteolysis leads to desquamation. Following this loss of cells from the surface, the cells migrating up through the layers of the epidermis replace the lost corneocytes. This replacement requires multiple changes in cell morphology; firstly the rounded basal keratinocytes become polygonal and flat with a permeable plasma membrane; enucleation occurs; finally a cornified envelope develops and matures around the cell (R. H. Rice & Green 1977). Lipids such as ceramide are present at this stage of differentiation and are covalently attached to the cornified envelope via interactions with proteins (Candi *et al.* 2005).

Together with the cells of the epidermis, the extracellular matrix (ECM) is also vital to the skin as a niche. The ECM fulfils roles in wound healing, for example, and the molecules present at different skin sites such as the forehead and axilla, help to create the different niches which can be exploited by different species of bacteria. The composition of the terminal layer of the skin, the stratum corneum, is a stacked bilayer of skin lipids with cholesterol, free fatty acids and ceramides in a ratio of 1:1:1 (Iwai *et al.* 2012).

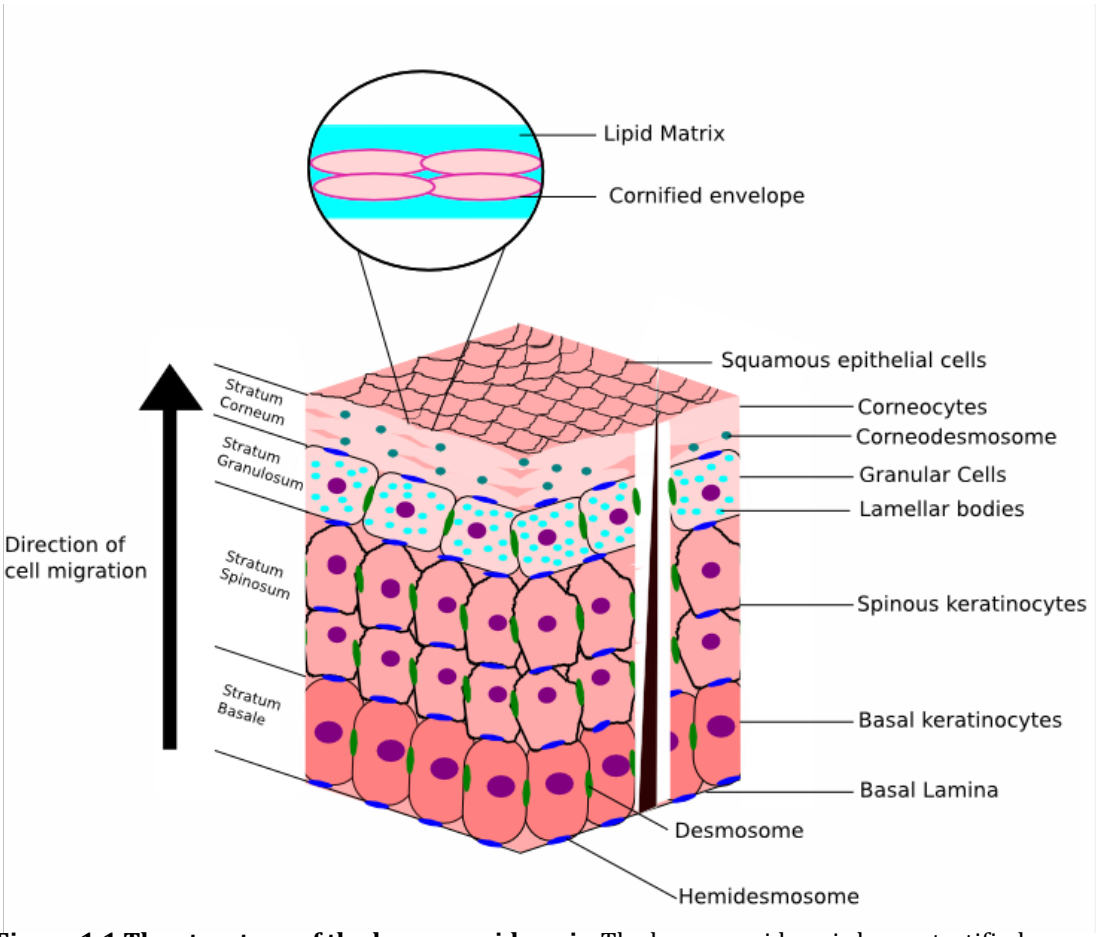


Figure 1.1 The structure of the human epidermis. The human epidermis has a stratified epithelial structure. As cells move up through the layers, as indicated by the arrow, they move toward terminal differentiation. The corneocytes of the terminal skin layer are surrounded by an extracellular matix of lipids.

1.4 The skin as a bacterial niche

As a barrier between the outside world and our body systems, the skin represents a harsh environment for bacteria. The very structure of the skin with its stratified composition and turnover of terminally differentiated corneocytes obstructs the initial attachment of bacteria to the skin surface. In addition to this fundamental barrier, the skin produces an arsenal of proteins, lipids and other antimicrobial molecules to prevent bacteria from gaining a foothold and subsequently thriving on its surface. Due to the presence of metabolites, however, the skin represents a source of nutrients for the growth and reproduction of bacterial species able to survive. This section will discuss the obstacles to skin survival for the resident skin microbiome and the mechanisms by which colonising bacteria overcome them (Figure 1.2).

1.4.1 Adhesins

Adhesins mediate attachment of bacterial cells to the host cells anchoring them in their niche. Staphylococci have a varied repertoire of adhesins which fall into the categories of microbial surface components recognising adhesive matrix molecules (MSCRAMMs), secretable expanded repertoire adhesive molecules (SERAMs) and non-proteinaceous adhesins. The different adhesins exhibit specificity for discrete host molecules and therefore facilitate colonisation of the variant niches of the body. This is discussed in further detail in chapter 3 (section 3.1.4).

In addition to niche specificity, adhesins and host receptor molecules may also differ between the initial attachment stages and long-term colonisation. In the nares, for example, *S. aureus* initial attachment to the epithelia is facilitated by wall teichoic acids (Weidenmaier *et al.* 2004) which have been shown to interact with fibronectin *in vitro* (Hussain *et al.* 2001). ClfB, on the other hand is implicated in long term nasal colonisation since *S. aureus* strains lacking the adhesin were cleared from a mouse nasal colonisation model more quickly than the wild type (Schaffer *et al.* 2006) (Weidenmaier *et al.* 2008).

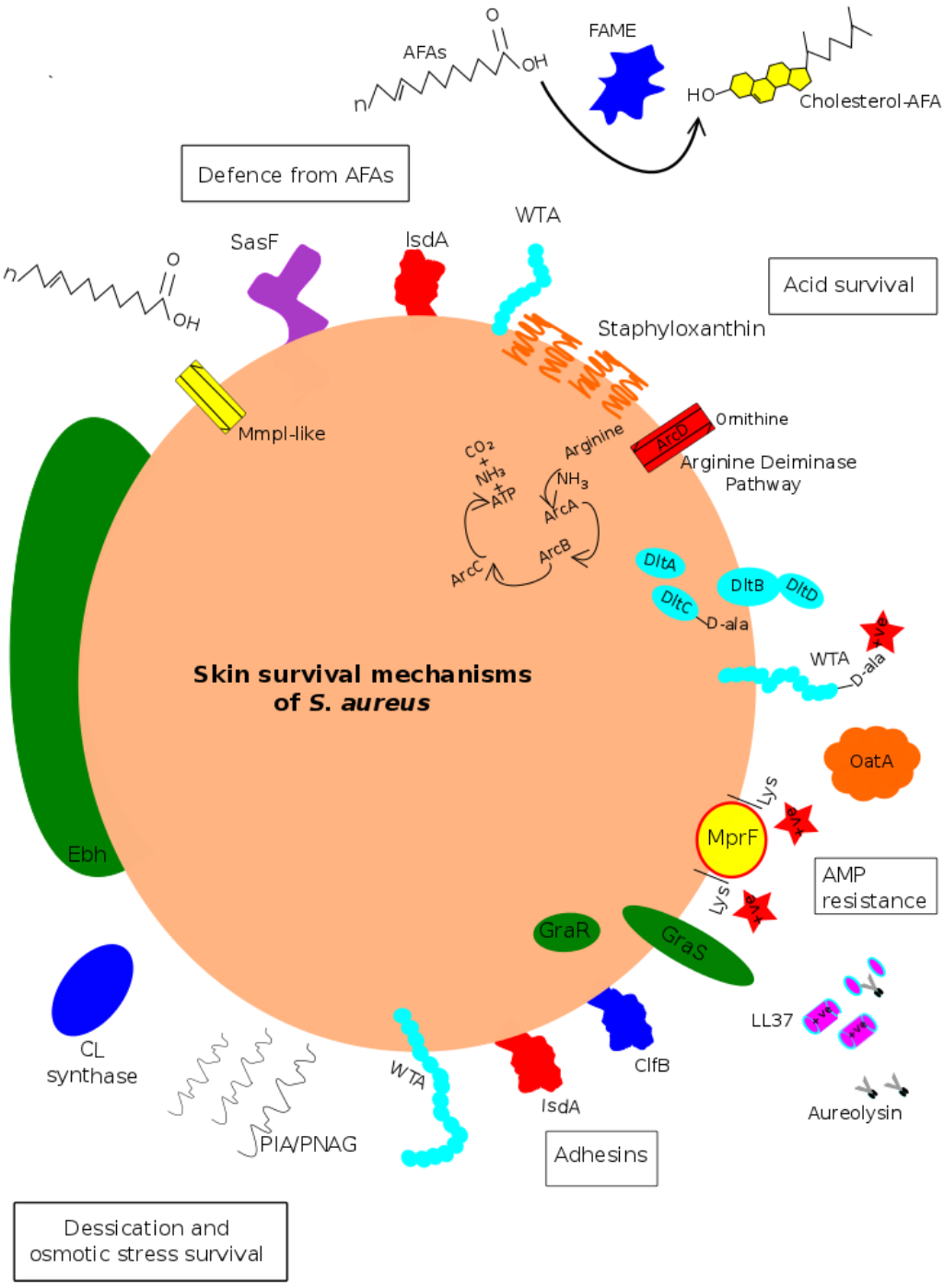


Figure 1.2 The skin survival mechanisms of *S. aureus*. The conditions on the skin such as dessication, the acid mantle and mechanical shear forces from desquamation, and the antimicrobial lipids and peptides produced, represent obstacles to survival on the skin. *S. aureus,* in turn, has evolved multiple defence mechanisms to combat these obstacles.

1.4.2 The acid mantle and acid tolerance

The acidic pH at the surface of the skin constitutes part of the innate immune system preventing colonisation of the skin. The epidermis becomes more acidic from the stratum granulosum which has a pH of 7.4, to the stratum corneum, which has a pH between 4 and 5. This acidic pH at the surface of the skin is due to lactate and fatty acids and is essential to the synthesis of ceramides which in turn are integral to barrier homeostasis (Feingold 2007). The acidic pH is also inhibitory to serine proteases able to compromise the barrier function by breaking down corneodesmosomes (Hachem *et al.* 2005). As a result of these regulatory functions, alteration of the skin's acidic gradient is implicated in skin disorders such as atopic dermatitis where the process of desquamation is disrupted and increased pH leads to elevated serine protease activity compromising the integrity of the epidermis (Cork *et al.* 2009). The acidity of the skin surface also enhances the antimicrobial properties of antimicrobial fatty acids (AFAs) (Cartron *et al.* 2014) and antimicrobial peptides (AMPs) (Walkenhorst *et al.* 2013), which will be discussed in later sections (1.4.3, 1.4.5). Finally, the acid pH has direct antimicrobial properties through its ability to denature DNA, disrupt bacterial membranes and damage proteins (Cotter & Hill 2003).

In acidic conditions, proton pumps serve to export protons from the intracellular environment by generating a proton motive force and thereby increasing the pH inside the bacterial cell. In *S. aureus* NADH dehydrogenase is upregulated in response to acid stress and is able to generate a proton motive force through converting NADH to NAD^+ (Bore *et al.* 2007). As well as working to increase their intracellular pH, skin staphylococci are also able to affect the pH of their local environment through systems such as the arginine deiminase pathway (Cotter & Hill 2003). Bore *et al* however did not find the genes of the arginine deiminase pathway to be upregulated in response to acid stress (Bore *et al.* 2007). The urease genes were however strongly upregulated and their function is known to result in the production of NH_3, which buffers the intracellular pH (Bore *et al.* 2007). Changes in gene regulation are also observed in response to acid tolerance. Expression of global regulators such as RsbU,

PhoB and the Fur-like regulator SACOL1541 is altered (Bore *et al.* 2007). Equivalent regulators are part of the acid stress response of *Salmonella enterica,* and gene mutations result in acid sensitivity (J. W. Foster 2004).

In addition to these direct mechanisms of acid tolerance, a range of indirect responses to acid stress also occur in *S. aureus.* The presence of low pH increases toxicity of metal ions so during acid stress metal ion export is upregulated. Conversely exporters of alkaline compounds such as spermidine and putrescine are downregulated. It is speculated that keeping these compounds inside the cell may act to raise intracellular pH (Bore *et al.* 2007). Systems important to DNA repair and protein degradation are also upregulated; DNA-damage-inducible protein P (*dinP*), DNA polymerase A (*polA*) and the exonucleases *rexA* and *rexB* are all upregulated (Bore *et al.* 2007). RexAB is necessary to repair double stranded breaks (Quiberoni *et al.* 2001). Finally, mechanisms associated with growth and repair are downregulated, which agrees with altered expression of ribosomes under stress conditions (Bollenbach *et al.* 2009). The duplication of the gene encoding cardiolipin is implicated in acid tolerance as one cardiolipin gene locus is not expressed under conditions of acid stress, whereas the other locus is (Ohniwa *et al.* 2013).

1.4.3 Antimicrobial fatty acids

In addition to contributing to the acid mantle of the human epidermis, antimicrobial fatty acids (AFAs) have other bactericidal effects. These AFAs interact with the staphylococcal membrane due to their lipophilic properties. Within the lipid bilayer the structure of the AFAs reduces the so-called packing density of the lipids in the bilayer thus reducing its stability, or in other words increasing membrane fluidity. In turn this results in the loss of proteins from the cell and disruption to metabolic processes such as the electron transport chain (Galbraith & Miller 1973) (Greenway & Dyke 1979) (Desbois & V. J. Smith 2009).

To combat the increased membrane fluidity caused by integration of AFAs into the lipid bilayer *S. aureus* synthesises the carotenoid pigment staphyloxanthin

from the *crt* operon (N. N. Mishra *et al.* 2011). The carotenoid operon is discussed in detail in chapter 3 (section 3.1.4.5). FAME, or fatty-acid modifying enzyme, activity inhibits the antimicrobial effects of AFAs by esterification with primary alcohols and cholesterol. FAME is present in a large proportion, but not all, strains of *S. epidermidis* found on the skin. This indicates that it is advantageous but not essential to skin survival. In addition to its presence in *S. epidermidis* FAME is also encoded and functional in *S. aureus, S. cohnii, S. schleiferi, S. saprophyticus, S. warneri, S. caprae, S. simulans, S. hominis* and *S. capitis*. FAME is absent from *S. lugdunensis* and *S. haemolyticus* (Long *et al.* 1992). Reduction of hydrophobicity is proposed as an important survival mechanism, for example via the MSCRAMM IsdA, which reduces the capacity of AFAs to interact with the cell surface (Clarke & S. J. Foster 2006). Upregulation of capsule and peptidoglycan genes may similarly result in reduced hydrophobicity (Kenny *et al.* 2009). Additional MSCRAMMs involved in defence against AFAs include SasF of *S. aureus* and the *S. saprophyticus* homologue SssF. SasF is upregulated in *S. aureus* in response to AFAs (Kenny *et al.* 2009), and clinical strains of *S. saprophyticus* expressing SssF have an increased resistance to AFAs compared to those which do not (King *et al.* 2012).

1.4.4 Osmotic stress and osmotic stress tolerance

The human epidermis has low relative humidity (water activity) and high salinity; these features lead to osmotic stress for the bacterial inhabitants. Both environmental stresses are counteracted by the skin resident *S. epidermidis* which has a requirement for relative humidity between 81-84% for growth. The transient species *S. aureus* requires a higher relative humidity of ~87% which is lower than environmental species such as *Pseudomonas fluorescens* which require a still higher level of 92-94.5% (de Goffau *et al.* 2009).

The staphylococci are characteristically capable of growth in high salinity. Mechanisms to overcome osmotic stress include morphological changes which occur if the relative humidity falls below the level tolerated. Both *S. aureus* and *S. epidermidis* increase their cell size and cell wall thickness, clustering in cuboidal groups of 8 cells in contrast to their usual “bunch of grapes” formation.

These changes probably serve to reduce the surface area to volume ratio to limit water loss and maintain the turgor pressure of the cell. As these morphological changes are also observed in autolysin mutants, these cell wall hydrolytic enzymes might be involved in bringing them about (de Goffau *et al.* 2011). Ebh is a further cell wall-located protein implicated in osmotic stress tolerance. Mutants of this giant protein result in invaginations along the septum of cells under conditions of high salinity (Kuroda *et al.* 2008). PIA, or polysaccharide intercellular adhesin, enables producer strains to form a biofilm. PIA producer strains of *S. epidermidis* have been found to have a high tolerance to low humidity conditions (de Goffau *et al.* 2009). PIA may act similarly to the exopolysaccharide (EPS) of *Pseudomonas aeruginosa* which slows the rate of both water loss during osmotic stress, and the rate of subsequent rehydration, to allow time for adaptation of the bacterial cell to the changing conditions (Roberson *et al.* 1992). Synthesis of cardiolipin in staphylococci has several roles during osmotic stress; it reduces membrane fluidity (Tsai *et al.* 2011) and reduces the rate of water loss from the cell by regulating solute efflux channels such as the mechanosensitive channel MscL (Romantsov *et al.* 2009).

1.4.5 Antimicrobial peptides of the epidermis and staphylococcal defence mechanisms

In addition to the acid mantle, the host's innate immune system also consists of antimicrobial peptides (AMPs) which function to prevent colonisation by bacteria. The epithelia produce two major classes of cationic AMPs: defensins and cathelicidins. A calcium gradient exists across the layers of the skin and increases towards the upper layers. The increased calcium concentration in turn leads to increased expression of cathelicidin processing enzymes; the serine proteases called kallikreins in the upper layers of the skin. Cathelicidins and defensins are both most abundant in the upper layers of the skin (Morizane *et al.* 2010). Human β-defensins (HBDs) 1 and 2, as well as cathelicidins, are secreted in the high salinity, low pH environment of the sweat gland. Cathelicin LL-37 is physiologically active at low concentrations here, however the activity of HBDs is uncertain (Murakami *et al.* 2002) (Rieg *et al.* 2005). HBDs and cathelicidins bind to the negatively charged surface of the staphylococcal cell

membrane through electrostatic interaction. Subsequent accumulated AMPs result in the formation of channels in the cell membrane which eventually effect increased permeability (Ganz 2003).

The staphylococci have evolved multiple signal transduction system to sense AMPs and regulate a downstream response to counteract their bactericidal activity. In *S. epidermidis* the two component system ApsS and the homologous system in *S. aureus*, GraRS, sense the presence of cationic AMPs (cAMP). This interaction results in signal transduction from the sensor protein to a regulator to control changes in expression of their cognate regulon of genes involved in resistance to antimicrobial peptides. The GraRS regulon of genes includes the *dlt* operon and *mprF*. Other TCS involved in resistance to AMPs and other antimicrobials of the skin include the GraRS homologue BraRS; these TCSs are discussed in detail in chapter 4 (section 4.1.4)

Susceptibility to cAMP is counteracted by the *dlt* operon, the gene products of which add D-alanine to wall teichoic acids. This reduces the negative charge of the staphylococcal cell surface which in turn reduces the electrostatic interaction between the cAMP and the cell (Peschel *et al.* 1999). MprF also causes a reduced surface negative charge through the addition of L-alanine and L-lysine to phospholipids of the bilayer as well as translocation of lysyl-phosphatidylglycerol to the outer cell membrane (Ernst & Peschel 2011).

Secreted proteases also play a role in increased persistence in the presence of AMPs. The *S. aureus* protease aureolysin, for example, degrades LL-37, as well as complement C3, therefore attenuating the immune response and allowing *S. aureus* to evade phagocytosis. LL-37 removal increases *S. aureus* persistence on the skin surface (Laarman *et al.* 2011).

Dermcidin is an anionic AMP and, due to its like charge with the staphylococcal membrane, its mode of action is differs from that of the cAMPs. The anionic AMP functions instead by inhibiting protein and RNA synthesis (Rieg *et al.* 2005) (Senyurek *et al.* 2009).

Proteases are produced by staphylococci as a defence mechanism against multiple AMPs. Production of SepA, a metalloprotease, is upregulated in the presence of dermcidin and is under the control of *sarA, agr* and *saeRS* (Lai *et al.* 2006). In addition to their role in resistance to cationic and anionic AMPs, secreted proteases also function to promote skin colonisation and destabilise the skin's barrier function. ET and V8 proteases break down desmoglein-1 in the stratum granulosum and therefore weaken cell-to-cell interactions (Hirasawa *et al.* 2009).

Lysozyme, an *N*-acetylmuramoylhydrolase, is produced by the epithelia in addition to many other tissues and secretions. Lysozyme interacts with cell wall peptidoglycans of Gram-positive bacteria and leads to cell lysis (Bera *et al.* 2004). Lysozyme resistance is mediated by OatA which O-acetylates cell wall peptidoglycans and inhibits interaction by lysozyme. Furthermore, crosslinked wall teichoic acids provide further steric hindrance to lysozyme interaction (Bera *et al.* 2006).

1.4.6 The antimicrobial RNase A superfamily

RNase 5 and 7 are both expressed by keratinocytes and have multiple roles, including ribonuclease activity. RNase 5 for example is involved in angiogenesis as a secondary function; it is also antimicrobial, and was shown to be active against *Candida albicans* and *Streptococcus pneumoniae* in similar physiological conditions to other AMPs (Hooper *et al.* 2003). Although the results of the study by Hooper et al could not be reproduced (Avdeeva *et al.* 2006), Abtin *et al* did confirm the antimicrobial activity of RNase 5 (Abtin *et al.* 2009).

The cationic RNase 7 exhibits broad spectrum antimicrobial activity and its expression in epithelial tissues is associated with *S. aureus* inhibition; anti-RNase 7 antibody-treated skin explants exhibited reduced killing of *S. aureus* (Simanski *et al.* 2010). Although the mechanism of antimicrobial action of RNase 7 remains unclear, Huang *et al* suggest that the cationic property of RNase 7

influences its ability to interact with *S. aureus* and bring about permeabilisation of the cell membrane (Y.-C. Huang *et al.* 2007).

1.4.7 Competition on the skin

As previously discussed, the staphylococci which inhabit the skin live as part of complex communities of microbes in the different niches of the body. The individuals in these communities must compete with one another for resources, such as nutrients and space.

There is evidence that there is competitive exclusion within skin niches; *S. aureus and S. epidermidis,* for example, are rarely isolated together from nasal epithelia. Exclusion mechanisms between *S. epidermidis* and *S. aureus* include stimulation of the host innate immune system. As a successful skin coloniser, *S. epidermidis* is resistant to the host antimicrobial peptides HBD-2 and HBD-3, whereas *S. aureus* is not. *S. epidermidis* produces the serine protease Esp which stimulates the production of these AMPs by keratinocytes and these AMPs in turn inhibit the colonisation of the epithelia by *S. aureus* (Iwase *et al.* 2010).

In addition to inhibiting *S. aureus* colonisation by secreting proteases, production of the phenol-soluble modulins (PSM)-γ and –δ by *S. epidermidis* also inhibits other species found on the skin. These two PSMs work in synergy with one another and the host immune system. PSM-γ –δ both work in concert with LL-37, and PSM-γ also works co-operatively with HBD-2 and HBD-3 (Cogen *et al.* 2010). Although *S. aureus* also produces PSMs these molecules have a reduced antimicrobial activity in comparison with *S. epidermidis* PSMs. Instead they are involved in biofilm production and haemolysis (Periasamy *et al.* 2012).

Several staphylococci are not often co-isolated with certain corynebacteria or *Propionibacterium acnes* (Frank *et al.* 2010) (Libberton *et al.* 2014). No antimicrobial production by *Corynebacterium* species was observed to explain the exclusion of staphylococci, however corynebacteria are proposed to bind more strongly to the nasal epithelia than the species of *Staphylococcus* which could explain their exclusion (Uehara *et al.* 2000). *S. epidermidis* is able to

convert glycerol to succinic acid which inhibits *P. acnes.* This may explain why *P. acnes* and *S. epidermidis* are not often co-isolated; succinic acid also ameliorated lesions in a *P. acnes* mouse model (Y. Wang *et al.* 2014).

The production of AMPs by skin commensals is also known to be a means of out-competing the competition. Many species of staphylococci, for example produce lantibiotics active against the members of their niche community. Staphylococcal lantibiotics will be discussed in detail in chapter 5 (section 5.1).

1.4.8 Staphylococcal disease

1.4.8.1 Atopic dermatitis

Atopic dermatitis is an inflammatory disease of the skin. Its symptoms include flaky lesions of the skin, infection and subsequent abscess formation. The disease has a high prevalence in a reported 97 countries across the world indicating that it is a significant problem in both developing and developed countries. Atopic dermatitis is associated with an increased production of the IgE antibody in response to food or environmental antigens usually otherwise well-tolerated. This is known as atopy, and can result in allergic disorders where sufferers may develop allergies and asthma, in addition to the atopic dermatitis, early in life (Zheng *et al.* 2011).

On atopic skin, dysregulation of the innate immune system results in reduced production of antimicrobial peptides, reduced recruitment of neutrophils to the skin and epidermal barrier dysfunction. These factors affect the skin microbiome. Patients with AD have high levels of *S. aureus* colonisation: up to 80% of the skin microflora may be comprised of *S. aureus*, with an associated reduction in overall microbiome diversity, in contrast with healthy skin where *S. epidermidis* comprises 90% of the resident microbiota (Baviera *et al.* 2014). In turn, the severity of atopic dermatitis is linked to *S. aureus* abundance and the concomitant loss of microbiome diversity (Kong *et al.* 2012).

A study by Nakamura *et al.* found that the *S. aureus* isolates present on atopic skin produce high levels of δ-toxin. These isolates promoted the production of the inflammatory cytokines IL-4 and IgE. These two inflammatory cytokines are known to play a role in the release of antimicrobial compounds from mast cells through degranulation. This increased mast cell degranulation by δ-toxin in *S. aureus* was shown through knock out and subsequent complementation of δ-toxin. Furthermore it was suggested that inhibition of δ-toxin in *S. aureus* may be a worthwhile treatment in atopic dermatitis as results from mouse models indicate that the presence of δ-toxin induces a Th2 immune response(Nakamura *et al.* 2013). The Th2 immune response is associated with allergic inflammatory responces which would induce inflammation of the skin (Paul & Zhu 2010).

In addition to the high prevalence of *S. aureus* on atopic skin, production of AMPs, such as hBD-2, was also proposed to be supressed in AD. Several studies by the same group of researchers suggested that a Th2 cytokine response may supress the production of AMPs in AD. This was shown by antibody-mediated neutralisation of Th2 cytokines which led to increased expression of hBD-3 (Howell *et al.* 2006). The AMPs hBD-2 and hBD-3 were also inhibited by the Th2 cytokines IL-4 and IL-13 (Nomura *et al.* 2003). It is possible, however, that inhibition of hBD-2 by corticosteroids routinely used in the treatment of AD results in this apparent lower level of hBD-2 on atopic skin. This outcome is supported by a study that showed lower levels of hBD-2 in AD lesions on skin treated with steroids compared with skin where the intervention was not used (Jensen *et al.* 2011). A separate study also found that an increased time interval since the last steroid treatment resulted in increased levels of hBD-2 in lesional and non-lesional AD skin (Clausen *et al.* 2013).

The levels of hBD-2 are higher in lesional AD skin than in control groups of healthy skin (Clausen *et al.* 2013). Expression of hBD-2 was also found to be elevated in AD by Harder *et al*, in addition to the increased expression of hBD-3 and RNase7, compared with healthy skin. This study did not, however, correlate AMP expression with levels of *S. aureus* colonisation (Harder *et al.* 2010).

Gambichler *et al* also found increased expression of hBD-3 and RNase7 in AD patients and so hypothesise that the increased colonisation by *S. aureus* associated with the disease does not result from reduced expression of AMPs. The data regarding AMP expression in atopic skin has been extensively reviewed recently (Kopfnagel *et al.* 2013).

The expression of AMPs also influences the immune response in AD. For example, hBDs are known to recruit cells of the innate and adaptive immune response, such as macrophages and neutrophils, by binding to CCR2 which is expressed on their surface (Röhrl *et al.* 2010). Furthermore, hBD-2 acts as a chemoattractant for T-helper cells and Th17 cells since it binds to CCR6, which is expressed by these cell types (Yang *et al.* 1999). Increased levels of hBDs on atopic skin may also result in increased levels of these immune cells.

Skin inflammation may be enhanced by hBD-2 as this AMP can stimulate the production of IFN-gamma by T-cells, thus favouring a Th1 cytokine response, which is associated with inflammation. Under some experimental conditions hBD-2 also enhances IL-22 and oncostatin M production and these cytokines in turn induce the production of hBD-2. Thus a positive feed-back loop may occur on atopic skin resulting in the increased production of hBD-2 that is observed. LL-37 also upregulates the cytokines IL-22 and oncostatin M (Kanda *et al.* 2011).

1.4.8.2 Abscess formation

One of the most common presentations of staphylococcal disease is abscess formation and an associated bacterial persistence. These abscess lesions are developed over a period of weeks and the disease forming bacteria are located at the centre of a lesion, separated by a pseudomembrane from the immune cells which infiltrate the abscess. The process of staphylococcal abscess formation after subcutaneous invasion is separated into four stages, each one categorised by a set of surface proteins and virulence factors (Cheng *et al.* 2011).

The first stage of subcutaneous abscess formation is the survival of *S. aureus* in the bloodstream. Upon inoculation into the bloodstream of a mouse model of infection ~ 99.9% of the staphylococcal inoculum disappears from the blood within around 6 h. Within 3 h, however, the bacteria will have reached the deep tissues where abscess formation will be established; replication of these seed populations is measurable within 24 hr (Cheng *et al.* 2009). Neutrophilic leukocytes are the hosts main defence against *S. aureus* circulating in the blood stream (Voyich *et al.* 2005). *S. aureus* in turn produces the antioxidant staphyloxanthin which acts to eliminate the toxic reactive oxygen species produced by the neutrophils (Pelz *et al.* 2005). *S. aureus* is able to survive within these neutrophils, and these infected immune cells are able to seed infections in naïve mice (Gresham *et al.* 2000). Production of phenol soluble modulins, secreted peptides which disrupt neutrophil membrane integrity, also increases the persistence of *S. aureus* in the blood stream (R. Wang *et al.* 2007). ClfA binds to fibrinogen and fibrin which, in the context of a staphylococcal blood stream infection, allows agglutination of *S. aureus* cells (McDevitt *et al.* 1994). This agglutination may act to prolong survival by inhibiting phagocytosis (Palmqvist *et al.* 2004). ClfA may also contribute to prolonged survival by allowing attachment of *S. aureus* to the epithelia of the blood vessel through interaction with fibrin (Moreillon *et al.* 1995) (Cheng *et al.* 2011). Finally, adenosine synthase A supresses the host inflammatory response by producing adenosine from AMP; adenosine is used by the host to signal the end of an inflammatory response (Thiel *et al.* 2003).

Establishing an infectious lesion is the next stage of staphylococcal abscess formation. Lesions are formed when immune cells such as neutrophils accumulate in a location at high enough frequency to replace normal epithelial tissue at this location. This is discernable ~48 hr post intravenous challenge (Cheng *et al.* 2009). The recruitment of large numbers of immune cells to lesions and small numbers of staphylococcal cells during the early stages of abscess formation are contradictory. It is possible that *S. aureus* may release lipoproteins which are recognised by Toll-like receptors, this would cause subsequent pro-inflammatory signals and lead to recruitment of immune cells

(Bubeck Wardenburg *et al.* 2006). The balance between fulfilling the metabolic need for haem, and reducing the haem-toxicity of the lesional environment requires the expression of the *isd* genes, which are involved in haem uptake, as well as the production of the HrtAB ABC transporter which exports excess haem from the cell (Mazmanian *et al.* 2003) (Torres *et al.* 2007).

The abscess formation progresses after several days to stage three which is characterised by a community of *S. aureus* replicating within a pseudocapsule of fibrin deposits (Cheng *et al.* 2009). These fibrin deposits are produced from the cleavage of fibrinogen by coagulase-prothrombin complexes (Cheng *et al.* 2011). This fibrin barrier prevents the immune cells from reaching the replicating *S. aureus* cells at the centre of the abscess (Viana *et al.* 2010). At this stage of abscess formation staphylococcal protein A (SpA) is also integral and prevents phagocytosis (Forsgren & Sjöquist 1966). Embp is also important during abscess maturation; the protein is expressed on the surface of the cells growing within the fibrin pseudomembrane suggesting it may promote biofilm-like growth (Cheng *et al.* 2009).

For persistence of the abscess expression of extracellular adhesion proteins (*eap*) and Ess, via a non-canonical protein secretion pathway, are both important. The *Mycobacterium tuberculosis* homologue of Ess is involved in aspects of immune evasion such as inhibiting phagocytosis (Davis & Ramakrishnan 2009). *S. aureus* mutants at this locus are able to establish an abscess but the abscess does not persist (Burts *et al.* 2008). Eap has both pro-inflammatory activity through induction of cytokines IL-6 and IFN-α from monocytes, as well as anti-inflammatory activity, by inhibiting T-cell activation and adherence of neutrophils to endothelial cells (Scriba *et al.* 2008) (Haggar *et al.* 2004).

1.5 Thesis aims

At the outset of this study there was an incomplete set of available genomes representing the major staphylococcal species that colonise humans. The sequencing of those species' genomes that were not available, or that were yet

to be sequenced, would be undertaken to enable the comparative study of the major human-associated staphylococcal species. The genomic study of the staphylococci has focussed mostly on the two clinically important species, *S. aureus* and *S. epidermidis*. This emphasis has limited the comparative study of the main species that colonise humans. To enlarge our knowledge of *S. hominis*, the second most frequent species of the genus colonising humans, a set of skin isolates would be sequenced and compared with the aforementioned species. A range of bioinformatics tools would then be used to identify components that were differentially present in staphylococcal species, including *S. hominis*, to gain new insights of colonisation and survival on human skin

Chapter 2

The sequencing, assembly and annotation of staphylococcal genomes

2.1 Introduction

Over the course of the last 40 years whole genome sequencing has exploded, starting from the sequencing of the first bacteriophage genomes in the late 1970s (Koonin & Galperin 2013). The shotgun sequencing of the first bacterial genome, *Haemophilus influenzae* in 1995, took 13 months to sequence and assemble into 140 contigs at a cost of around $900 000 (Fleischmann *et al.* 1995). Afterwards the draft human genome sequence was published in 2001 taking a large team around 10 years work (Lander *et al.* 2001). Today we find ourselves at the point where bench-top sequencing machines, such as the Illumina MiSeq can output a bacterial genome in a matter of days at a cost of around $100. The near future may hold further improvements as nanopore technology matures. Nanopore technology directly sequences long, native DNA fragments without amplification over a short time-frame. This sequencing uses a device the size of an average smart phone and the technique has the potential to yield whole bacterial chromosomes assembled in a single contig (Loman *et al.* 2015).

Here, the significant sequencing and assembly methods spanning the lifetime of whole genome sequencing are briefly reviewed.

2.1.2 Sequencing technologies

2.1.2.1 Sanger sequencing

The conception of whole genome sequencing was born from the Sanger sequencing method. This method was developed by Freidrich Sanger in 1977, and has since been automated using capillary electrophoresis and optical reading methods rather than manual sequencing (Sanger *et al.* 1977) (VARALDO & Kilpper-Bälz 1988). Today, the method is based on a single-stranded DNA template in a sequencing reaction which contains four dideoxynucleotides labelled with different fluorescent tags. Deoxynucleotides are present in excess to allow coverage of the whole DNA template which is amplified in a polymerase chain reaction (PCR) whereby the incorporation of a dideoxynucleotide in the place of a deoxynucleotide terminates the extension

step. The resulting fragments after the rounds of extension are then separated by capillary electrophoresis, and the fluorescent tag from each fragment is detected thus allowing the sequence to be read (L. M. Smith *et al.* 1986) (Prober *et al.* 1987).

Although Sanger sequencing is still considered to be the gold standard for DNA sequencing, other sequencing chemistries and platforms have been developed with the aim of making DNA sequencing faster and cheaper without sacrificing accuracy.

2.1.2.2 454 Pyrosequencing

454 Life Science released the first commercially available genome sequencer to the market in 2005 and were soon after acquired by Roche Diagnostics in 2007. Sequencing involves fragmentation of the DNA template and ligation of an adapter sequence to each end of these short fragments to act as both PCR and sequencing primers. The fragments are then attached to streptavidin-coated beads via a biotin tag at the 5' end of one of the adaptors. Only one fragment is intended to anneal to one bead, and these bead bound DNA fragments constitute the DNA sequencing library (Margulies *et al.* 2005). This method of library preparation was considerably quicker and easier than existing library preparation. Whole genome sequencing using the Sanger sequencing method instead involved shotgun sequencing of fragments of DNA cloned into BAC or PAC large-fragment cloning vectors (Lander *et al.* 2001).

Following library preparation the bead bound fragments are used in emulsion PCR, which sees the amplification of bead bound fragments within micro-reactors of the PCR mixture in an emulsion oil. The amplified fragments are immobilised on the beads. Beads without bound DNA are then filtered. The subsequent sequencing reaction proceeded on a so-called sequencing by synthesis basis. In this reaction, single beads are loaded into the wells of a Pico TiterPlate and the previously amplified fragments are exposed to DNA bases sequentially washed over the plates. These sequential washes occur 100 times during sequencing and the incorporation of a complimentary nucleotide results

in the emission of a recorded light signal. Signal strength relates to the number of complimentary bases incorporated (Margulies *et al.* 2005).

One of the advantages to 454 sequencing lies in the long read length produced; the 454 GS FLX+ system could output 1,000,000 reads of around 1000 bp in a run. The platform has a relatively fast throughput time compared to whole genome sequencing using Sanger methodology. However the major drawbacks of the system include the high cost of reagents and the high error rate surrounding runs of the same base longer than 6bp (L. Liu *et al.* 2012) (El-Metwally 2014).

2.1.2.3 ABI SOLiD

Sequencing by Oligonucleotide Ligation and Detection (SOLiD) became commercially available in 2007 following acquisition in 2006 by Applied Biosystems (now a division of Life Technologies). This sequencing platform is based on sequencing by ligation.

DNA libraries are prepared by annealing SOLiD-specific adapters to the template DNA fragments and adapters complimentary to those found on magnetic beads. Template DNA fragments therefore anneal to the magnetic beads via these complimentary adapter sequences and clonal amplification of the fragments proceeds via the SOLiD-specific adapters. The initial amplification steps are based on the same principle of emulsion PCR found in the 454 sequencing chemistry. The beads are then introduced to a glass slide, treated to allow them to covalently attach to its surface. The SOLiD chemistry differs from other platforms chemistries, since it progresses through incorporation of octamer probe sets, rather than the addition of individual nucleotide bases. These octamer probes are grouped into 4 sets, each set has a different fluorescent dye and represents 4 of 16 possible di-nucleotide sequences. The complementary probe hybridises to the DNA template after the sequencing primer and the dye is cleaved when the fluorescence has been read, exposing a free phosphate group. The next octamer probe is then hybridised and ligated via this free phosphate group. After 7 cycles of probe addition the read length is

35bp. More cycles can be added to yield a longer read. After the required number of cycles, the synthesised strand is cleaved and new sequencing primer is annealed in the 2nd base position of the first sequencing primer. Incorporation of the probes proceeds again for the requisite number of cycles. This is repeated several times to ensure multiple coverage of each base (Mardis 2008) www.appliedbiosystems.com.

The benefit of the SOLiD sequencing platform is that it can be scaled to allow 256 indexed samples per run, as well as the presence of 6 isolated lanes per glass slide, allowing multiple chemistries per run, which permits flexibility and reduces machine down time. The drawback of the SOLiD sequencing chemistry has remained the final read length. The SOLiD 5500 xlw released in 2010 had a read length of only 85 bp (Barba *et al.* 2014).

2.1.2.4 Illumina Sequencing

The first Genome Analyzer sequencer entered the market in 2006 launched by Solexa which was capable of sequencing 1 Gb of data per run. This technology was built by bringing together clustering, or colony DNA sequencing technology from Manteia Predictive medicine, acquired by Solexa in 2004, and technology and skills from Lynx Therapeutics acquired in 2005. This combination enabled the Solexa prototype to be developed into a commercially viable sequencing machine. Solexa was in turn acquired by Illumina in 2007 which optimised the technology to produce the HiSeq platform series which includes the HiSeq 2500 with a read length of 200 bp and the potential to sequence a whole human genome within 24 h (Barba *et al.* 2014).

The chemistry associated with the Illumina sequencers relies on sequencing by synthesis and clonal amplification of DNA fragments, resulting in cluster generation. Fragment libraries are prepared by mechanical or enzymatic fragmentation of template DNA. The final DNA fragment library is single-stranded and fragments have adapter sequences ligated to their 3' and 5' ends. The first of these adapters is complimentary to oligonucleotides present on the surface of the sequencing flow cell, and therefore anneals to this oligo. In this

way the DNA fragment library is bound to the flow cell. The oligonucleotide then acts as a sequencing primer and a complement to the single-stranded DNA template fragment is produced. Denaturation then occurs and the original DNA fragment template is removed. Clonal expansion then occurs via bridge amplification during which the second adapter sequence of the DNA fragment template anneals to the complimentary oligonucleotide on the flow cell. The complimentary DNA strand is synthesised and a so-called double stranded bridge structure is produced. When this structure is denatured it reveals the forward and reverse strand of the same DNA fragment. This cycle is then repeated until there are many copies of the DNA fragment and occurs concomitantly with the clonal amplification of the rest of the fragments in the library (Turcatti *et al.* 2008) (Shendure & Ji 2008).

Sequencing initially requires only the forward strand to produce read 1 therefore after clonal expansion the reverse strands are removed and the 3 prime end of the forward strands are blocked. Extension of the first sequencing primer occurs through competition of the fluorescently labelled nucleotides. Upon incorporation of a nucleotide a fluorescent light excites the expanded clusters of DNA fragments. This causes the emission of a signal characteristic of the nucleotide incorporated therefore the sequence of the DNA fragment is deduced, and the number of cycles of nucleotide additions controls the length of the read. Each read in a cluster and in turn each cluster on a flow cell are read concurrently. The reverse strand is then generated for each fragment in each cluster and the process is repeated to generate read 2 (Turcatti *et al.* 2008) (Shendure & Ji 2008).

Due to the high throughput and relatively low cost of the Illumina sequencing platform and its chemistry, as well as the versatility of the technology which allows its utilisation in many different applications, Illumina has become the market leader in the DNA sequencing arena. Long read technologies such as nanopore and PacBio, previously plagued by high error rates, are however being developed and refined, and are able to produce small genomes in a single

contig (Loman *et al.* 2015). The improvements to long read sequencing are increasing the popularity and usefulness of these technologies.

2.1.2.4 PacBio

The benefits of long read sequencing are in the ease of genome assembly. Larger reads allow longer overlapping regions and thus can be matched into contiguous sequences more easily than short reads with shorter overlapping regions. The difficulties in obtaining these long reads have centred on maintaining read accuracy and the comparatively lower cost per base offered by short read sequencing technologies. Although the per base cost of these short read sequencing efforts is now relatively low, the cost of producing finished genomes from the resulting data, that is genomes with all gaps closed and >99.99% accuracy per base, remains high. As a result it is now de rigour to submit genomes in an unfinished format reducing the average quality of publically available genomes. Long read sequencing is proposed as a way to combat the compromise between quality and cost (Koren & Phillippy 2015).

In 2011 the PacBio RS was released as the first commercially available long read sequencer that produced read lengths in the order of kilobases, as opposed to the few hundred bases previously available. Accuracy, however, was reduced at around 82% as opposed to the 99% offered by Illumina platforms (Koren & Phillippy 2015). These read lengths and overall accuracy meant that PacBio could not be used as a stand-alone method of sequencing and had to be used in conjunction with higher accuracy technologies such as Illumina HiSeq. The combination of the reads types did, however reduce the monetary and time cost of assembling finished bacterial genomes (Ribeiro *et al.* 2012).

With improvements to the sequencing chemistry, and the release of the second generation PacBio RS II sequencing machine, reads up to 50 kbp are now possible with an accuracy of around 87% per base. The per-read error can be further corrected through statistical methods due to the random, non-biased nature of the sequencing errors. This in turn results in a consensus sequence

with >99.99% accuracy, the accuracy required for a finished quality genome (Chin *et al.* 2013).

The associated sequencing chemistry is known as SMRT sequencing, or Single Molecule, Real-Time DNA sequencing. The reaction is conducted on a SMRT cell which contains ~150,000 Zero-Mode Waveguides where a DNA polymerase associated with a DNA template is immobilised. Phospholinked nucleotides are then added, and their incorporation into the growing complementary DNA strand being synthesised is captured in real time. By linking the fluorescent dye to the nucleotide's phosphate group the dye is cleaved by the natural process of translation, so translation does not need to be halted in order to cleave the dye. Other sequencing chemistries such as Illumina use a base linked dye moiety which requires dye cleavage before translation can progress. The SMRT cells allow this process to occur in parallel for DNA fragments covering an entire genome (Eid *et al.* 2009). In addition to the sequence data generated by PacBio, the methylation status of the DNA template can be inferred through the polymerase kinetics (Flusberg *et al.* 2010).

2.1.2.6 Oxford Nanopore

New technology being developed by Oxford Nanopore is called strand sequencing. In this method a protein nanopore is inserted into a synthetic polymer membrane with high electronic resistance. When an electronic potential is passed across the membrane it causes a current to flow through the hole in the nanopore. A sequence can be identified as a molecule of DNA is passed through the nanopore; combinations of nucleotide bases cause distinctive disruptions to the current flowing through the nanopore, which can be used to define the sequence of bases. A hairpin structure at the end of the double stranded DNA template enables both the forward and reverse sequence to be read (www.nanoporetech.com)

The sequencer hardware, the grid ION node, has the smallest footprint of any sequencer available thus far. Furthermore, there is a miniaturised version of the instrument, the MinIon, which is the size of a smart phone. The current

drawback of this platform is that it is in the early stages of release, and the chemistry and technology are in their infancy. In the future, however, this technology could have many benefits such as fast, automated library preparation and truly portable and low cost hardware (Eisenstein 2015).

2.1.3 Genome assembly

2.1.3.1 Short read data

The initial method of genome assembly involved taking a read and aligning it to the other sequence reads until a fitting overlap between the ends of the sequence occurred. This was the method employed by early assembly software, such as Celera. Although these assemblers are effective in assembling long, uncomplicated reads, they fail in the face of confounding factors such as repeat sequences which prevent detection of the correct path through the overlapping reads. This results in misassembled contigs which required manual finishing and correction (Myers *et al.* 2000) (Myers *et al.* 2000) (Pevzner *et al.* 2001).

A computational solution to the problem of repeats in a genome is the use and adaptation of a de Bruijn graph. The premise of de Bruijn assembly is to assemble reads into the fewest possible contiguous sequences of reads, or contigs, by overlapping these sequences. This method was made feasible by Pevzner *et al* with the addition of an error correction step prior to assembly, resulting in a high proportion of error free reads (Pevzner *et al.* 2001). Velvet is one such assembler which exploits de Bruijn graphs for whole genome assembly. The overlaps are controlled by defining the *K-mer* size, *K-mer* refers to the length of the overlapping words which make up a read sequence and adjacent *K-mers* within a read overlap by *K*-1. Common *K-1* sequences within the genome are known as nodes and these nodes, therefore, represent a set of overlapping *K*-mers. The complementary DNA sequence is represented by a twin node, and together these make a block. Changes to one node in a block are synonymously applied to the other. These nodes are connected by edges if the nodes overlap by *K*-2 (Zerbino & Birney 2008). VelvetOptimiser is available as a wrapper script for Velvet and functions to automatically select optimal *K-mer*

size, as well as other variable, user defined parameters with which to perform the subsequent Velvet assembly (www.bioinformatics.net.au/software.shtml).

The SOAPdenovo assembler also uses a de Bruijn based approach. The pipeline first corrects the reads to reduce the overall number of reads and therefore the memory usage of the assembly. Contig assembly then proceeds and nodes of less than 50 bp or of low coverage are removed. Similar bubbles and parallel paths are merged to further simplify the graph. The contigs are then scaffolded by mapping short reads back to them and gap closure is attempted by extracting reads which extend into the gap and performing local assembly of the region (R. Li *et al.* 2010).

A later assembler based upon the de Bruijn methodology is SPAdes, however it is described as being a universal A-Bruijn assembler. *K-mers* are established in order to draw an initial de Bruijn graph, and subsequently "forgotten" in later stages of the assembly. Downstream operations are performed based on the spatial arrangement of the graph, read coverage and sequence length. Another difference in SPAdes is the ability to identify the distance between paired reads and thus utilise the pairing in assembly (Bankevich *et al.* 2012) (Nurk *et al.* 2013).

The quest for gap closure and resolution of repeats also generated software such as SSPACE, a standalone scaffolder for pre-assembled genomes. SSPACE (SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension) allows the user to select their assembly tool and perform subsequent contig extension and/or scaffolding steps with the independent tool. The performance of SSPACE was found to be superior to that of the similar tool Abyss. The tool allows read correction, mapping of reads to contigs using Bowtie, duplicate reads are filtered, and contig extension can be carried out as an optional step using unmapped reads. After read filtering and contig extension, contigs are paired using paired reads which map to the ends of different contigs at a specified distance set by the user. Following this pairing the contigs are scaffolded if end alignments of a set *K* length support the link. The program allows the use of

both small and large insert libraries and scaffolding occurs in a hierarchical manner with small insert libraries first (Boetzer *et al.* 2011).

CLC de novo assembler creates de Bruijn graphs from *K-mer* word tables which allows matching, neighbouring words of length K-1 to be identified forwards and backwards. Bubbles are resolved by storing information about *K-mer* frequency. Since increased *K-mer* length can help to resolve repeat regions, CLC de novo assembler automatically selects the optimal *K-mer* based on the data input. Contig lengths are increased by mapping reads back to the assembled sequences and a scaffolding step is also performed. The major advantage of this assembler over the others discussed is that it eliminates the need to select *K-mer* size by trial and error, however the other software discussed is available on an opensource basis, whereas CLC de novo assembler is available to purchase (www.clcbio.com/files/whitepapers/whitepaper-denovo-assembly-4.pdf).

As the PacBio platform and long read data becomes more widely used, assembly software is being updated to support long read data as well as short read data; SPAdes and SSPACE both have the capacity to support long read data in their most recent releases. This data can be used to scaffold pre-existing contigs to "upgrade" draft genomes to produce a more complete assembly with reduced contig number (Boetzer & Pirovano 2014) (Nurk *et al.* 2013).

2.1.3.2 Integrating long and short read data

Due to the error prone nature of long read data when PacBio was first released, the data was initially applied to enhance and finish assemblies from short read data. Several assembly platforms were developed to integrate the two types of data. The benefits of integrating the two types of data offer a solution to the problem which has hindered computational assembly of short read data: repeat regions. The goal of short read assembly pipelines has remained the pursuit of genomes assembled in the fewest number of contigs, which have the longest possible sequence. This is described by the N50 statistic which states that half of the length of the sum of all contigs is contained within contigs of this length or longer.

PBJelly was launched in 2012 to improve the quality of draft genome sequences. Although the platform was designed to facilitate using PacBio data, any long reads could be supported. PBJelly improved gaps between contigs by extending neighbouring contigs into the gaps between them, a closed gap refers to two neighbouring contigs joined by sequence constructed from the long read data. The assembly of the draft genome is taken to be accurate and any gap fills can be accepted or rejected through the verbose logging of any changes made. Another useful feature of this improving assembler is that annotations from the draft genome are stored in co-ordinate tables so they can be easily superimposed upon the improved genome, further minimizing the workload required for improving and closing existing draft genomes (English *et al.* 2012).

One of the first assemblers to generate a finished quality microbial genome through single molecule sequencing was ALLPATHS-LG. At the time the software was published the PacBio read length was around 2-3 kb. The pipeline combined Illumina mate pair and paired end libraries to produce a de Bruijn assembly of the genome and gaps and repeats were subsequently covered and resolved by the long read PacBio data (Ribeiro *et al.* 2012). The downside to this method is that it required three different sequencing libraries to produce a finished genome, which leads to prohibitive costs.

2.1.3.3 Long read data

High quality assemblies from long read data require an adequate number of redundant long reads from the spectrum of read sizes produced to span and therefore resolve all repeat regions in a genome. With the current PacBio sequencing chemistry the coverage required to achieve the above criteria, and therefore a finished quality assembly is ~50x. Although early releases of the PacBio chemistry, C1 and C2 versions, did not allow for sufficiently long or accurate reads to allow for read self-correction and assembly of data without inclusion of short read data, subsequent releases allowed stand alone assembly of PacBio data.

PBcR is a correction and assembly pipeline which deals specifically with the high degree of noise per read associated with long read data. The pipeline improves accuracy per single-molecule sequence and then assembles these error-corrected consensus sequences using the Celera Assembler. The program exploits genome coverage using the short reads produced by PacBio sequencing to produce a consensus sequence. These sequences are then assembled through overlapping regions, and the long reads generated are used to resolve repeat regions. Quiver is then used to polish by mapping the raw reads to the assembly to correct indels and substitutions. PBcR represents an improvement on the ALLPATHS-LG program because complete genomes could be assembled from only one library. The HGAP assembly software distributed by PacBio with its genome sequencers is a variation of the pipeline, although without the added capability of PBcR to support the use of short read data in data correction as well as data generated from the Oxford Nanopore sequencing platform (Koren *et al.* 2013).

Most recently MHAP, for MinHash Alignment Process, has been developed which allows rapid assembly of large genomes, up to 600-fold faster than PBcR and HGAP. The MHAP uses a probabilistic method to create assemblies based on overlap between reads. MHAP is integrated into the PBcR pipeline (Berlin *et al.* 2015).

2.1.4 Chapter Aims

In this chapter the key aim was to generate sequences of several staphylococcal species for which no genome was available to use for comparative analysis. A second aim was to sequence a collection of *S. hominis* isolates to enable a subsequent investigation of genomic features of this species. Within these aims a selection of the sequencing technologies and assembly softwares described above were tested and compared to examine the quality of the assemblies of staphylococcal genomes. Furthermore, in an effort to improve the quality of the *S. hominis* genome, PacBio sequencing was used to resolve gaps for the strain J31.

2.2 Methods

2.2.1 Bacterial strains and culturing

A collection of skin isolates was established through sampling of the inner elbow (provided by Dr Jennifer Kelly (Kelly 2013))and the anterior nares (provided by Dr Benjamin Liberton (Libberton *et al.* 2014)). All isolates of *S. hominis* and *S. haemolyticus* sequenced as part of this study were drawn from these strain collections. *S. simulans* ATCC27848, *S. cohnii* ATCC29974 and *S. xylosus* ATCC29971 type strains were provided by Ross Fitzgerald (Laboratory for Bacterial Evolution and Pathogenesis, University of Edinburgh).

Overnight cultures of all strains were grown for 18 h at 37 °C in 10 ml Oxoid Brain Heart Infusion broth (3.7 % w/v BHI) (Thermo Scientific) with shaking at 200 rpm. Glycerol stocks of all strains were maintained by adding 700 µl of overnight culture to 300 µl 50% (v/v) glycerol. Stocks were stored at -80 °C.

2.2.2 Extraction of genomic DNA

Cells were harvested from a 1.5 ml sample of a 10 ml overnight culture using centrifugation at 13000 rpm for 5 min; harvested cells were then resuspended in lysis buffer containing 12.5 µg ml^{-1} lysostaphin (Sigma-Aldrich) and 2 µl of 5 kU mL^{-1} mutanolysin (Sigma-Aldrich). A DNeasy Blood and Tissue Kit (Qiagen) was used to purify the DNA according to manufacturer's instructions for Gram-positive bacteria. Protocol alterations were made: Both lysis and proteinase K incubation steps were extended to 1 h; an on column RNA digest was added as specified by the DNeasy Blood and Tissue Kit manufacturers instructions using 16 µL solution of RNase A (Invitrogen) at 20 mg mL^{1}; an extra wash step was performed with wash buffer 1.

2.2.3 DNA Quality Checking

The purity of DNA was measured using a Nanodrop spectrophotometer (Thermo scientific); samples with 260/280 and 260/230 ratios above 1.8 were assumed to have minimal protein and RNA contamination and were used for sequencing.

DNA molecular weight was assessed after separation using agarose gel electrophoresis with ~2mm thick 0.8% agarose gels run at a current of 80 V for 1 h to ensure a tight band of DNA. Smearing of the DNA indicated the sample was fragmented by the extraction protocol and was not selected for sequencing. In all cases where agarose gel electrophoresis was applied TAE buffer, UV transillumination and GeneSnap gel imaging software (Syngene) were used. 2 µl of Midori green (Nippon Genetics) DNA stain was used in 100 mL of agarose.

DNA quantification of purified samples was carried out using Qubit (Invitrogen) according to manufacturers instructions.

2.2.4 Bacterial species confirmation

Prior to every whole genome sequencing isolate speciation was confirmed using 16S rDNA sequencing (GATC-Biotech). The 16S region was amplified using the primers pA and pH. PCR and thermocycling conditions were used as follows: initial denaturation for 5 min at 95 °C was followed by 30 cycles of denaturation at 95 °C for 60 s, annealing at temperatures specific to individual primers for 60 s and extension at 72 °C for 1 min per 1 kb of product. Lastly a final extension at 72 °C for 7 min.

Agarose gel electrophoresis was used to separate DNA bands to ensure lack of contamination in the reaction and specificity. Each PCR product was then purified before using ExoSAP-IT to dephosphorylate dNTPs and primers and degrade the latter. Specifically, 4 µl of ExoSAP-IT (Affymetrix) was added to every 15 µl of PCR product. The mixture was vortexed for 10 s and incubated at 37 °C for 30 min, followed by a 15 min inactivation of the enzymes at 80 °C. Sanger sequencing of the 16S rDNA PCR products was then carried out by GATC-Biotech. The DNA sequences were analysed using both BLASTn and the RDP seqmatch program (http://rdp.cme.msu.edu/). The closest matching sequences were determined taking into account blast score, blast E-value, blast maximum identity and RDP e-value.

2.2.5 Library preparation for whole genome sequencing

2.2.5.1 Illumina TRUSEQ

Illumina TRUSEQ small insert libraries (paired end) were prepared with 300-500 bp insert sizes and were prepared by myself according to manufactuer's instructions (Illumina). Next, Illumina TRUSEQ large insert libraries (mate pair) were prepared with an insert size of 8 kb by the Centre for Genomic Research, University of Liverpool. All TRUSEQ libraries were sequenced on the HiSeq 2500 platform.

2.2.5.2 Illumina Nextera

Illumina NEXTERA libraries were prepared by myself according to manufacturer's instructions (Illumina). Library QC was carried out using the 2100 Bioanalyser (Agilent Technologies) and High Sensitivity DNA chips (Agilent Technologies). Nextera libraries were sequenced on both the HiSeq 2500 and the MiSeq personal sequencer platforms as specified by the Centre for Genomic Research, University of Liverpool.

2.2.5.3 PacBio Sequencing libraries

PacBio sequencing libraries were prepared and sequenced on the PacBio RS II platform by the Centre for Genomic Research, University of Liverpool.

Table 2.1 Comparison of the quality of genome assemblies using the assemblers Velvet, CLC workbench and SPAdes. Scaffolding with SSPACE was performed on the contigs produced from all assemblies.

Species	Assembler	Scaffold number	N50 (bp)	% mapping of MP reads (reads which mapped 1X)
S. cohnii	VelvetOptimiser	84	578,102	69.81
	CLC workbench	85	902,047	69.95
	SPAdes	764	487,394	68.56
S. simulans	VelvetOptimiser	47	1,589,405	69.92
	CLC workbench	41	2,420,180	70.65
	SPAdes	240	833,978	70.65
S. xylosus	VelvetOptimiser	53	775,104	78.13
	CLC workbench	30	2,063,813	78.98
	SPAdes	1000	939,092	76.21

2.2.6 Genome assembly

Sequence reads were adapter trimmed and quality filtered by the CGR (University of Liverpool).

Genome assemblers were compared for *de novo assemblies* of staphylococcal species from Illumina TruSEQ data generated in the initial round of sequencing. See Table 2.1 for strain information. VelvetOptimiser (Velvet version 1.2.06) was used with the *K-mer* hash of values between 19 and 99. Velvet parameters are automatically optimised and assigned by VelvetOptimiser. Forward and reverse reads in fasta format were used as input for this assembler. SPAdes (version 2.2.1) was run with the *K-mer* sizes 21,33 and 55. Forward and reverse reads in fastq format were used as input. CLC genomics workbench (version 5.0) genome assembler was used and *K-mer* length is automatically optimised. Forward and reverse reads in fastq format were used as input for this assembler.

Scaffolding was carried out using SSPACE-BASIC (version 2.0) for initial *de novo* sequencing with contig extension parameters of –m = 50 and –o = 15. A contig file in fasta format and a library file in tab delimited text format are required as input for this scaffolder. The library file specifies the library types and signposts the scaffolder to the libraries to use for scaffolding and extending the contigs. Information about insert size is also included.

Scaffolding of resequenced *S. hominis* genomes was carried out using the perl script *actcompare.pl* (developed by Roy Chaudhuri for the CGR, University of Liverpool). Actcompare uses NUCmer to align query genomes to a reference and outputs a GenBank file of concatenated contigs which can be launched in the Artemis Comparison Tool (ACT). Contig files in fasta format were used as the input for this scaffolder.

2.2.7 Genome annotation

Annotation of all genomes was carried out using PROKKA version 1.5.2. PROKKA is a command line tool, implemented in Perl, which is capable of

annotating a bacterial genome on a desktop computer in ~10 min. Genome features are predicted using the external programs Prodigal, RNAmmer, Aragorn, SignalP and Infernal (Seemann 2014). A genome in concatenated fasta format was used as input for this program.

2.2.8 *S. hominis* genome assembly improvement

The *S. hominis* strain J31 was sequenced using the PacBio platform as described above. The genome was assembled by the CGR (University of Liverpool) using the PacBio distributed assembler HGAP.

Circularity of the contigs was assessed by aligning several kb from the beginning of the contig back to the contig itself using command line blast. Significant hits to the beginning and the end of the contig indicated that the contig was circular. This was used to assess if the genome of *S. hominis* J31 was present in a single contig. Smaller contigs present in the *S. hominis* J31 PacBio assembly were also assessed for circularity in the same way and designated as putative plasmid contigs if circularity was observed.

This single contig alignment was then used to improve the contig arrangement of all other *S. hominis* genomes, both those sequenced as part of this project and those downloaded from public databases. This improvement was achieved by rearranging the contigs of the *S. hominis* genomes with *actcompare* using the J31 PacBio assembly as a reference genome.

2.3 Results and Discussion

2.3.1 Assembly of staphylococcal genomes and comparison of assemblers

The primary aim of this thesis was the comparative genomic analysis of the major staphylococcal species. At the beginning of the project there were no complete reference genome sequences or draft genome sequences for *S. cohnii, S. simulans* or *S. xylosus.* Sequence data was available for *S. hominis* and *S. haemolyticus*. A previous comparative genomic analysis of the *Staphylococcus* genus included only 12 of the 51 staphylococcal species (Suzuki *et al.* 2012). This study aimed to encompass a larger amount of strains and include more species known to be found on human skin (Kloos 1980), therefore *S. simulans, S.cohnii* and *S. xylosus* type strains were sequenced. Type strains were selected for these *de novo* sequencing efforts since they are readily available to the scientific community. Furthermore, as no reference genomes were available for these species, both mate pair and paired end sequencing libraries were used to scaffold contigs assembled from paired end reads in order to better resolve the macro structure of the genome. As reference assemblies were available for both *S. hominis* and *S. haemolyticus,* only paired end reads were used to assemble contigs and these contigs were rearranged in relation to these reference sequences.

During the course of this research study, in 2013, the whole genome sequence of *S. simulans* was published (Calcutt *et al.* 2013) and in 2014 the genome sequences of *S. cohnii* (Hu *et al.* 2014) and *S. xylosus* (Labrie *et al.* 2014) were also published.

As described, there is a plethora of assembly platforms available to assemble and scaffold short read data. Each of them has its advantages and disadvantages, however there is no gold standard for genome assembly. Furthermore, different species have different genome features such as relatively high prevalence of repetitive sequences, or an unusual GC content, which affect the fidelity of sequencing and the ease of assembly.

To assess which available assembler produced the optimal assemblies of the staphylococcal genomes sequenced here for use in comparative genomic analysis, three platforms were tested and compared: SPAdes, CLC and Velvet. Both the N50 and scaffold number are accepted bench marks of genome assembly quality and so these are the parameters considered here. It must be cautioned, however, that these measures of assembly success focus on contig length and not necessarily accuracy. A high N50 and low contig or scaffold number is assumed to represent an assembly in which contigs traverse and therefore resolve repeat regions. An alternative possibility is that such an assembly contains large, misassembled contigs (Narzisi & B. Mishra 2011). In this study, mapping of mate pair (MP), or large insert reads, back to the assembly was used as an assessment of the resolution of the macrostructure of the genome. A high degree of contig misassembly would result in low percentage mapping of the MP reads.

The differences in assembly quality bench marks between the three assemblers were compared as part of this thesis. The three assemblers were compared for the three de novo sequencing efforts; the *S. cohnii, S. simulans* and *S. xylosus* genomes (Table 2.1). In the absence of a reference genome to scaffold against, contig extension and scaffolding was performed using SSPACE.

In all cases SPAdes with contig extension and scaffolding performed with SSPACE performed the most poorly where contig number is concerned producing 764, 240 and 1000 contigs respectively for *S. cohnii*, *S. simulans* and *S. xylosus*, respectively. With *S. cohnii* and *S. simulans,* SPAdes also produced the lowest N50 of 487,394 bp and 833,978 bp, respectively. For *S. xylosus* however, the N50 produced by SPAdes was 939,092 bp. This is higher than the N50 produced by Velvet which was 775,104 bp (Table 2.2).

With all three genomes CLC workbench produced the best assembly in terms of N50 and scaffold number. The N50 was 902,047 bp, 2,420,180 bp and 2, 063,813 bp for *S. cohnii*, *S. simulans* and *S. xylosus*, respectively. The number of scaffolds was 41 and 30 for *S. simulans* and *S. xylosus*, while *S. cohnii* was

assembled into 85 scaffolds. The extent of mapping MP reads was >70% for both *S. simulans* and *S. xylosus*, but below this value for *S. cohnii* (Table 2.2).

With all three assemblers *S. cohnii* underperformed in comparison to *S. simulans* and *S. xylosus* for all three bench marks (Table 2.2). This differential could indicate a higher proportion of repetitive sequence in the *S. cohnii* genome which is difficult to resolve and can result in more fragmented assemblies.

From these comparative genome assembly results CLC workbench was considered to be the best platform of the three tested for the assembly of staphylococcal genomes from Illumina sequence data. This assembler was therefore used to assemble all genomes in this study used for inter- and intraspecies genome comparisons.

Table 2.2 Sequencing and assembly statistics for genomes sequencing as part of genome comparison efforts								
Species	Isolate	Assembler	Scaffolding/ rearrangement method	Sequencing platform	Library preparation	No. of contigs/scaffolds	N50 (bp)	Coverage (X)
S. hominis	I4	CLC workbench	*actcompare*	Illumina HiSeq 2500	TRUSEQ (PE)	83	167,692	277
S. hominis	J6	CLC workbench	*actcompare*	Illumina HiSeq 2500	NEXTERA	45	298,133	833
S. hominis	J11	CLC workbench	*actcompare*	Illumina HiSeq 2500	NEXTERA	51	298,149	1036
S. hominis	J23	CLC workbench	*actcompare*	Illumina HiSeq 2500	NEXTERA	50	193,853	760
S. hominis	J27	CLC workbench	*actcompare*	Illumina HiSeq 2500	NEXTERA	40	230,743	499
S. hominis	J31	CLC workbench	*actcompare*	Illumina HiSeq 2500	NEXTERA	633	313,790	1101
S. hominis	J31	HGAP			PacBio	5	2,188,325	205
S. hominis	B10	CLC workbench	*actcompare*	Illumina HiSeq 2500	NEXTERA	37	1,174,157	1117
S. haemolyticus	K8	CLC workbench	*actcompare*	Illumina HiSeq 2500	TRUSEQ (PE)	125	54,979	195
S. cohnii	ATCC29974	CLC workbench	SSPACE	Illumina HiSeq 2500	TRUSEQ (PE + MP)	85	902,047	115
S. simulans	ATCC27848	CLC workbench	SSPACE	Illumina HiSeq 2500	TRUSEQ (PE + MP)	41	2,420,180	115
S. xylosus	ATCC29971	CLC workbench	SSPACE	Illumina HiSeq 2500	TRUSEQ (PE + MP)	30	2,063,813	133

2.3.2 Annotation of sequenced staphylococcal genomes

The genomes of all staphylococcal species sequenced were annotated using PROKKA: a command line tool, implemented in Perl, which is capable of annotating a bacterial genome on a desktop computer in ~10 minutes. The number of protein coding sequences (CDS) ranged from 2187 in *S. hominis* J23 to 2642 in *S. simulans* (Table 2.3).

Table 2.3 Genome features of sequenced and assembled staphylococcal genomes. Staphylococcal genomes sequenced as part of the genomic comparison were assembled and annotated by Prokka.

Species	Isolate	Protein CDS	tRNA	rRNA	GC %	Source
S. hominis	I4	2202	54	4	31.3	Skin
S. hominis	J6	2193	50	3	31.2	Skin
S. hominis	J11	2192	52	3	31.2	Skin
S. hominis	J23	2187	55	4	31.2	Skin
S. hominis	J27	2188	54	5	31.3	Skin
S. hominis	B10	2253	55	3	31.3	Skin
S. cohnii	ATCC 49330	2611	58	8	32.3	Skin
S. haemolyticus	K8	2382	54	3	32.7	Skin
S. simulans	ATCC 27851	2642	59	7	35.9	Skin
S. xylosus	ATCC 29971	2541	50	5	32.7	Skin

The tRNA gene family is the largest known gene family and its repertoire in a species is important as it affects the major codon bias observed in bacterial genomes (Schattner *et al.* 2005). This major codon bias is implicated in the global optimisation of cell growth efficiency, translation efficiency and control of cellular processes (Quax *et al.* 2015). The tRNA repertoire ranged from 50 genes in *S. hominis* J6 and *S. xylosus* to 59 genes in *S. simulans.* The tRNA repertoire is expanded in the PacBio assembly of *S. hominis* J31 and contains 70 genes.

The rRNA gene operon poorly assembles in many whole genome sequences, however there are often multiple copies of this operon in bacterial genomes. At high growth rates multiple copies of the rRNA operon and the resulting ribosomes are thought to offer an advantage by enabling the bacterial cell to respond quickly to favourable environmental conditions by increasing translation. Seven copies of the rRNA operon are observed in *E. coli* and *S.*

enterica serovar typhimurium, and 10 copies are found in the *B. subtillis* genome (Klappenbach *et al.* 2000). In the Illumina sequenced staphylococci the number of rRNA genes annotated ranges from three copies in *S. hominis* strain J6 to eight copies in *S. cohnii.* The number of copies of rRNA genes was compared between the Illumina assembly of J31 and the PacBio assembly. In the Illumina assembly one 16S rRNA, one 23S rRNA and two 5S rRNA genes were annotated. In contrast the PacBio assembly contained eight 16S, eight 5S and seven 23S rRNA genes annotated (Table 2.3). Furthermore, these annotated rRNA genes are distributed throughout the genome indicating that they have been correctly assembled. Clustering of these genes at the end of the genome sequence among genes associated with plasmids and phage would indicate that these genes had been poorly assembled. This indicates that assemblies from short read data lead to underestimation of rRNA gene copy number in the staphylococci

It is important to take into consideration the GC content of a genome given that the GC content bias may affect read coverage. GC rich locations of a genome can be prone to high coverage whereas low GC areas can exhibit a deficit in read coverage. This sequencing artefact can confound results if gene copy number or expression levels are being investigated (Benjamini & Speed 2012). In genomes with low GC content, such as staphylococcal genomes, this GC bias can result in fragmented assemblies due to the fact that many assemblers assume an even GC coverage across the genome (Chen *et al.* 2013). Table 2.3 demonstrates that all staphylococcal genomes sequenced here have a consistent GC content between 31 - 36% and that the use of PacBio sequencing, which does not exhibit the GC sequencing bias present in the Illumina sequencing chemistry, does reveal a significantly different GC content for *S. hominis* strain J31 when compared to Illumina sequence data (Table 2.4).

Table 2.4 Comparison of genome features of *S. hominis* J31. Assemblies from Illumina and PacBio sequence data were annotated by Prokka.		
Genome feature	Illumina	PacBio
Protein CDS	2635	2233
tRNA	54	70
rRNA	5	23
GC %	31.4	31.5
Source	Skin	Skin

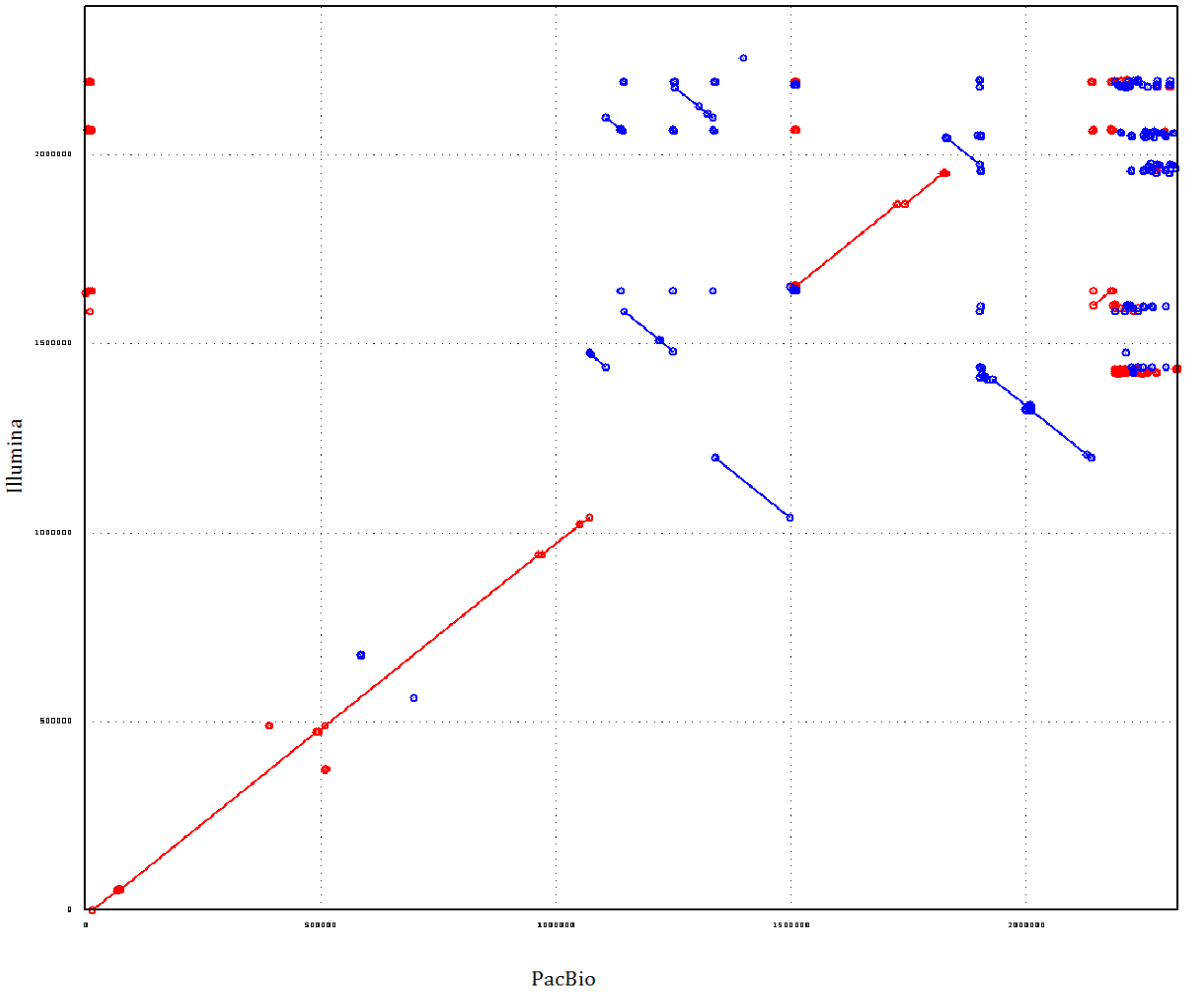


Figure 2.1 Genome synteny between *S. hominis* strain J31 PacBio asembly and Illumina assembly. Synteny is shown by a nucleotide mummer plot where blue lines indicate inversions and red lines indicate forward alignments between the two assemblies. Translocations are indicated by deviations from the hypothetical line f(x)=x

2.3.3 Improvements to the *S. hominis* genome using long read sequence data

PacBio sequencing was used to produce an improved assembly of *S. hominis* strain J31. Pacbio sequencing of this strain resulted in one large contig, plus a further 4 smaller contigs (Table 2.5). The PacBio assembly was compared to the Illumina assembly by aligning the two using a nucleotide mummer plot (Figure 2.1). Regions of similarity between the two genomes, that is forward alignments with the same topography, are shown in red on the hypothetical line f(x)=x. Reverse alignments are shown in blue and translocated regions are represented by deviations from f(x)=x. The mummer plot shows that many of the Illumina contigs are assembled differently compared to the regions in the PacBio genome.

Table 2.5 Contig lengths and read coverage per contig resulting from the PacBio assembly of *S. hominis* J31

Contig	Length	Coverage (X)
0	2,188,298	207
1	60,147	256
2	28,669	92.5
3	40,632	117.5
4	6417	45

Further investigation of the PacBio contigs showed that the longest contig, numbered 0, had coverage of 207X and represents the bacterial chromosome (Table 2.5). Evidence for this contig being the entire, circular bacterial chromosome is found in a 15000 bp region from the start of the contig being repeated at the end. The gene content and synteny of the region is conserved and the presence of an rRNA gene operon is observed. This indicates that some of the tRNAs in the expanded repertoire observed in the J31 assembly of PacBio data, and 4 of the rRNA genes annotated, may be as a result of a sequencing artefact, in addition to the cluster of *fol* genes present at the end of the chromosome contig (Figure 2.2). When this 15000 bp region is blasted against contig 0 there are an additional 4 significant hits along the contig (Figure 2.3). When the genes in the region of these blast hits were investigated the rRNA operon was found to be the only coding sequences present.

The 4 smaller contigs were also examined to investigate if these contigs might represent complete circular plasmid DNA. Bases from the beginning of the contig were blasted against the contig itself and the sequence was assumed to be circular if this produced significant blast hits to both the start and end. Further evidence of the circularity of these contigs can be found in the annotation of the plasmids discussed in chapter 3 (3.3.15.1) as gene annotation from the beginning of the plasmids is repeated at the end. The exception to this is contig 4 which represents the smallest contig. Although there are significant blast hits, of the first 1000 bases to the beginning and the end of the contig, there is also a significant blast hit of the same region equidistant between the two (Figure 2.4 D). One explanation for this may be misassembly of the contig. Another, however, is that this contig represents two copies of a small plasmid. This contig also exhibits lower read coverage (Table 2.5), which may be as a result of the fact that the contig is approaching read length for PacBio sequencing.

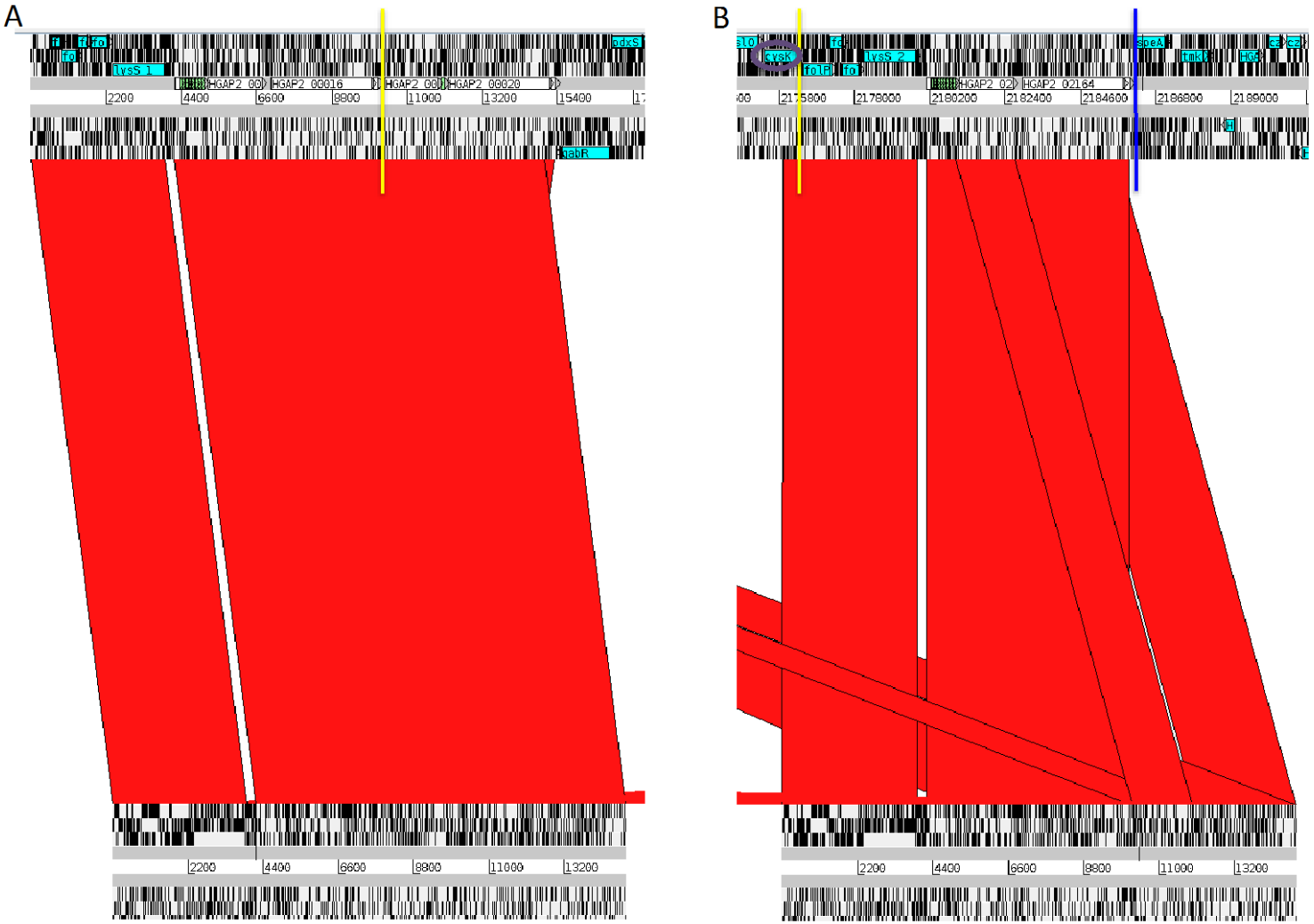


Figure 2.2 Detailed analysis of the alignment of the first 15 000 bp of contig 0 with all contigs of the J31 PacBio assembly. A) the first 15 000 bases of contig 0 aligned with the start of contig 0. The yellow line indicates the end of the speculated overlap region. Genes in the region are annotated (left to right) folP_1, FolP_2, FolB_1, FolK_1 and LysS_1. These genes are followed by rRNA 5S gene, a cluster of rRNAs, rRNA 16S, 23S and 5S genes. B) the first 15 000 bases of contig 0 aligned with the end of contig 0. The yellow line indicates the start of the speculated overlap region, the blue line indicates the end of contig 0. Genes in the region are annotated (left to right) FolP_3, FolB_2, FolK_2 and LysS_2. These genes are followed by rRNA 5S gene, a cluster of rRNAs, rRNA 16S, 23S and 5S genes.

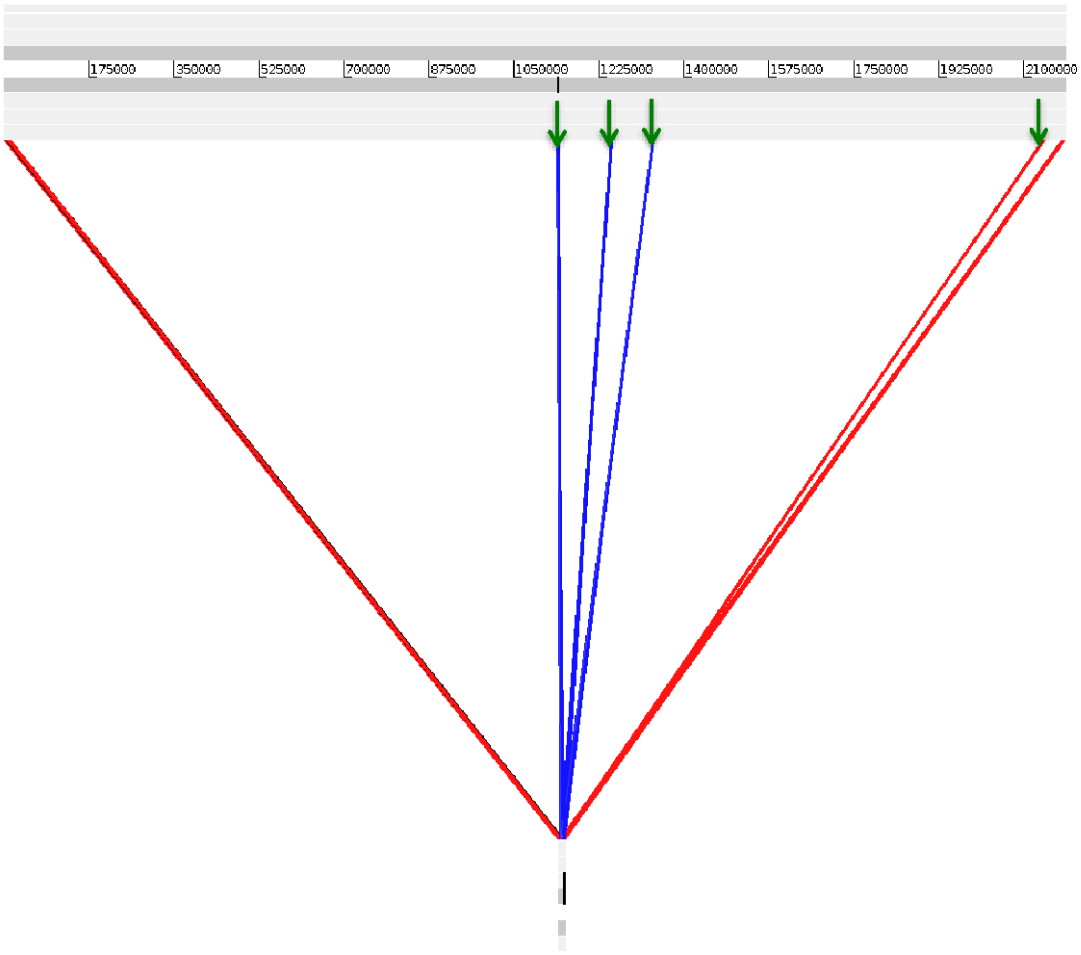


Figure 2.3 Alignment of the first 15000 bases of contig 0 with the rest of contig 0. Green arrows indicate the locations of distributed rRNA operons which occur as matches to a region of the first 15000 bases containing rRNA genes. The regions indicated by the green arrows are around 5500 bp in length and contain only rRNA genes.

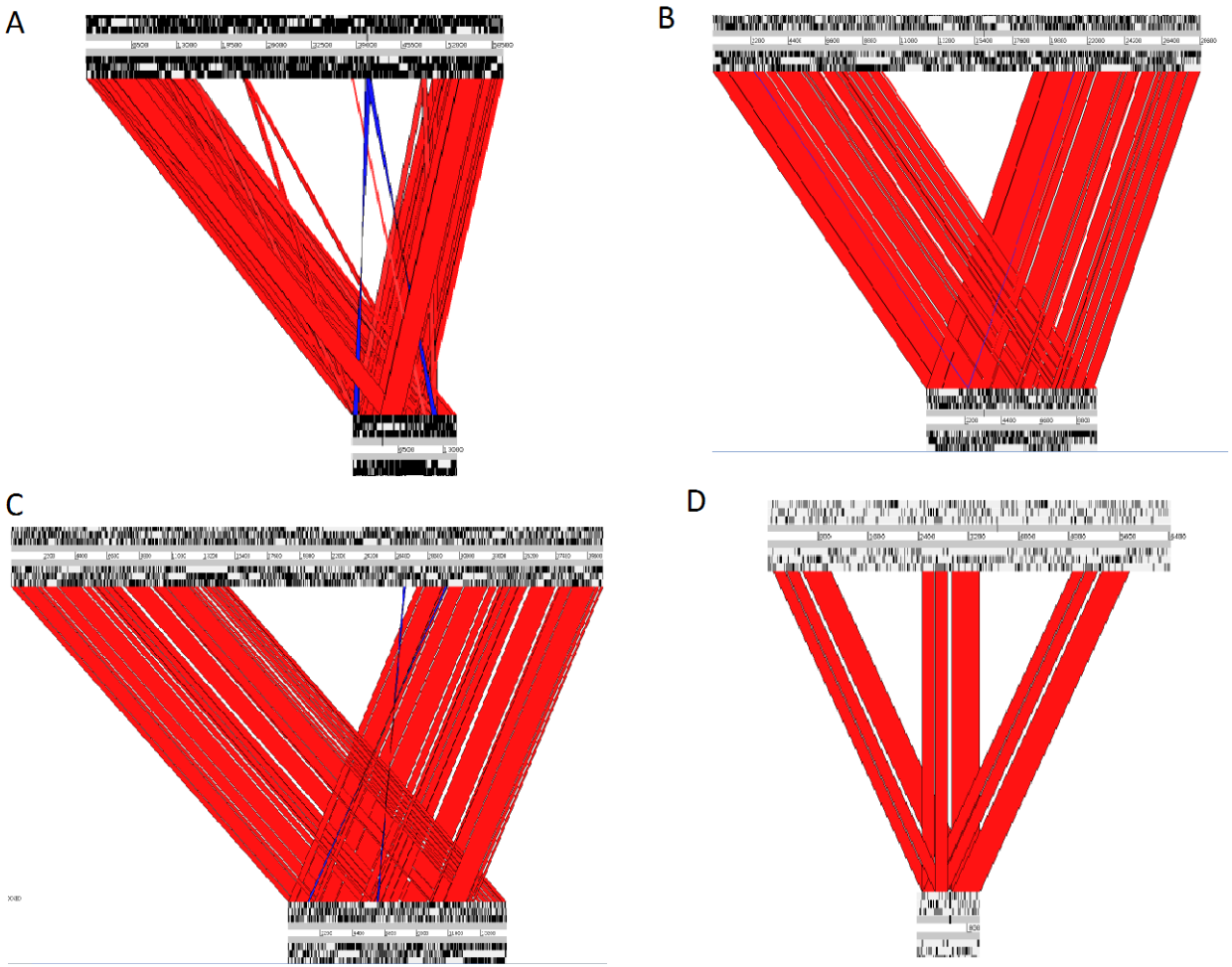


Figure 2.4 Visualisation of the circularity of the contigs produced from the PacBio assembly of S. hominis strain J31. A) The first 15,000 bp of contig 1 aligned against contig 1 B) The first 10,000 bp of contig 2 aligned against contig 2 C) The first 15,000 bp of contig 3 aligned against contig 3 D) The first 1000 bp of contig 4 aligned against contig 4

2.3.4 Gene synteny of the *S. hominis* genomes sequenced

Since the large contig resulting from PacBio sequencing was taken to be the circular bacterial chromosome, this assembly is considered to be of better quality compared to the Illumina assembly of the J31 genome by the metrics stated earlier, namely, a large N50 and low contig number. Consequently, the contigs resulting from CLC workbench assembly of the *S. hominis* genomes sequenced on the Illumina platform were reordered relative to the PacBio HGAP assembly of *S. hominis* J31.

Alignments of the reordered assemblies were performed using Progressive Mauve and visualised using Mauve software (version snapshot 2015-01-25 (1)). Local collinear blocks (LCBs) are used to represent homologous backbone sequence. Rearrangements such as translocations and inversions can be identified by LCBs in different locations on the backbone. The high degree of synteny among the *S. hominis* genomes is indicated by the conservation of the arrangements of the LCBs (Figure 2.5). This synteny, together with the differential assembly observed between the J31 Illumina and PacBio assemblies (Figure 2.1), indicates that rearrangements of contigs against a PacBio reference can result in an improved arrangement of assembled contigs.

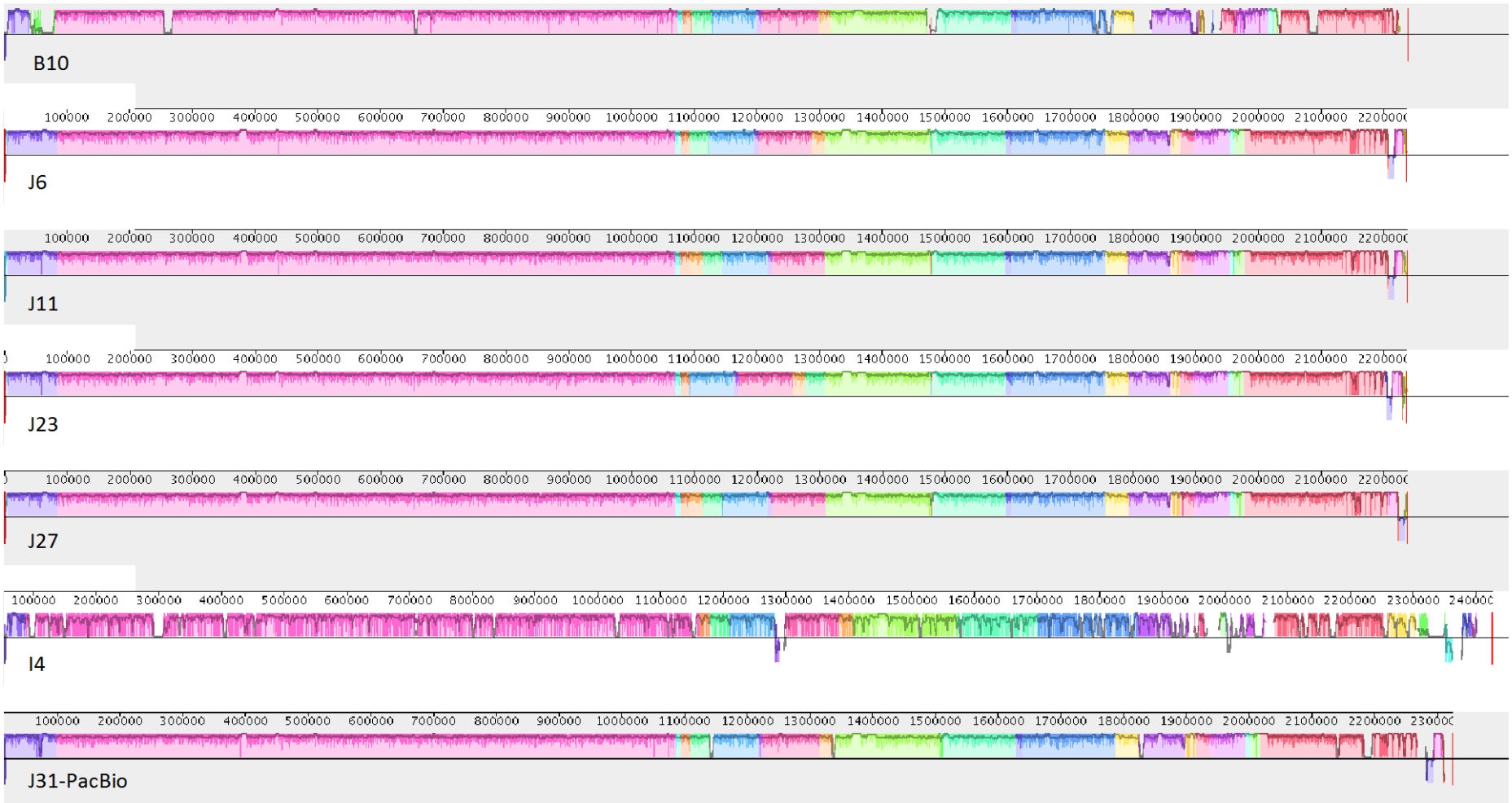


Figure 2.5 Progressive Mauve genome alignment of the *S. hominis* genomes. Contigs resulting from CLC workbench assemblies of Illumina sequence data rearranged using HGAP alignment of PacBio sequence data from strain J31 . The colour bloack indicate homologous regions without any internal rearrangements. Known as Locally Colinear Blocks.

2.4 Conclusion

Over the course of this project improving sequencing technologies have been exploited and have demonstrated some of the principles known about the growing field of whole genome sequencing. The Illumina platform has improved to give longer read length and a greater depth of coverage. In the first round of Illumina sequencing ~200X was considered to be deep coverage. In the second round of Illumina sequencing this coverage was greatly exceeded (Table 2.2) however the assembly quality did not improve in proportion to the vastly increased read coverage. This assembly limitation reveals the fundamental downside of short reads given the vast coverage increases, a finding highlighted by other studies (Koren *et al.* 2013).

Moreover, the benefits of long read sequencing are clear to see with the resolution of the *S. hominis* J31 genome into a single chromosome, compared to the assembly from Illumina short read data which resulted in 633 contigs (Table 2.2). Despite this great advantage of long read sequencing over short read sequencing there are also drawbacks. Due to the nature of the large fragment sizes and long read length which enable long read sequencing to produce automated assemblies of whole genomes, small extra-genomic features such as small plasmids and free phage are more difficult to resolve as described with contig 4 of the J31 PacBio genome assembly (Figure 2.4).

As large banks of species-wide whole genome sequence data from short reads have been collected, perhaps long read data can help researchers to fully exploit these resources. This project has reinforced that contig arrangements of assemblies from Illumina data can be improved by alignment to a PacBio genome.

Chapter 3

Intraspecies comparative genomics of *Staphylococcus hominis*

3.1 Background

3.1.1 Staphylococcal speciation

Since the late 19th century staphylococcal classification has been inextricably linked with the ability to cause disease and the aetiology of this disease. *Staphylococcus aureus* was first designated *S. pyogenes aureus* and *Staphylococcus epidermidis* was known as *S. pyogenes albus* for the colour of the pus associated with infection by each organism(Kloos 1980). Despite some bacteriologists at the time questioning the innocence of commensal staphylococci emphasis continued to be placed on distinguishing the culprit of the yellow pus, believed to be a single species of staphylococci now known as *S. aureus*, from the rest. For almost a hundred years the genus of *Staphylococcus* was considered to contain only two species, *S. aureus* and *S. epidermidis*. This opinion was challenged by Baird-Parker who began to subdivide *S. epidermidis* and *S. aureus* into biotypes. He posited that these biotypes may indeed be defined species. Although Baird-Parker refrained from evangelising this opinion he laid the foundation work for the modern taxonomy of the staphylococci. However, even as *S. warneri*, *S. epidermidis*, *S. haemolyticus* and *S. hominis* were designated as new species, overlapping characteristics were observed among them (Kloos & Schleifer 1975).

Bacterial species designation has progressed from being purely phenotypic, through being driven by pairwise DNA-DNA hybridisation to taking a systematic approach based on 16S rDNA clustering methods (Cohan 2001). A protein sequence clustering approach has now shown a higher resolution for distinguishing between closely related species where the slow rate of evolution of 16S rDNA has failed to differentiate the two species (Palys *et al.* 2000).

Together with the presence of closely related genes, the question of niche selection is also relevant where species evolution is concerned. Investigating the possibilities that acquisition of new genes or changes to existing ones facilitate the jump to new niches is being aided by increasingly large collections of geographically and ecologically distinct isolates of a given species and the availability of their genome sequences. Large repositories of genome sequences

facilitate the identification of genes important to niche adaptation in collections of bacterial species (Gupta *et al.* 2015) (Y. Zhang & Sievert 2014) (Lowder *et al.* 2009). It is hoped that the genome sequences of *S. hominis* sequenced as part of this work will contribute to the understanding of genes important for survival on human skin.

3.1.2 Pan-genome analysis and speciation

While genome sequencing continues to become increasingly accessible the growing number of bacterial whole genome sequences has begun to highlight the sequence variation between isolates of the same species. In the face of this breadth of genomic information, pan-genome analyses contribute to the definition of a species and have quantified the extent to which intra-specific genetic diversity can occur. Standard microbiological techniques to identify isolates, in the clinic for example, mostly use a gene marker-based approach, most commonly 16S rDNA sequencing. Although this approach is able to discriminate to a species level, as previously stated, it is not capable of distinguishing very similar species. Furthermore it does not take into account the degree of genetic variability of the strain and its possible phenotypic implications. This is illustrated by the pan-genome investigation of *Bacillus anthracis*. This species exhibits a closed pan-genome indicating a low degree of variability and appears phylogenetically to be a clone of *B. cereus*, not a distinct species. The discriminatory phenotypic trait, production of the anthrax toxin, is encoded on one of the two plasmids that differentiate *B. anthracis* from *B. cereus* (Rasko *et al.* 2005). Despite such important phenotypic differences between each species, it could be argued that this is simply minor variability in the accessory genome of *B. cereus*. This is an example of the discrepancies in current methods of speciation (Medini *et al.* 2005). By the same token 16S rDNA based species identification uses a single gene snapshot and lacks the resolving power to classify isolates with a highly variable genome such as *S. epidermidis* (Conlan *et al.* 2012). Multilocus sequence typing offers a more global understanding of strain differences within a species. By amplification and sequencing of fragments from a set of seven housekeeping genes, an allelic profile can be assigned that is specific to each clonal group (Enright *et al.* 2000).

This scheme can be used to understand the evolutionary patterns and population structure of species of bacteria. As such it can be more informative than 16S rDNA sequencing. MLST, however is still plagued by significant drawbacks. MLST relies on the core genome of the species, which is often stable within a species, and large percentages can be shared by species within a genus; *S. aureus* and *S. epidermidis* for example share a large proportion of their core genomes. Phenotypes resulting from the sometimes hyper mobile accessory genome, which can be shared through horizontal gene transfer, is not taken into account (Turner & Feil 2007). Pan-genome analyses have highlighted the inadequacies of the previously accepted practice of having a single reference genome sequence to represent a species, since it fails to accommodate the full extent of genetic diversity within a species. MLST typing begins to explore the genetic diversity of a species in more depth than 16S sequencing however, as previously stated, it still only utilises a small fraction of the available genetic variability.

An MLST scheme has been developed for *S. hominis* by Zhang *et al*. This MLST scheme categorised 108 isolates collected over the course of 40 years from 10 different countries into 40 sequence types. This indicated a higher degree of genetic diversity among the *S. hominis* population relative to that of *S. aureus* and *S. epidermidis* (Zhang *et al.* 2013). There are two proposed subspecies of *S. hominis*; *S. hominis* subsp. *hominis* and *S. hominis* subsp. *novobiosepticus*. The phenotypic traits which distinguish the two subspecies are novibiocin resistance and a lack of acid production from anaerobically metabolised D-trehalose and N-acetyl-D-glucosamine (Kloos *et al.* 1998). It is difficult, however, to distinguish the two subspecies from one another through phenotypic and sequence based approaches. Furthermore, Zhang *et al* found that MLST anaylsis of the 108 isolates of *S. hominis* did not reveal a distinct sequence type cluster for the presumed *S. hominis* subsp. *novobiosepticus* isolates. This indicates that this may be an artificial taxon (Zhang *et al.* 2013).

Although the concept of the pan-genome began as a way to refer to the full complement of genes in a single species it has now been extended to encompass

all genes present in a group of species of bacteria in a given niche, for example (Y. Zhang & Sievert 2014). As such the pan-genome, representing the total possible genetic diversity, can help in the understanding of the evolution of species and the divergence within them (Donati *et al.* 2010). The pan-genome can be sub-divided into core genes, present in all (n) isolates of a species, and accessory genes present in ≤n-1 isolates. The core genome describes the essential biological processes of a species whereas, most often, traits such as virulence, antibiotic resistance profiles and niche adaptation can be found in the accessory genome(R. A. Mann *et al.* 2013). Mathematical modelling of the pan-genome using Heaps law can be used to describe a pan-genome as either open or closed. Heaps law was first developed in the linguistics arena and was used to describe the number of words used in a document as a function of the document length, the use of Heaps law in a genomic context was pioneered by Tettelin *et al* (Tettelin *et al.* 2008). An open pan-genome refers to a genetic repertoire that is not fully described by the sequences used in the current analysis, whereby successive genome sequences will contribute additional genes. Moreover, an open pan-genome indicates a high degree of variability within a species, which can be accounted for by a large accessory genome.

3.1.3 Horizontal gene transfer and the accessory genome

The movement of genetic material into and out of bacterial genomes, termed horizontal gene transfer (HGT), plays a major role in the composition of the accessory genome. The process of horizontal gene transfer has the potential to have a negative outcome if genes leaving or entering the chromosome reduce the fitness of a bacterium. In this scenario the bacterium in question will fail to persist in the population and the individual, thus the genetic change will be lost. Neutral effects rely on subsequent events in the population to dictate whether these mutations will be maintained, and HGT that confers increased fitness on a bacterium increases the chance of these individuals persisting in the population and spreading the mutation (C. M. Thomas & Nielsen 2005). The range of *SCCmec* types present in the staphylococci are a prime example of the fitness implications of MGE, with some larger *SCCmec* types conveying reduced fitness

in the community environment compared to a hospital where antibiotic pressures are present (Ma *et al.* 2002).

There are three key mechanisms by which HGT occurs and these vary in their specificity and efficiency. Natural transformation exhibits a low specificity and accounts for the ability of certain bacteria to acquire genes from outside of their own species (Johnsborg *et al.* 2007). It is estimated that around 1% of described bacteria can readily incorporate naked, foreign DNA, including plasmid DNA or mobile genetic elements, and the staphylococci are among this number (Morikawa *et al.* 2012). The HGT process more commonly associated with plasmid DNA uptake is conjugative transfer mediated by cell-to-cell junctions, although other mobile genetic elements such as conjugative transposons are mobilised via this mechanism. Antibiotic resistance genes are frequently transferred by conjugation if the genes involved are found on sections of the chromosome capable of excising themselves and facilitating intercellular relocation (Courvalin 1994). This mechanism has a higher degree of specificity than natural transformation since plasmid transfer systems typically have a defined host range, although the host range of some plasmids will be broader than others (Grohmann *et al.* 2003). HGT via phage transduction similarly exhibits varying degrees of specificity in line with the host range of the infecting bacteriophage. This virus vector is capable of integrating into the bacterial chromosome, leading to transcription of phage cargo genes. Once integrated into the host genome these foreign DNA elements are termed genomic islands. If these genomic islands contribute to pathogenesis of their host they become known as pathogenicity islands (Novick 2003). The enrichment of genes of unknown function in the accessory genome is often a result of the presence of genomic islands of phage origin bearing such genes. Generalised transduction, whereby bacterial DNA is packaged instead of phage DNA in error, contributes further to the spread of genes in the absence of phage genomes (Canchaya *et al.* 2003).

The possibility for horizontal gene transfer between species is reported to be a key factor defining human skin as a niche which promotes adaptive evolution

(Dethlefsen *et al.* 2007). An almost ubiquitous presence of antibiotics in host populations has helped to drive this adaptive evolution in populations of skin resident bacteria through the spread of antibiotic resistance determinants by HGT. This is an important consideration in the *Staphylococcus* genus as the gain of certain genes has been linked to clinically important phenotypes. There are several notable examples of clinically important HGT events in staphylococci. The acquisition of Panton-Valentine leukocidin (PVL) by the USA300 epidemic clone of community acquired *S. aureus* has been associated with the rapid spread of clones able to cause antibiotic resistant skin and soft tissue infection (Tenover & Goering 2009), though more recent debate has questioned the importance of PVL. The differential effects of PVL in mouse and rabbit infection models triggered this debate. In a mose acute pneumonia model PVL was indicated to result in transcriptional changes of cell wall and secreted proteins of *S. aureus* and recruitment of neutrophils to the lungs, inflammation of the parenchyma and other tissue damage. These pathologies were absent in mice infected with the PVL-negative strain (Labandeira-Rey *et al.* 2007). These findings were however contradicted by a further two studies which found no decreased virulence in PVL knockout strains of *S. aureus* (Bubeck Wardenburg *et al.* 2007); (Voyich *et al.* 2006). A further study showed that mouse neutrophils are infact resistant to PVL, and that the actions of the toxin are species specific with both human and rabbit neutrophils showing susceptibility (Löffler *et al.* 2010). A rabbit infection model indicated that PVL may have a 'modest and transient' effect during acute stages of infection (Diep *et al.* 2008). As a result of this controversy greater significance was afforded to the arginine catabolic mobile element (ACME) in the evolution of CA-MRSA (Thurlow *et al.* 2012).

It has been suggested that the significance of ACME in the virulence of CA-MRSA is as a result of the polyamine resistance determinants it carries. Host polyamines have been indicated to play a role in wound healing in a mouse skin and soft tissue infection model. These compounds are also known to be peculiarly toxic to *S. aureus* and are thought to inhibit its proliferation (Joshi *et al.* 2011). Excess host polyamines are produced as a result of the arginine to

ornithine metabolism mediated by the *arc* arginine deiminase genes present on ACME. The presence of *speG* which encodes a polyamine-resistance enzyme counteracts this excess of host polyamines and in turn results in S. aureus resistance to the compound (Thurlow *et al.* 2013).

3.1.4 Genome features associated with niche adaptation

3.1.4.1 Cell wall anchored proteins

Cell wall anchored (CWA) proteins in staphylococci and other Gram-positive bacteria are exposed proteins on the cell surface. Due to their exposed nature they are often found to interact with target ligands on host cells and the extracellular matrix. These interactions can promote adherence and so persistence in a given niche, or evasion of the host immune system. Key features of CWA proteins include an N-terminal secretion signal peptide, a region containing multiple domains which often fold into IgG-like structures and a tandem repeat region. Also required is a wall spanning region that tethers the protein to the cell and an LPXTG-motif required for processing by the sortase transpeptidase enzyme producing covalent linkage to the peptidoglycan wall (Bowden *et al.* 2005).

3.1.4.2 MSCRAMMs

MSCRAMMS are the best-characterised group of staphylococcal adhesin proteins and mediate attachment to the host extracellular matrix. The diversity of tissue specificity across the repertoire of MSCRAMMs and other adhesins within a given genome is proposed to dictate the range of niches the organism can inhabit. It is known that the *S. aureus* genome, for example, encodes 20 MSCRAMMs and is capable of colonising multiple niches, from asymptomatic colonisation of the anterior nares through ClfB-mediated loricrin binding (Mulcahy *et al.* 2012), to post-surgical infections resulting from the colonisation of implanted medical devices and prostheses mediated by the fibronectin binding protein genes *fnbA* and *fnbB* (Lower *et al.* 2011).

The ability to colonise various niches has been implicated in the lifestyle of the staphylococci as opportunistic pathogens, particularly *S. aureus*. Although colonisation of the anterior nares is in itself asymptomatic it is known to be a risk factor for nosocomial infection within particular patient groups. The decolonisation of surgical patients carrying meticillin-resistant strains of *S. aureus* often occurring as a pre-operative measure to control post-surgical infection (Simor 2011).

Coagulase-negative staphylococci are thought to have a reduced repertoire of adhesins in their genome with 12 MSCRAMMS having been characterised in the *S. epidermidis* genome, fewer than the 20 encoded in the *S. aureus* genome. The complement of MSCRAMMs in other coagulase-negative staphylococci is however poorly understood. *S. saprophyticus* is the next best characterised staphylococcal genome in terms of its adhesins and has so far been found to express three genes associated with attachment to host molecules; SdrI (Sakinç *et al.* 2009), UafB (King *et al.* 2011) and Aas (Hell *et al.* 1998). Other staphylococcal genomes found to contain adhesins include *S. lugdunensis* and *S. caprae* (Coates *et al.* 2014). To my knowledge the *S. hominis* genome has not been interrogated for the presence of adhesins.

The presence of a sortase enzyme in the genome can be used as a method for predicting the presence of MSCRAMMS. Sortases catalyse attachment of MSCRAMMS to the cell surface by cleavage between the threonine and glycine residues of the conserved LPXTG motif and amide-linking the protein to the bacterial cell surface during the synthesis of the peptidoglycan cell wall (Mazmanian *et al.* 1999).

3.1.4.3 Other adhesins associated with niche selection

Biofilm production requires the production of an intercellular adhesin to facilitate cell-to-cell contact of bacteria within the biofilm structure. One such adhesin is synthesised upon expression of the *ica* operon, the intercellular adhesion locus, and is a polysaccharide comprised of linear β-1,6-linked glucosaminylglycans known as PNAG or PIA. Expression of *icaA* alone is capable

of producing low activity of the enzyme N-acetylglucosaminyltransferase required to synthesize the polysaccharide. Concomitant expression of the *icaD* gene however results in increased activity of N-acetylglucosaminyltransferase and is thus associated with a biofilm forming phenotype (Arciola & Baldassarri 2001). Due to its association with phenotypic biofilm production and the implications of biofilm production in antibiotic resistant nosocomial infections the *ica* operon has become a commonly used marker for invasiveness in the clinic.

EbpS, a giant, membrane-spanning protein of *S. epidermidis* has also been associated with the accumulation of *S. epidermidis* biofilms (Christner et al. 2010). EbpS and its homologue Ebh in the *S. aureus* genome are non-covalently attached surface proteins which bind fibronectin. The cognate genes are the largest found in the *S. epidermidis* and *S. aureus* genomes andthe encoded megadalton proteins have an extracellular domain, a membrane spanning domain located at the C-terminus and an intracellular, positively charged repeat region (Linke & Goldman 2011) (Heilmann 2011)

The SERAMS or secretable expanded repertoire adhesive molecules comprise a further subgroup of adhesive molecules which includes the protein coagulase (Friedrich *et al.* 2003). The SERAMS are secreted, surface-associated proteins which bind to the bacterial cell in an as yet undefined manner as opposed to the cell wall anchored proteins which interact with the cell wall in a defined and characterised way (Guggenberger *et al.* 2012). These SERAMS, Eap and Emp in particular, are proposed to have a lower binding specificity for the bacterial cell surface than the MSCRAMMs that might relate to their described roles in infection of the blood vessels through interaction with both the extracellular matrix and directly with the surface of endothelial cells. Emp and Eap also facilitate uptake of staphylococci by eukaryotic cells and inhibit phagocytic activity and neutrophil binding to the endothelium (Linke & Goldman 2011) (Heilmann 2011)

3.1.4.4 Biofilm forming capabilities

The formation of a biofilm involves four defined stages, however the precise mechanisms of these four stages are as yet unknown. The first stage is adherence to a surface, which could be either host cells or a synthetic surface such as catheters, artificial joints and other indwelling medical devices. This first step is proposed to occur through non-specific hydrophobic interactions (Vacheethasanee *et al.* 1998). Certain cell surface adhesins are also thought to facilitate these interactions. AtlE, for example, may play a dual role as an adhesin by binding to vitronectin in addition to its described autolytic activity which releases extracellular DNA (eDNA) (Heilmann *et al.* 1997); the presence of eDNA was shown to contribute to biofilm adherence (E. E. Mann *et al.* 2009). The extracellular matrix serum proteins and platelets that coat indwelling medical devices can be bound by other staphylococcal surface proteins and MSCRAMMS, such as SdrG that is known to bind fibrinogen (Sellman *et al.* 2008), and Embp known to bind fibronectin. Studies have shown that Embp is sufficient to establish a functional and significant biofilm (Christner *et al.* 2010). Other uncharacterised LPXTG-motif containing proteins were proposed to play a role in staphylococcal biofilms and may influence adherence (Bowden *et al.* 2005).

Accumulation is the second stage of biofilm formation and can be facilitated through polysaccharides by polysaccharide intercellular adhesin (PIA) encoded by the *icaADBC* locus, or proteinaceous factors, e.g. Bhp, Aap, Embp and their homologues. The *ica* locus is important in initial adherence of cells, cell-to-cell adhesion as well as subsequent accumulation of cells within the biofilm (Fey & Olson 2010). Proteinaceous biofilm accumulation through Aap is mediated by the proteins B domain which dimerises in a zinc-dependent manner with the B domain of a second Aap on a proximal cell (Conrady *et al.* 2008). The corneocyte binding capabilities of the N-terminal A domain of Aap have been said to implicate Aap in skin colonisation (Macintosh *et al.* 2009). The *S. epidermidis* protein Bhp is capable of supporting biofilm accumulation in the absence of PIA (Tormo *et al.* 2005). However *bhp* is only found in less than 45% of *S.*

epidermidis isolates and more rarely on a pathogenicity island of *S. aureus* (Tormo *et al.* 2005).

Maturation of a biofilm is the process by which growing and dividing cells are linked to one another and by which channels form in the layers of cells to allow nutrients to permeate the biofilm. Formation of these channels also leads to the detachment of clusters of cells which can cause biofilm expansion (Otto 2013). A host of transcriptional changes also occur during this stage of the biofilm and cells within the same biofilm can exist in different metabolic states (Rani *et al.* 2007). The arginine deiminase operon was found to be consistently upregulated in both *S. aureus* and *S. epidermidis* biofilms. It is hypothesised that the arginine deiminase operon may aid biofilm maturation by enabling pH homeostasis and yielding ATP (Fey & Olson 2010).

Finally, biofilm detachment occurs. This event can occur piecemeal as a result of the formation of channels in the biofilm, discussed previously, or it can occur as a dramatic event mediated by auto-inducing peptide (AIP). Proteins such as delta-toxin and some phenol soluble modulins (PSMs) may act a surfactants and thus hinder bacterial surface interactions. In the detachment of proteinaceous biofilms the action of metalloproteases and serine proteases are capable of destroying these structures (Boles & Horswill 2008).

The ability of a staphylococcal species to form a biofilm is important due to their clinical significance. Biofilms are known to reduce the antibiotic susceptibility of the infection and can result in persistent infections (Høiby *et al.* 2010). The ability of *S. epidermidis* to form a biofilm is the key factor associated with virulence and isolate invasiveness, with PIA being the major clinical marker for this phenotype. Since other CoNS are more frequently being isolated from infections it is important to understand the process of biofilm formation in these species and determine how it may differ from the better studied *S. aureus* and *S. epidermidis*.

3.1.4.5 Staphylococcal pigmentation

The staphylococcal pigment is a carotenoid, staphyloxanthin, associated with the golden colour that gives *S. aureus* its species name. Bacterial pigments can serve various protective functions across a wide range of species and can act as virulence factors (G. Y. Liu & Nizet 2009). The physiological role of staphyloxanthin is proposed as being required for maximal resistance to reactive oxygen species such as hydrogen peroxide which are associated with survival in the host macrophage (Olivier *et al.* 2009) (Clauditz *et al.* 2006) and increased survival of neutrophil killing (C.-I. Liu *et al.* 2008). It is also known that *S. aureus* strains which lack the *crtOPQMN* operon have reduced ability to form abcesses in a mouse infection model thus implicating staphyloxanthin in the virulence of the species(G. Y. Liu *et al.* 2005). The pigment also has a role in regulating membrane fluidity, which conveys resistance to cationic antimicrobial peptides. The mechanism by which this is achieved is not known as phospholipid composition, surface charge, fatty acid complement or thickness of the cell wall did not appear to change in a *S. aureus* mutant overexpressing crtMN of the *crtOPQMN* operon (N. N. Mishra *et al.* 2011). A study by Grinsted & Lacey found that pigmented strains of *S. aureus* had increased survival in the presence of linoleic acid and under conditions of desiccation (Grinsted & Lacey 1973). A *S. aureus crtM* mutant was also demonstrated to have reduced linoleic acid survival (Kenny *et al.*, 2009). Another study found that variation in pigmentation could not be correlated with survival differences (Noble 1977). The same study posited that the carotenoid pigment could afford pigmented bacteria protection from DNA damage by UV sunlight. There is evidence for bacterial protection against UV DNA damage in *Sporothrix schenckii* (Romero-Martinez *et al.* 2000).

3.1.5 Staphylococcus hominis

S. hominis was first isolated from healthy human skin (Kloos & Schleifer 1975) and is now known to be a consistent member of the human microflora. Investigation of the anterior nares showed 10% of people to be persistently colonised (Faria *et al.* 2014) and that *S. hominis* is the second most abundant staphylococcal species on human skin, making its genome an important part of

the bacterial pan-genome potentially available for horizontal gene transfer in this niche. There is evidence to suggest that *S. hominis* acts as a reservoir for mobile genetic elements, such as the *SCCmec* element which carries the genes involved in meticillin resistance (Faria *et al.* 2014).

Despite its commensal status, *S. hominis*, along with other coagulase negative staphylococci, is an emerging clinical pathogen capable of causing infection in a variety of niches, such as the blood (Bouchami *et al.* 2011), particularly in the presence of long term indwelling medical devices (Weinstein *et al.* 1998), and the urogenital tract (Orrett & Shurland 1998) (John *et al.* 1978) . Furthermore, the presence of multidrug resistance phenotypes in the *S. hominis* population were identified (Mendoza-Olazarán *et al.* 2013) (Jiang *et al.* 2012), which is significant given the increasing rates of nosocomial *S. hominis* infections (Macía-Heras & Donate-Correa 2013) (Chaves *et al.* 2005; Mendoza-Olazarán *et al.* 2013).

Although it is undeniable that *S. hominis* constitutes a significant resident of the human microbiome efforts to understand its role in human health and disease are significantly earlier in their infancy than those of *S. aureus*, *S. epidermidis* and even *S. haemolyticus*. At the time of writing there were around 138 publically available *S. epidermidis* genomes, 13 *S. haemolyticus* genomes but only 5 *S. hominis* genomes with none of these being a complete reference genome (GOLD genomes online database, accessed 28.4.15). As part of this study this deficit of *S. hominis* genome availability has begun to be addressed by sequencing and annotation of 7 recent skin isolates. The addition of these genomes to those already publically available will enhance the knowledge of *S. hominis* as an important species of the human skin niche.

3.1.6 Aims

The specific aims of the research presented in this chapter were to genetically characterise *S. hominis* as a species through interrogation of the entire genome sequences of both clinical and commensal isolates. The data from intra-species comparisons and pan-genome investigations will increase understanding of *S.*

hominis and its little studied genetic repertoire, which is available as a reservoir of genes for other species of bacteria which share the skin niche, such as the transient and pathogenic species *S. aureus*.

3.2 Methods

3.2.1 Assembly of the *S. hominis* pan-genome

In order to understand the genetic relatedness of the group of *S. hominis* isolates used in the pan-genome analysis a tree based on the housekeeping sense *aroE, gmk, pta* and *tpi* was produced. The housekeeping genes were pulled out of all 11 *S. hominis* genomes using Artemis (version 14.0.0) and concatenated together. A multiple sequence alignment was produced using a Geneious alignment and the Blosum62 cost matrix. The Geneious tree builder was then used to produce a consensus tree based on 1000 bootstrap replicates. Geneious version 8.1.7 was used.

Intraspecies analysis of *S. hominis* required the use of all available genomes. Therefore all *S. hominis* isolate sequences from this study and those genomes publically available on the NCBI-FTP website were downloaded for incorporation into the study. Strain IDs and EBI project accession numbers are listed in Table 3.1. All read, contig and annotation information is available within the EBI project.

Table 3.1 List of strains. The ID and EBI project accession number of the strains used in this study, where the genome sequence is available the public

Species	**Strain ID**	**Accession number**
S. hominis	B10	PRJEB10524
S. hominis	C80	UID61127
S. hominis	J6	PRJEB10524
S. hominis	J11	PRJEB10524
S. hominis	J23	PRJEB10524
S. hominis	J27	PRJEB10524
S. hominis	J31	PRJEB10524
S. hominis	I4	PRJEB10524
S. hominis	SK119	UID55861
S. hominis	VCU122	UID180067
S. hominis	ZBW5	UID200270
S. epidermidis	RP62A	UID57663
S. aureus	Newman	UID58839

The contigs from publically available draft genomes were scaffolded against the PacBio assembly of the *S. hominis* J31 genome using the script ACT compare (Table S2). Gene calling and annotation was then carried out using PROKKA version 1.5.2. OrthoMCL version 1.4 was used to cluster the predicted genes from all isolates into orthologous groups; the following parameters were used, e-value cut-off: 1e-5, percentage identity cut-off: 30, percentage match cut off: 20 (L. Li *et al.* 2003).

The number of singleton genes, defined as being present in only 1 genome, shared genes, defined as being present in 2 or more genomes, and the total number of all genes was calculated for each combination of 2 to 11 genomes. This was achieved using the R scripts pan_and_core_genome and new_genes (Table S2).

The size of the pan-genome was estimated using Heaps' law, which is described by the equation $n = \kappa \times N^{\gamma}$. κ and γ are the intercept and slope of the line, respectively, n is the total number of genes and N is the number of genomes. γ is used to solve $\alpha = 1 - \gamma$ to determine if the pan-genome is open or closed. $\alpha \geq 1$ indicates a closed pan-genome and $\alpha < 1$ indicates an open pan-genome. The number of new genes added and the size of the core genome was estimated using an exponential decay model described by $n = \kappa * \exp(-N/\gamma) + tg(\Theta)$, where n is the number of genes, N the number of genomes and $tg(\Theta)$ is the predicted number of new genes added and the size of the core genome, respectively.

3.2.2 Functional annotation of the pan-genome

The web server WebMGA (Wu *et al.* 2011) was used to assign all of the predicted coding genes in all of the isolates to COG (clusters of orthologous genes) categories. WebMGA aligns all genes against the COG database using RPSblast with an e-value cut-off of 0.001. OrthoMCL clusters of core and accessory genes were then annotated with the corresponding COG category if all genes within a cluster were consistently annotated. This was achieved using the perl script *extract_COGID.pl* (Table S2).

3.2.3 Investigation of biofilm forming capabilities

10 ml of BHI broth was inoculated with the *S. hominis* isolates and the cultures were incubated overnight at 37°C with shaking. OD_{600} of these cultures was adjusted to 0.1 in BHI broth supplemented with 0.25% glucose. These cultures were then incubated at 37°C until OD_{600} was 1.0 then 100 μ□of these cultures were used to inoculate the wells of a polystyrene microtiter plate (Nunclon tissue culture-treated) in triplicate. The plates were incubated for 24 h at 37°C without shaking and then washed three times with distilled water. After drying for 1 h at 65°C the wells were stained for 10 min with 0.4% (w/v) crystal violet solution. The plates were again washed three times with distilled water, 100 μl of 33% (w/v) sodium acetate was added to the wells and the absorbance measured at 490 nm.

3.2.4 Investigation of putative cell wall-associated proteins

A modification of the method by Bowden *et al* was used to determine the set of putative cell wall-associated proteins of *S. hominis* (Bowden *et al.* 2005). Firstly, the perl script *LPXTG_capture.pl* (Table S2) was used to search for the extended LPXTG motif in all *S. hominis* genomes. Next, a locally installed stand alone version of PSI-blast, provided by the NCBI, was used to investigate this set of proteins for homology to a set of staphylococcal adhesins defined by Coates et al, 2014. Homology between two sequences is indicated by a low e-value obtained for a PSI-blast alignment. This was achieved using the script *psi-blast.sh* where the adhesin protein sequence was the query, and the *S. hominis* genomes were formatted as the database. Results were then concatenated into a single file for each isolate. The expanded set of proteins was filtered on the criteria that the proteins contained a signal peptide (http://www.predisi.de/) and more than 1 transmembrane domain (http://www.cbs.dtu.dk/services/TMHMM/). LPXTG-motif proteins which fulfilled all three criteria of homology to known staphylococcal adhesins, presence of a signal peptide and presence of transmembrane domains were defined as putative cell wall-associated proteins. LPXTG-motif proteins which did not fulfil the criteria of homology by PSI-blast to known staphylococcal

adhesins were defined as discrete *S. hominis* cell wall-associated proteins. The rationale for this strategy was that *S. hominis* may have CWA proteins without significant homology to CWA proteins in other species, therefore these may not be identified by PSI-blast. Proteins may not fulfil the criteria of having an LPXTG motif or transmembrane domains if they are attached to the cell wall by alternative mechanisms such as interacting with specific domains of cell wall components (Desvaux *et al.* 2006).

3.2.5 Carotenoid Assay

S. hominis isolates, together with *S. aureus* control strains SH1000 and Newman, were grown for 48 h in 5 ml of BHI broth at 37 °C with shaking at 200 rpm. Lids were loosened to allow maximum aeration of the culture. Ten-fold diluted cells were then plated on BHI agar and incubated for 48h at 37C. Plates were then incubated for a further 48h at room temperature to allow maximal pigment expression

Methanol extraction was carried out by first harvesting the bacterial cells from 2 mL of culture followed by resuspension in 200 µL of 100% (v/v) methanol. The methanol suspended cells were then incubated overnight at 37°C without shaking.

3.2.6 Antibiotic resistance gene profiles

The antibiotic resistance gene profiles were investigated using ARDB, the antibiotic resistance database (http://ardb.cbcb.umd.edu/). The antibiotic resistance database is a curated database which collates publically available information regarding antibiotic resistance including mechanisms of action, COG annotation and sequence data (B. Liu & Pop 2009).

3.2.7 Investigation of *S. hominis* plasmids

Putative plasmid encoding contigs that were identified in the PacBio sequencing of *S. hominis* J31 (section 2.3.3) were extracted from the genome assembly and reannotated using PROKKA (Seemann 2014). The circularised sequence of

plasmids was confirmed through the observation of duplicated genes at the beginning and ends of the plasmid contigs.

The conservation of plasmid genes across the isolates was investigated by clustering the coding sequences found on the plasmid contigs with the coding sequences of the *S. hominis* genomes. Orthologous groups containing plasmid genes were extracted from the OrthoMCL output and a presence/absence matrix of these genes across the *S. hominis* isolates was created using the python script *table_alt_cluster-1.py* (Table SI). The percentage of clusters from each plasmid present in each of the genomes was then calculated.

3.2.8 Insertion sequences

Transposable elements in each genome were identified and annotated using ISFinder; this is an annotated database of insertion sequences which provides an online tool to allow the user to BLAST search sequences of interest against this database to identify possible insertion sequences (Siguier *et al.* 2006).

3.2.9 Bacteriophage annotation

Bacteriophages were annotated using PHAST, which searches a genome for phage-like genes with homology to a phage/prophage protein database and uses the DBSCAN algorithm to determine the clustering of these genes in the database. The density of phage-related proteins across these clusters is then identified together with the presence of essential phage protein families. Performance of a cluster using these criteria generates a phage completeness score out of 150. Less than 60 results in a cluster being designated as an incomplete prophage, 60-90 is a questionable prophage, while above a score of 90 represents a likely complete prophage (Y. Zhou *et al.* 2011).

3.2.10 Genome manipulation and interrogation

Genome manipulation and interrogation was carried out using a locally installed version of Artemis (version 14.0.0). The Artemis tool was developed by The

Sanger Institute to enable visualisation of sequencing data (Rutherford *et al.* 2000) (Carver *et al.* 2012).

3.2.11 Protein sequence aligments

Multiple sequence alignments were carried out using ClustalW2, which is an open source program hosted by the European Bioinformatics Institute. Average percentage sequence identity was calculated for the alignments. Sequence identity refers to the amount of exact character matches between two sequences. Gaps are not considered and the score is relative to the shortest sequence.

BLAST (basic local alignment search tool) is a tool used to find pairwise local sequence similarity. Similarity is the degree to which two sequences resemble one another. This measure takes into account identity, chemical properties of the amino acids and the number of insertions, deletions and substitutions (mutations) that occur between sequences (Altschul *et al.* 1990). Multiple algorithms (BLASTN, BLASTP, psi-BLAST) were used in this analysis in both web based and locally installed forms.

3.3 Results and discussion

3.3.1 The *S. hominis* pan-genome

A pan-genome analysis is influenced by the genomes included. For this reason the diversity of the isolates used in the *S. hominis* analysis presented here was investigated using multiple alignment of the housekeeping genes *aroE, gmk, pta* and *tpi*. Isolates B10, I4 and the "J" isolates were found to be closely related at these housekeeping gene loci branching together in a neighbour joining tree. (Figure 3.0) This close relationship of 7 of the 11 isolates used in the pan-genome analysis could result in a lower estimated repertoire of genes for *S. hominis.*

To identify the *S. hominis* pan-genome and its genetic repertoire, the number of genes present in all available strains was quantified and plotted as a function of the number of genomes added for all possible combinations of 2-11 genomes. As genomes were sequentially added to the analysis, the size of the *S. hominis* pan-genome increased until it reached ~3115 genes at 11 genomes (Figure 3.1, panel A). This is an estimated number of genes due to the presence of paralogues in 1% of the clusters from the orthoMCL analysis. Heaps' law, a sub-linear power law of frequency distribution, was used to perform a regression analysis. This statistical process is described by the equation $n = \kappa \times N^{\gamma}$, where κ and γ describe the intercept and slope, respectively, n is the total number of genes and N the number of genomes. $\alpha = 1 - \gamma$ was then used to determine if the pan-genome is open or closed, where $\alpha \geq 1$ indicates a closed pan-genome and $\alpha < 1$ indicates an open pan-genome. An open pan-genome will increase in size with the addition of new genomes. For *S. hominis* Heaps' law predicts $\kappa = 7.661506$, $\gamma = 0.125405$ therefore $\alpha = 0.874595$ indicating an open pan-genome with respect to the number of input genomes studied (Figure 3.1, panel A).

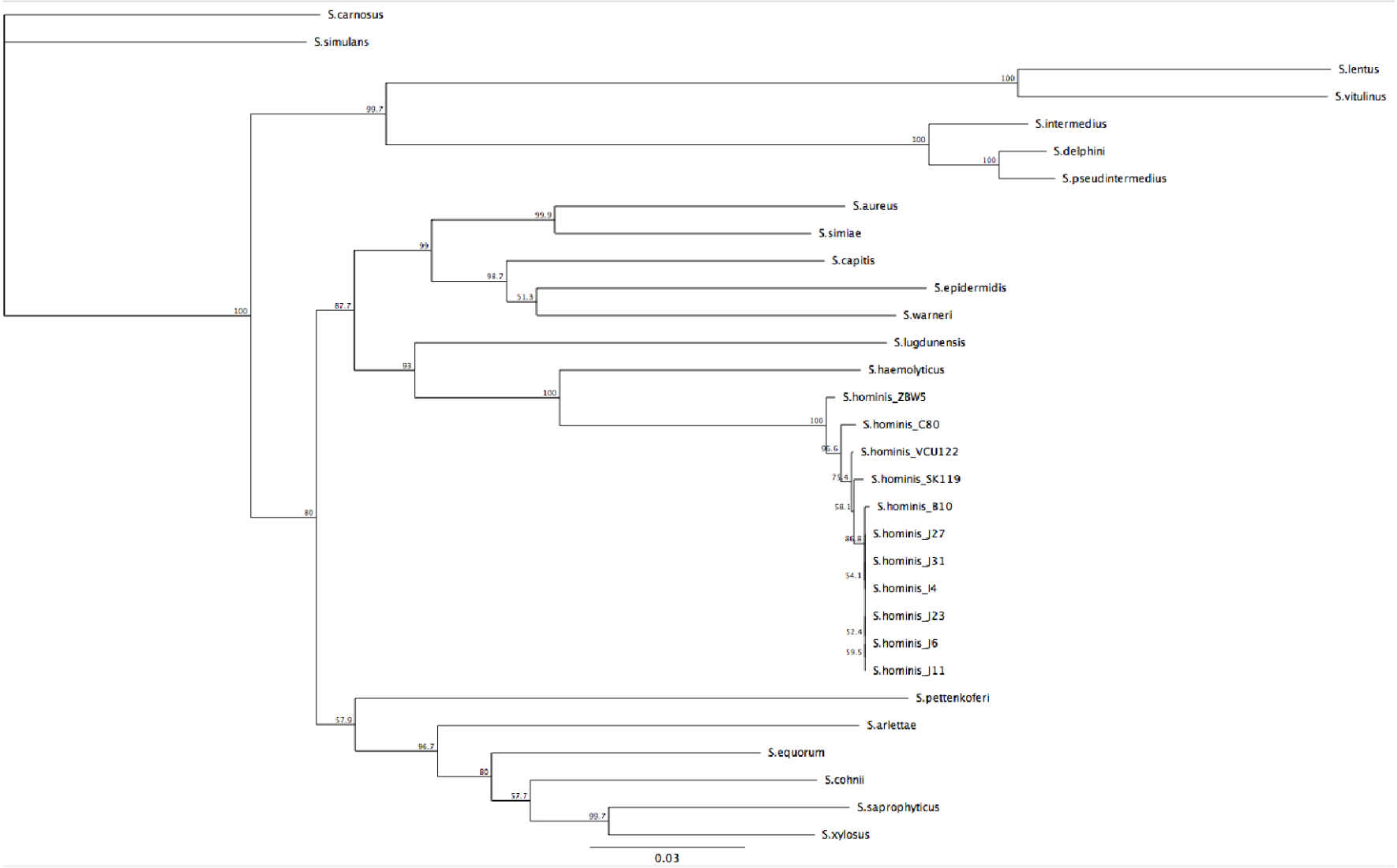
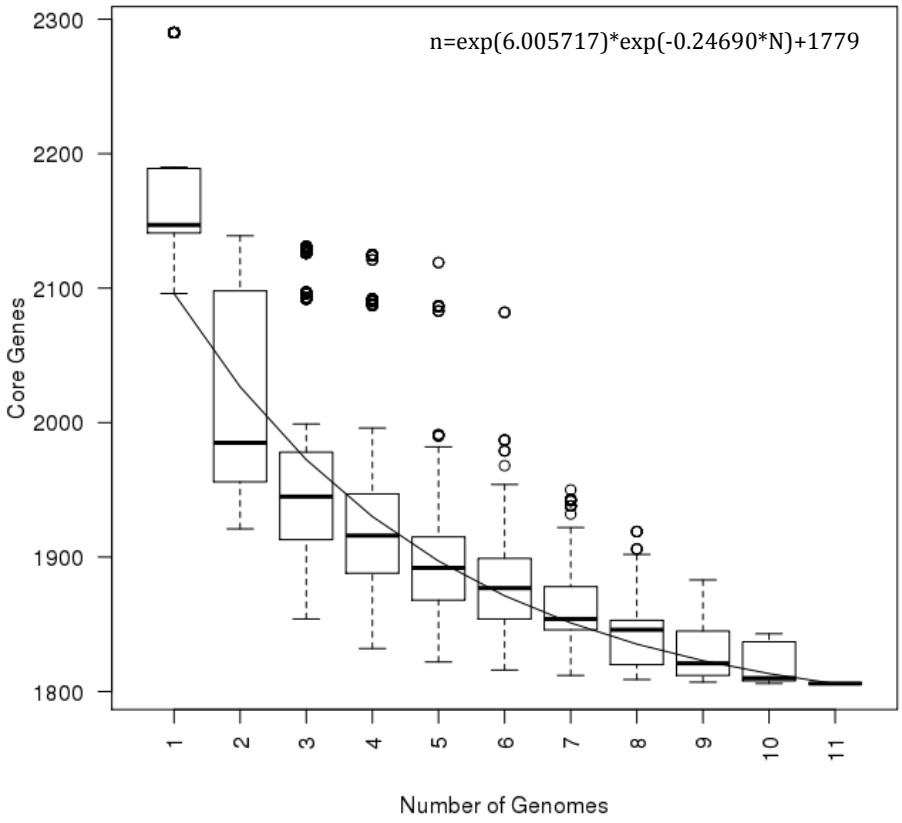
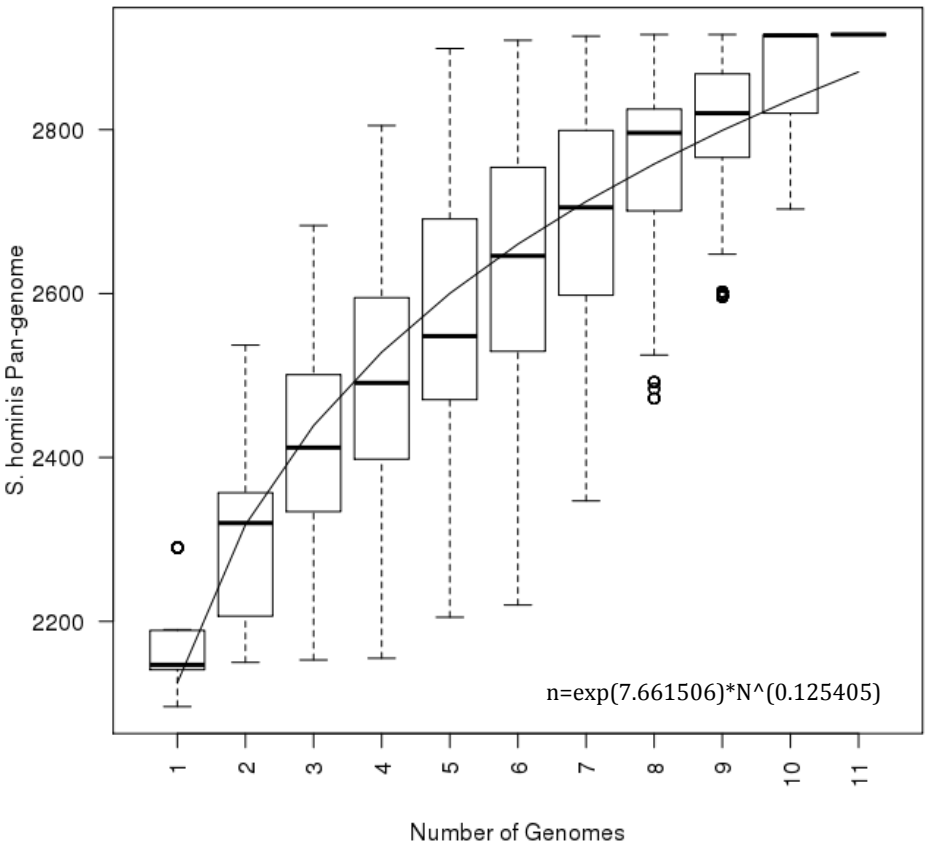


Figure 3.0 Neighbour joining tree of housekeeping genes. The basis of the tree is the concatenated protein sequences of the housekeeping genes *aroE, gmk, pta* and *tpi.* Bootstrapping using 1000 repetitions was carried out.

A


n=exp(6.005717)*exp(-0.24690*N)+1779
Core Genes
2300
2200
2100
2000
1900
1800
1
2
3
4
5
6
7
8
9
10
11
Number of Genomes

B


n=exp(7.661506)*N^(0.125405)
S. hominis Pan-genome
2800
2600
2400
2200
1
2
3
4
5
6
7
8
9
10
11
Number of Genomes

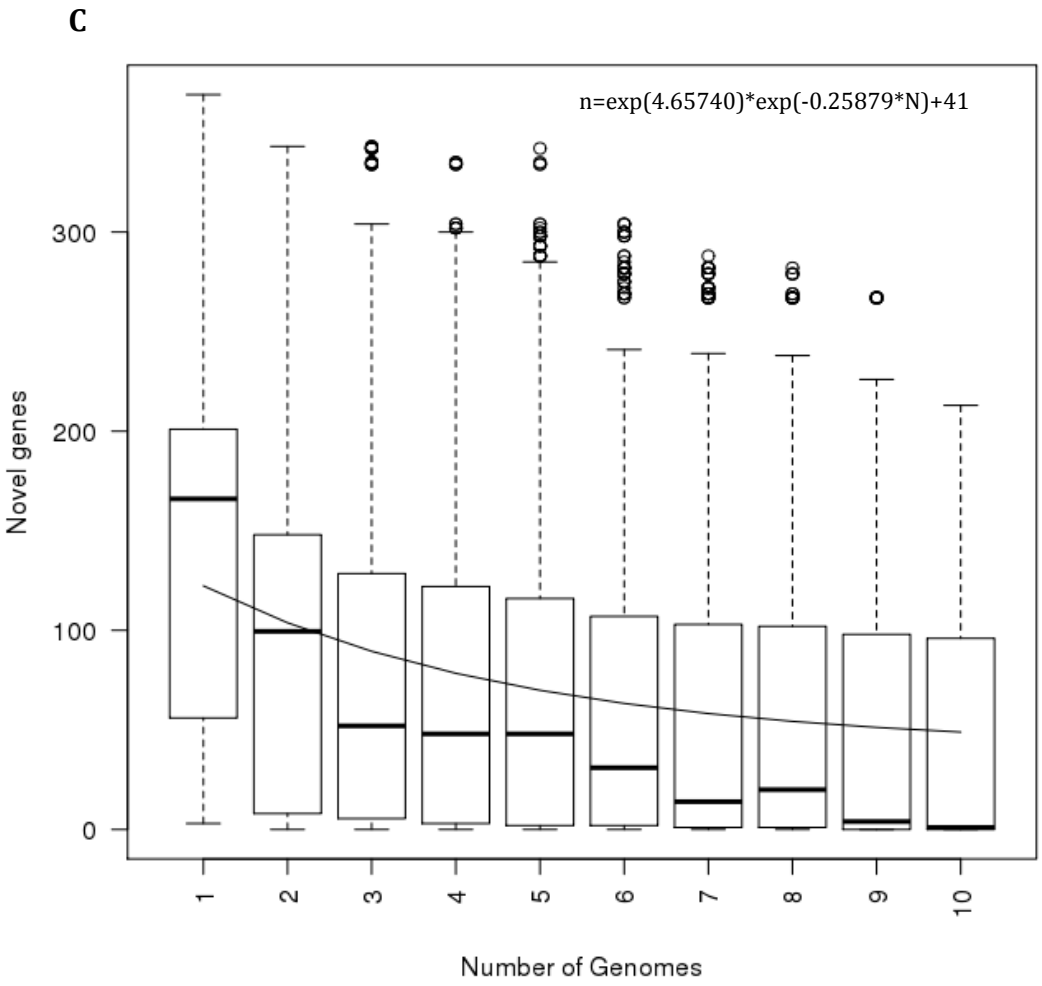


Figure 3.1 Core genome, pan genome and new genes added plots for *S. hominis*. A) Core genome plot represents the n number of shared genes among N number of genomes. The equation of the line, fitted to the median of data, is the exponential decay model n = ! * exp(-N/'') + tg(#) B)Pan genome plot represents the total number all genes (n) in the (N) number of genomes. The equation of the line, fitted to the median of the data, is a modified non-linear least squares known as Heaps' law with the formula n = ! x N^''. C) New genes added represents the number of new genes added (n) per new genome (N). The equation of the line, fitted to the median of data, is the exponential decay model n = ! * exp(-N/'') + tg(#).

3.3.2 Growth of the *S. hominis* pan-genome

The number of new genes added to the pan-genome with the addition of every new genome was investigated for all possible combinations of 2-11 genomes. A non-linear least squares model was fitted to the data using an exponential decay model, which is defined by $n = \kappa * \exp(-N/\gamma) + tg(\Theta)$, where n is the number of genes, N the number of genomes and $tg(\Theta)$ is the predicted number of new genes accrued with the addition of each new genome. For this analysis $tg(\Theta) = 41$, predicting that 41 new genes were added to the pan-genome with each additional *S. hominis* genome. This agrees with the prediction that the pan-genome is open and supports that for the collection of isolates in this analysis the pan-genome is approaching closure, with a value of $\alpha = 0.874595$. This α value agrees with previous studies containing similar numbers of strains of other species of bacteria (Tomida *et al.* 2013; Conlan *et al.* 2012). Soares *et al* in particular found an α value of 0.89 for a pan-genome analysis completed with 15 strains of *Corynebacterium pseudotuberculosis* (Soares *et al.* 2013). Despite this having 7 isolates that appear closely related at housekeeping gene loci may reduce the diversity of the strain set investigated here and so indicate a higher α value which is not representative of the species.

3.3.3 The *S. hominis* core genome

The pan-genome of a species comprises both the core genome, containing genes required for essential biological and metabolic processes, and the accessory genome representing genes for specific adaptations. The size of the *S. hominis* core genome was determined by calculating the number of genes shared by all strains. This value was then plotted as a function of the number of genomes (Figure 3.1, panel B). A non-linear least squares model was fitted to these data using an exponential decay model, which is defined by $n = \kappa * \exp(-N/\gamma) + tg(\Theta)$ where n is the number of genes, N the number of genomes and $tg(\Theta)$ is the predicted size of the core genome. Analysis identified that the predicted size of the *S. hominis* core genome is 1770 genes, which is equivalent to the asymptote of the regression line (Figure 3.1, panel B). At 11 genomes the size of the core genome is 1879 genes, or 60% of the total pan-genome, which is approaching the predicted total size.

The discrepancy between the predicted core genome size and the core genome size for 11 strains indicates that sequencing of additional isolates of *S. hominis* from diverse sequence types or ecological and geographical locations will be required to fully define the core *S. hominis* genome. For individual strains the core genome comprises an average of 83% of the total genome, 17% is therefore comprised of accessory genes. These genome proportions of individual strains, attributable to the core and accessory genomes, agrees with a previous study of staphylococci (Conlan *et al.* 2012). These values indicate that the pan-genome overestimates the proportion of genes in the *S. hominis* genome contributed from the accessory genome. Within individual strains it is clear that the majority of the genome is comprised of essential core genes, with only a small percentage being drawn from that variable accessory genome. This relatively larger pool of non-essential accessory genes in the pan-genome speaks to the variability present between different strains.

3.3.4 The *S. hominis* accessory genome

The accessory portion of the genome can be further subdivided into shared accessory genes present in 2 or more, but not all strains, and strain-specific singleton genes present in only one strain. The *S. hominis* singleton accessory genome was found to contain 541 genes across the 11 genomes, whereas the shared accessory genome contained 659 genes. The number of strain specific genes present can be used as a measure of the divergence between the strains present in the analysis. As the number of singletons is not normally distributed across the dataset it can be stated that particular strains are more divergent than others (Figure 3.2). For example, the drug resistant isolate ZBW5 contains 208 singletons, whereas all "J" isolates plus strain B10 contain no singletons. It was postulated that the accessory genome could contain genes designated as sub-core genes for strains isolated from different niches, and that these sub-core genes might contribute to niche adaptation and survival (Gupta *et al.* 2015). It is evident that additional *S. hominis* isolates from different environments must be added to the pan-genome analysis to enable conclusions

to be drawn regarding niche specific adaptations.

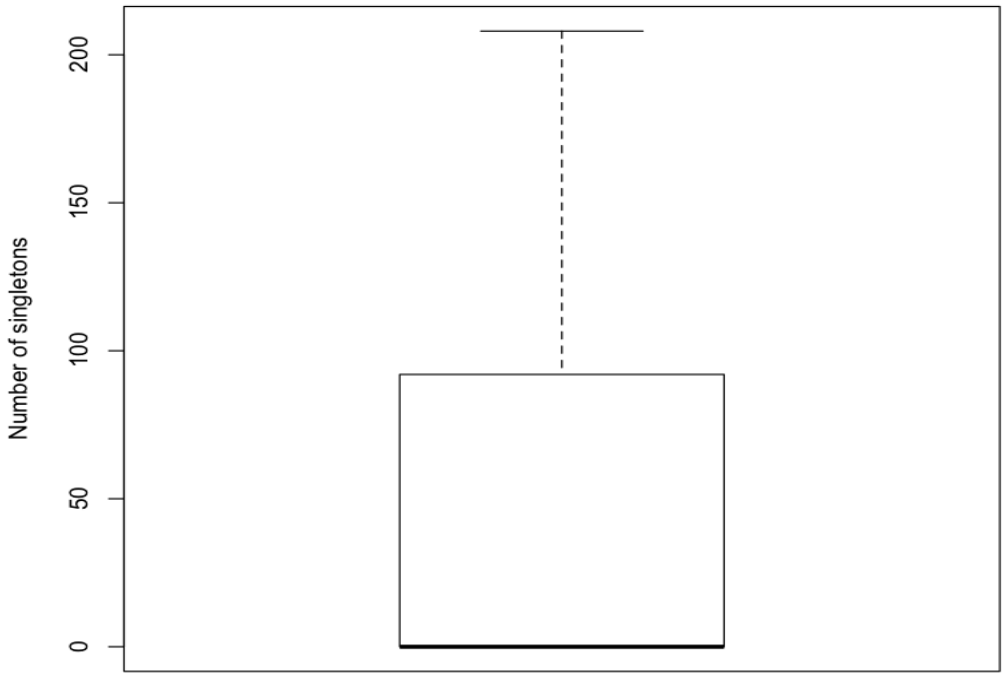


Figure 3.2 Distribution of strain specific genes assigned by orthoMCL analysis of the *S. hominis* genomes. Whiskers represent the highest and lowest values, the perimeters of the box show the 1st and 3rd quartile range while the thick black line shows the median.

3.3.5 Functional diversity of the core and accessory genomes.

The functions of genes within the core and accessory genomes of *S. hominis* were investigated by assigning all gene clusters discerned by OrthoMCL analysis to clusters of orthologous groups (COGs).

The majority of genes in the accessory genome (~64%) could be assigned to the COG categories general function prediction only, function unknown or unassigned function. This group of genes was enriched in comparison to the core genome where only 33% of genes were similarly assigned (Figure 3.3). This indicated the presence of novel gene clusters in the *S. hominis* genome.

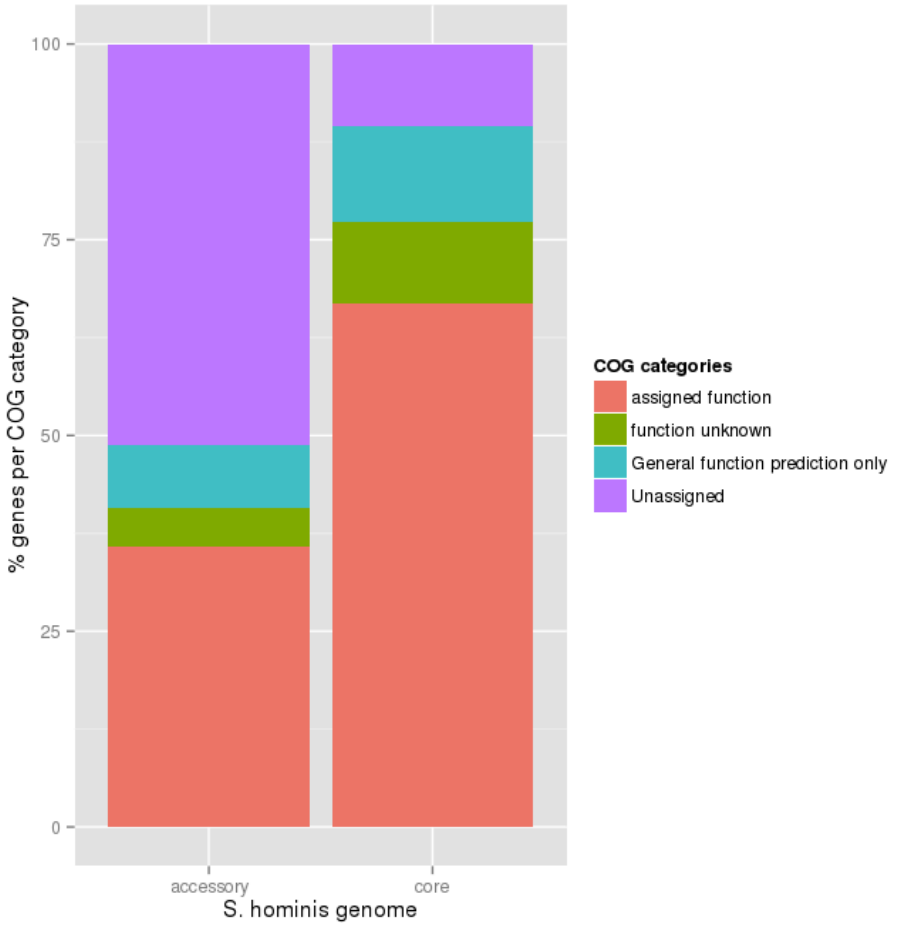


Figure 3.3 Functional annotation of the core and accessory genome. Percentages of the *S. hominis* accessory and core genomes annotated as either function unknown, general function prediction only or unassigned function, as well as the percentage of genes annotated with an assigned COG function.

Genes associated with both of the categories, defense mechanisms and replication, recombination and repair were found to be enriched in the accessory genome (Figure 3.4), which agrees with analyses of other species of staphylococci (Conlan *et al.* 2012). Despite the enrichment in the replication, recombination and repair COG category there is limited diversity in this group of genes. Approximately 30% of these gene clusters are involved in the integration and excision of mobile genetic elements from the genome. A further 29% of the gene clusters in the recombination, replication and repair group encode recombinases. It has been suggested that movement of genes on mobile genetic elements and through genetic recombination are important in the diversity of staphylococcal phenotypes.

Around 58% of the genes in the enriched defence mechanism COG category of the accessory genome were associated with drug resistance. This identifies that 2% of the accessory genome accounts for COG groups associated with virulence in species that are designated as pathogenic species. This relatively low percentage of virulence-associated genes is in accordance with the lifestyle of *S. hominis* being predominantly commensal, however their presence also corroborates the potential of *S. hominis* as an opportunistic pathogen.

All other COG clusters were enriched in the core genome potentially indicating greater functional diversity in the core genome compared with the accessory portion. However, 64% of the genes in the accessory genome were unassigned or poorly assigned to COG categories, and it is likely that these unassigned clusters account for the apparent drop in functional diversity between the core and accessory genome.

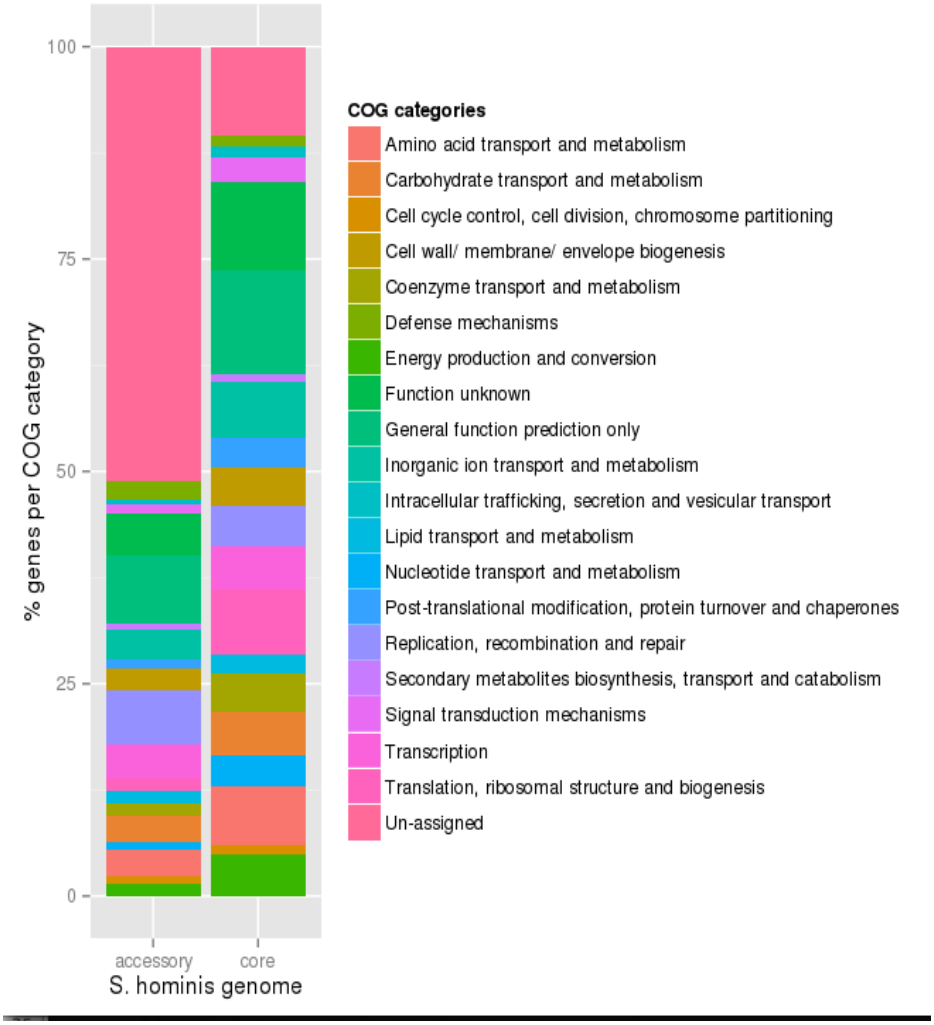


Figure 3.4 COG assignment of the core and accessory genome. Percentage of genes in each of the COG categories in the *S. hominis* accessory genome and core genome.

COG annotation of the *S. hominis* core genome reveals that the majority of the genes are associated with housekeeping functions including essential processes of metabolism and cell replication. The most abundant COG category in the core genome is translation, ribosomal structure and biogenesis in contrast to the accessory genome where replication, recombination and repair corresponds to the most abundant COG category.

3.3.6 Crt operon

In *S. aureus* the *crtOPQMN* operon encodes the five enzymes known to be necessary and sufficient in the biosynthesis of the triterpenoid carotenoid staphyloxanthin (Pelz *et al.* 2005). The *crtOPQMN* operon is regulated by a σ^B dependent promoter (Morikawa *et al.* 2001). Production of staphyloxanthin is considered to be a virulence factor in *S. aureus* due to the association between the pigment, antioxidant defence and establishment of infecton (C.-I. Liu *et al.* 2008).

8 out of the 11 isolates of *S. hominis* examined in this study contain the *crtPQMN* genes of the staphyloxanthin carotenoid operon. In isolates where the *crtPQMN* genes are present the protein sequences were conserved across all isolates with greater than 99% sequence identity. This agrees with the level of sequence identity observed in the *crtPQMN* operon across different *S. aureus* genomes. This suggests, within the subset of *S.* hominis isolates investigated here, previous findings that the gene sequences of the *crt* operon are conserved across a species with respect to gene synteny and genomic location (Pelz *et al.* 2005). Two isolates, B10 and VCU122, contained only a fragment of *crtN*; this fragment also has >99% sequence identity with *crtN* of other *S. hominis* isolates. Uniquely, the multidrug resistant isolate ZBW5, did not contain an identifiable *crt* operon or any gene fragments.

The sequence identity between the CrtPQMN proteins present in *S. hominis* and *S. aureus* was investigated. CrtN and CrtP were the most highly conserved between the two species sharing 69.5% and 70.3% amino acid sequence

identity, respectively. CrtM and CrtQ shared 51.4 and 49.9% identity, respectively.

Culture of *S. hominis* strains revealed that the isolates of *S. hominis* containing *crtPQMN* could produce pigment suggesting the operon may be expressed in these isolates (Figure 3.6). This pigment cannot, however, be extracted using a standard protocol for staphyloxanthin extraction (Morikawa *et al.* 2001). This is consistent with the fact that the *crtO* gene is not present in any of the *S. hominis* genomes. CrtO is an acyltransferase which esterifies glucose to form glycosyl-4,4′-diaponeurosporenoate, the last step in the staphyloxanthin biosynthesis pathway (Pelz *et al.* 2005). Lack of *crtO* results in the accumulation of the yellow pigment glycosyl-4,4′-diaponeurosporenoate without the development of the orange colour associated with staphyloxanthin. CrtO was found to have no sequence similarity to any as yet characterised carotenoid biosynthesis genes (Pelz *et al.* 2005). Perhaps the yellow pigment glycosyl-4,4′-

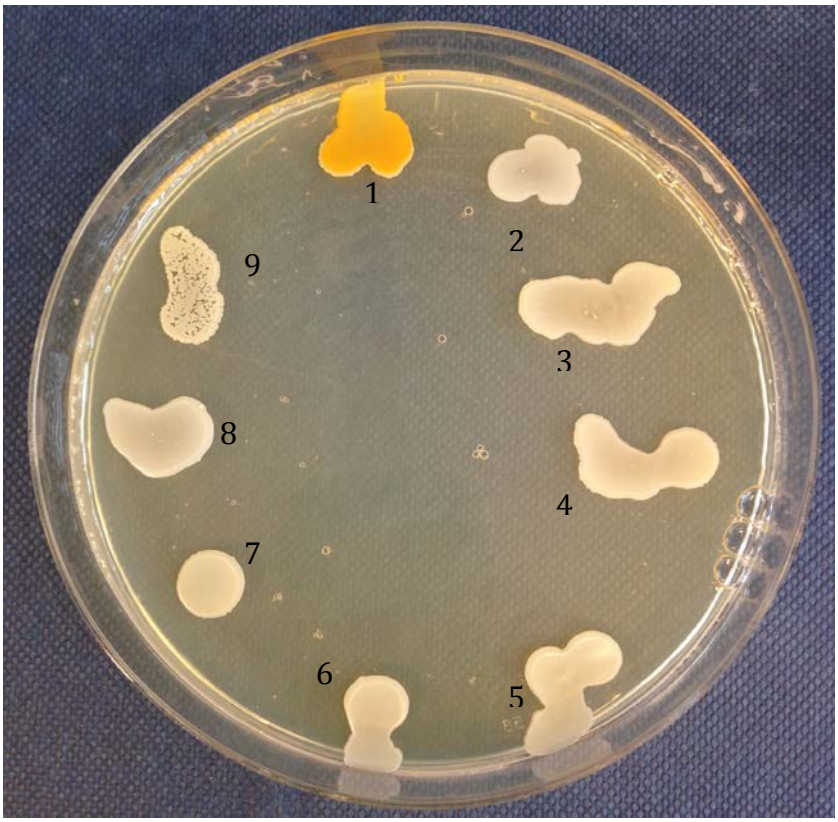


Figure 3.6 expression of the *crtOPQMN* operon in *S. hominis* 1) *S. aureus* SH1000 positive control, known to produce staphyloxanthin 2) *S. hominis* B10 3) *S. hominis* J6 4) *S. hominis* J11 5) *S. hominis* J11 6) *S. hominis* J23 7) *S. hominis* J27 8) *S. hominis* J31 9) *S. hominis* I4

diaponeurosporenoate of *S. hominis* fulfils the role of membrane stabilisation in the face of antimicrobial peptides and fatty acids which will be encountered by commensal bacteria such as *S. hominis* on the skin. The terminal step in the biosynthetic pathway for staphyloxanthin may have evolved as a result of the *S. aureus* lifecycle as a more virulent opportunist pathogen, compared with the less virulent commensal staphylococci with respect to oxidative stress resistance or stabilising the cell membrane. Further investigation is required to understand 1) the membrane function of glycosyl-4,4′-diaponeurosporenoate and how it is localised and 2) the degree of protection offered by this molecule for bacterial fatty acid and antimicrobial peptide survival. Heterologous expression of *crtO* in *S. hominis* would also provide insights via potential gain of function.

3.3.7 Sortases

The LPXTG motif is conserved in MSCRAMMs across the staphylococci and it acts to both direct cell wall sorting and promote attachment of MSCRAMMs to cell wall peptidoglycan catalysed by the enzyme sortase (SrtA) (Navarre & Schneewind 1999). The presence of a sortase enzyme indicates the presence of cognate proteins containing the LPXTG motif. A *srtA* gene is present in all 11 *S. hominis* isolate genomes which indicates that this mechanism of covalently attaching protein to the cell surface is also conserved in this species.

A multiple sequence alignment of the *S. hominis* SrtA protein sequence showed SrtA to be conserved among the *S. hominis* isolates with an overall average identity of >99% across the 11 *S. hominis* genomes. There were 9 amino acid changes between the isolate ZBW5 and the rest of the isolates, and 1 amino acid change between B10 and the other 10 *S. hominis* isolates. Average identity to *S. epidermidis* UniProt SrtA protein Q5HL98_STAEQ is 82%, and identity beween the *S. hominis* SrtA and the *S. aureus* UniProt SrtA protein Q9S446_STAAU is 85%. *S. aureus* and *S. epidermidis* have a distance of 1.507 showing that both the *S. epidermidis* SrtA protein sequence and the *S. aureus* SrtA protein sequence have a higher identity to that of *S. hominis* SrtA than to each other. This is

supported with Blast comparisons which show *S. epidermidis* SrtA has 72% identity with *S. aureus* whereas *S. hominis* shares 76% identity.

Table 3.2 Putative *S. hominis* cell wall associated proteins. Homologous genes are grouped together and the *S. hominis* surface protein nomenclature is shown in bold. The closest match by blast is the sequence with the highest blast score after searching the cell wall associated protein against the BLAST database. Average identity is the average protein sequence identity between the protein sequences from the isolates which contain the cell wall associated protein.

Protein ID	Prokka annotation	Closest match by BLAST	Average identity
ShsA			
B10-PB_00037	Trifunctional nucleotide phosphoesterase protein YfkN precursor	5'-nucleosidase, *S. hominis*	99.7
C80-PB_00043			
J6-PB_00037			
J11-PB_00037			
J23-PB_00036			
J27-PB_00038			
J31-PB_00038			
I4-PB_00037			
SK119-PB_00038			
VCU122-PB_00035			
ShsB			
B10-PB_00206	Lowaffinity inorganic phosphate transporter 1, PitA	Inorganic phosphate transporter, *S. hominis*	99.9
C80-PB_00156			
J6-PB_00133			
J11-PB_00133			
J23-PB_00131			
J27-PB_00133			
J31-PB_00133			
I4-PB_00132			
SK119-PB_00139			
VCU122-PB_00138			
ZBW5-PB_01046			
ShsC			
B10-PB_01086	Hypothetical protein	Hypothetical protein, predicted metal-dependent hydrolase domain	100%
J6-PB_01008			
J11-PB_01009			
J23-PB_01005			
J27-PB_01007			
J31-PB_01009			
I4-PB_01006			
SK110-PB_00993			
VCU122-PB_01012			
ShsD			
B10-PB_01322	Rod shape-determining protein RodA	Cell division protein FtsW	99.7
C80-PB_01272			
J6-PB_01237			
J11-PB_01266			
J23-PB_01178			
J27-PB_01265			
J31-PB_01270			
I4_PB_01244			
SK119-PB_01224			
VCU122-PB_01252			
ZBW5-PB_01274			
ShsE			

B10-PB_01471	Swarming motility protein SwrC	Multidrug transporter	99.2
C80-PB_01425			
J6-PB_01407			
J11-PB_01410			
J23-PB_01406			
J27-PB_01409			
J31-PB_01415			
I4-PB_01413			
SK119-PB_01393			
VCU122-PB_01422			
ZBW5-PB_01453			
ShsF			
B10-PB_01722	EIICBA Glc	PTS glucose transporter subunit IIABC	99.8
C80-PB_01681			
J6-PB_01643			
J11-PB_01644			
J23-PB_01639			
J27-PB_01644			
J31-PB_01651			
I4-PB_01647			
SK119-PB_01646			
VCU122-PB_01676			
ZBW5-PB_01708			
ShsG			
J6-PB_01850	Lipase precursor	Hypothetical protein, *S. haemolyticus*	100
J11-PB_01850			
J23-PB_01846			
J27-PB_01851			
J31-PB_01857			
I4-PB_01854			
ShsH			
B10-PB_02302	Hypothetical protein	Staphylococcal peptidase	100
J6-PB_02165			
J11-PB_02163			
J23-PB_02165			
J27-PB_02169			
J31-PB_02177			
I4-PB_02159			
ZBW5-PB_02260			
ShsI			
B10-PB_00702	Glyceroaquaporin	Glycerol transporter, *S. hominis*	99.6
C80-PB_00658			
J6-PB_00629*			
J11-PB_00630			
J23-PB_00627			
J27-PB_00629			
J31-PB_00629			
I4-PB_00629*			
SK119-PB_00616			
VCU122-PB_00625			
ZBW5-PB_00669*			
ShsJ			
B10-PB_02072	Citrate/succinate antiporter	Staphylococcal membrane protein	99.2
C80-PB_01983*			
J6-PB_01934*			
J11-PB_01934*			
J23-PB_01929*			

J27-PB_01936*			
J31-PB_01942*			
I4-PB_01933*			
SK119-PB_01911*			
VCU122-PB_01960*			
ZBW5-PB_01969			
ShsK			
B10-PB_01223*	Sensor protein VraS	Staphylococcal sensor histidine kinase	99.9
C80-PB_01173*			
J6-PB_01140*			
J11-PB_01169*			
J23-PB_01081*			
J27-PB_01168*			
J31-PB_01173*			
I4-PB_-01147*			
SK119-PB_01126*			
VCU122-PB_01151*			
ZBW5-PB_01176			
***absent from analysis but present on interrogation of the genome**			
Singleton LPXTG-containing proteins			
Protein ID	**Putative surface protein**	**Prokka annotation**	**Closest match by BLAST**
B10-PB_01770	ShsL	Hypothetical protein	Cell wall surface anchor protein
ZBW5-PB_01352	ShsM	Hypothetical protein	Staphylococcal hemolysin III
ZBW5-PB_01982	ShsN	Inner membrane transport permease YbhR	Antibiotic ABC transporter permease

3.3.8 Putative cell wall associated proteins

To investigate putative *S. hominis* MSCRAMMS and adhesins all possible cell wall-associated (CWA) proteins encoded in the *S. hominis* genomes were first located. *In silico* analysis predicted 14 proteins containing the LPXTG motif, an N terminal signal peptide, a hydrophobic transmembrane domain and a relationship with known staphylococcal MSCRAMMS determined through PSI-blast (Table 3.2). These criteria represent the key features of CWA proteins. The number of LPXTG motif-containing proteins found in *S. hominis* is close to the 11 predicted in the *S. epidermidis* genome (Bowden *et al.* 2005). Although more LPXTG proteins were predicted in *S. hominis* than in *S. epidermidis* this may be explained by the fact that 11 isolates were used in this analysis whereas only isolate RP62A was used in the analysis of *S. epidermidis* CWA proteins.

Of the 14 LPXTG motif-containing proteins identified, 7 were conserved across the 11 *S. hominis* isolates with greater than 99 % identity (Table 3.2). The rest of the predicted proteins were conserved among subsets of the 11 isolates. Conservation of a core group of MSCRAMMS among all isolates and distribution of an accessory set of MSCRAMMs was also seen by McCarthy and Lindsay when the distribution of surface proteins among a group of *S. aureus* isolates was investigated (McCarthy & Lindsay 2010). 3 of the proteins were identified in only one isolate and designated as singleton LPXTG-containing proteins; The multidrug resistant isolate ZBW5 showed the most divergence in its repertoire of LPXTG-motif containing proteins sharing 8 of its 10 LPXTG-motif containing proteins with two or more other isolates. It does not however display an expanded repertoire of putative CWA proteins as it contains 10 in line with the rest of the isolates which also contain between 8 and 11 proteins each. Since no *S. hominis* surface proteins have been identified in the literature the putative surface proteins proposed

here have been designated ShsA-N according to the nomenclature proposed by (Mazmanian *et al.* 2001) and will be referred to as such from here onwards.

3.3.9.1 ShsA

5'- nucleosidases have been identified as substrates for the sortase enzyme through investigation of LPXTG-motif containing proteins in the Gram-positive species *S. aureus*, *Streptococcus pyogenes*, *Bacillus anthracis*, *B. subtilis*, *Clostridium difficile* and *C. acetobutylicum* (Pallen *et al.* 2001). The *S. aureus* surface protein SasH has 32% homology with a 5' nucleotidase from *Bacillus halodurans* (Roche, Massey, *et al.* 2003) and so has been designated as a putative 5' nucleotidase. It has also been found to be associated with invasive disease (Roche, Massey, *et al.* 2003). Across all isolates found to contain ShsA the putative *S. hominis* CWA protein shares an average of 47 % identity with *S. aureus* SasH, which has been renamed AdsA due to the protein function to synthesise adenosine from 5'-AMP. This synthesis is mediated by the 5'-nucleosidase activity of the protein and is necessary for *S. aureus* survival in blood, reducing neutrophil killing of *S. aureus* in the bloodstream. As a result of these functions, AdsA was designated as a virulence factor (Thammavongsa *et*

al. 2009) (Thammavongsa *et al.* 2011). Thammavongsa *et al* found an AdsA homologue in the pathogen *Bacillus anthracis*; sequence identity between B. *anthracis* adenosine synthase (accession no. BAS4031) and the USA300 AdsA/SasH protein was 15 %. Identity of the ShsA protein across the studied *S. hominis* species was 16%. There are differences in the length of the *S. hominis* protein (958 amino acids), and the *S. aureus* and *B. anthracis* proteins (786 and 790 amino acids, respectively). Thammavongsa *et al* also identified the presence of 5'-nucleosidases with LPXTG motifs in other Gram-positive species (Thammavongsa *et al.* 2009), and still others secrete 5'-nucleosidases (Punj *et al.* 2000).

During genome annotation ShsA was annotated by Prokka as a trifunctional nucleotide phosphoesterase protein (YfkN) precursor. YfkN is secreted by *B. subtilis* and has 5' nucleotidase activity together with 2', 3' cyclic nucleotide phosphodiesterase and 2' nucleotidase activities (Chambert *et al.* 2003). YfkN has a large reading frame (4386 nucleotides) that is proposed to have arisen from a gene fusion event between genes encoding the 2', 3' cyclic nucleotide phosphodiesterase activity of the enzyme, with another possessing 5' nucleotidase activity (Yamamoto *et al.* 1996). The putative *S. hominis* CWA protein ShsA is comprised of 962 amino acids and so falls far short of the expected length of YfkN. A region of 522 residues at the C-terminus of YfkN shares 26.8% identity with a 5' nucleotidase precursor of *Homo sapiens* (Yamamoto *et al.* 1996). This agrees with an average 22.6% sequence identity between *S. hominis* ShsA and a 574 residue fragment of the *Homo sapiens* 5' nucleotidase.

Taken together, the identity between *S. aureus* AdsA and the *B. subtilis* YfkN 5' nucleotidase precursor proteins indicate that ShsA may be a surface anchored 5' nucleotidase. The presence of this protein in *S. hominis*, a non-pathogenic commensal bacteria, calls into question the designation of AdsA as a virulence factor (Thammavongsa *et al.* 2009).

3.3.9.2 ShsC

From sequence comparisons using BLAST, ShsC contains a predicted metal-dependent hydrolase domain. Staphylococcal cell wall hydrolases are typically peptidoglycan, (murein) hydrolases. Peptidoglycan hydrolases have many housekeeping roles in cell division and homeostasis of the peptidoglycan layer (Frirdich & Gaynor 2013). Cell wall peptidoglycan hydrolases are implicated in pathogenicity of staphylococci with the murein hydrolase regulator CidA contributing to DNA release and biofilm development and LrgAB regulating penicillin sensitivity (K. C. Rice *et al.* 2007). The *S. aureus* surface protein AtlA is the most studied murein hydrolase of the staphylococci together with its *S. epidermidis* and *S. caprae* homologues AtlE and AtlC, respectively. The repeat regions of Atl proteins are know to bind fibronectin. ShsC shares little sequence identity with AtlA or AtlC therefore ShsC may represent a novel *S. hominis* CWA hydrolase. Further study is required to predict its substrate specificity.

3.3.9.3 ShsD

ShsD was revealed using BLAST to share similarity with the essential cell division protein FtsW. The membrane protein FtsW was shown in *E. coli* to connect cell wall synthesis and the cell division machinery and it is localised at the septum during cell division where it interacts with PBP3. More recently FtsW was revealed to transport the lipid-linked peptidoglycan precursor, lipid II, across the cytoplasmic membrane (Mohammadi *et al.* 2011).

3.3.9.4 ShsE

ShsE was determined to share similarity with SwrC. In *B. subtillis* SwrC is known to convey resistance to surfactin production and is thought to belong to the resistance-nodulation cell division (RND) family of transmembrane efflux transporters (X. Li *et al.* 2015). In *S. aureus* an RND transporter with some similarity to SwrC has been shown to interact with other proteins of the cell wall synthesis pathway (Quiblier *et al.* 2013).

3.3.9.5 ShsK

ShsK was annotated by Prokka as VraS and confirmed by BLAST similarity as a staphylococcal histidine kinase. VraS forms part of a two component sensor-regulator system VraRS and a member of the cell wall stress stimulon (Utaida *et al.* 2003). It is also known to have a role in the cell wall biosynthesis pathway (Kuroda *et al.* 2003). Expression of VraRS is upregulated in *S. aureus* in response to antibiotics whose action targets the cell wall (McAleese *et al.* 2006). Kuroda et al proposed that VraRS controlled the transcription of a set of 46 genes in the meticillin-resistant N315 strain of *S. aureus*, and that the cell wall stimulon including VraRS was induced as a result of peptidoglycan synthesis inhibition as opposed to the presence of cell wall active antibiotics (Kuroda *et al.* 2003). In *S. hominis* VraS was conserved across all 11 isolates, although sequence diversity was observed as only VraS in the multidrug resistant isolate ZBW5 was present in the PSI-blast analysis of the LPXTG-motif containing proteins (Table 3.2).

3.3.9.6 ShsN

ShsN was found to have evolutionary relationship to the DrrB family of ATP transporters. This family of ABC transporters was identified to have specificity for drugs, however some are also known to transport other substrates such as lipids and steroids (Borges-Walmsley & Walmsley 2001). ShsN lies immediately upstream of a homologous gene with an overlapping reading frame. This is in line with the expected structure of an ABC drug transporter which typically comprises two similar subunits.

3.3.9.7 Cell Wall-Associated Lipases

Lipases are commonly isolated from the staphylococci and are proposed to have roles in virulence and skin colonisation with evidence that surface bound lipases may interact with the extracelluar matrix(Sakinç *et al.* 2005). GehD *of S. epidermidis* is an example of one such surface-associated lipase shown to bind collagen (Bowden *et al.* 2002). One study found lipase production in 9 out of a panel of 11 species of CoNS tested; *S. hominis* exhibited lipase production in 50% of strains (Long *et al.* 1992). In the strains investigated here 6 of the 11

isolates were predicted to contain a possible lipase precursor (ShsG, Table 3.2), which agrees with the results of the previous study. Of note, the two species described to be negative for lipase production in 100% of strains tested were *S. haemolyticus* and *S. lugdunensis*, both of which are recognised as increasingly important opportunistic pathogens (Ravaioli *et al.* 2012) (Barros *et al.* 2012) .

3.3.9.8 CWA Peptidase and Haemolysin in *S. hominis*

The presence of a possible haemolysin in one of the 11 isolates from this study is not without precedence. Kloos & Schleifer found that 5 out of 20 isolates of *S. hominis* from two geographically distinct locations showed weak haemolytic activity (Kloos & Schleifer 1975).

Membrane-associated peptidases are known to occur in staphylococci, most notably the acquired transpeptidase PBP2a, which is encoded by the *mecA* gene, that confers resistance to meticillin in several species of staphylococci (Łeski & Tomasz 2005). The lateral transfer of PBP2a sets a precedent for peptidases as part of the accessory genome which is interesting since ShsH, a putative staphylococcal peptidase, is not universal but is present in a majority subset of 8 of the 11 isolates.

3.3.9.9 Putative CWA transporter proteins

The putative CWA proteins ShsB, ShsD, ShsI, ShsJ and ShsN were all identified as likely surface-associated transporter proteins. ShsB was annotated by Prokka as PitA, a low affinity inorganic phosphate transporter and this annotation was supported by BLAST. PitA is conserved across several species of bacteria including *E. coli*, *Listeria monocytogenes* and *S. aureus*. In *E. coli* PitA, as part of the PitAB transporter is concerned with the transport of inorganic phosphates as well as efflux of zinc(II) (Beard *et al.* 2000). In *S. aureus* PitA is a member of the GraRS regulon and shown to be upregulated by GraRS through expression profiling (Falord *et al.* 2011). In this study ShsB was conserved across all 11 isolates.

ShsF was found to encode a phosphotransferase system (PTS) glucose transporter, E11CBA. PTS systems are important in the regulation of many catabolic pathways and consist of the two general proteins cytoplasmic soluble enzyme I (EI) and a histidine-containing phosphocarrier protein. These general proteins are involved in the transport of all PTS sugars (Reizer *et al.* 1999). The specificity of the system comes from enzyme II (EII) which is a permease specific for its substrate comprising three domains. EIIA and EIIB are both hydrophilic domains located in the cytoplasm. EIIC is a hydrophobic transmembrane domain which interacts with and translocates the sugar substrate into the cell (Christiansen & Hengstenberg 1999). ShsF is conserved across all 11 isolates of *S. hominis*.

ShsJ was annotated by Prokka as a citrate/succinate antiporter. Citrate/succinate antiporters participate in the tricarboxylic acid (TCA) cycle which is accepted as a metabolic hub in aerobes and a system which undergoes regulation in pathogenic bacteria to modulate stress tolerance (Somerville & Proctor 2009). In *S. epidermidis* iron limitation and the presence of ethanol depress the TCA cycle and activate biofilm formation (Sadykov *et al.* 2010). It has also been suggested that mutations in the TCA cycle found in clinical isolates of *S. epidermidis* could contribute to increased survival in the presence of some β-lactam antibiotics (V. C. Thomas *et al.* 2013). Although ShsJ only fulfilled all 4 criteria for designation as a putative surface protein, in 2 of the 11 isolates tested it was conserved across all 11 isolates (Table 3.2). In the remaining 9 isolates ShsJ contained a signal peptide, a transmembrane domain and an LPXTG motif however it was not found to have homology with any of the staphylococcal surface proteins tested through PSI-blast. Average identity between the remaining 9 isolates was 99.8%.

3.3.10 Discrete *S. homnis* CWA proteins

Alongside the previously described putative CWA proteins there was another set of LPXTG-motif containing proteins which did not fulfil the criteria of homology to the panel of adhesins used as templates for PSI-blast (Table 3.3). These additional proteins did, however fulfil all other criteria for the analysis

and so were investigated as a secondary, discrete group of putative CWA proteins.

Table 3.3 Discrete *S. hominis* cell wall associated proteins. Homologous genes are grouped together. The closest match by blast is the sequence with the highest blast score after searching the cell wall associated protein against the BLAST database. Average identity is the average protein sequence identity between the protein sequences from the isolates which contain the cell wall associated protein.

Protein ID	Prokka annotation	Closest match by BLAST	Average identity
1			
B10-PB_01650	Antiholin-like protein LrgB	LrgB	99.9
C80-PB_01608			
J6-PB_01569			
J11-PB_01570			
J23-PB_01565			
J27-PB_01570			
J31-PB_01577			
I4-PB_01573			
SK119-PB_01572			
VCU122-PB_01601			
ZBW5-PB_01629			
2			
B10-PB_00664	Phosphatidate cytidylyltransferase	Phosphatidate cytidylyltransferase	99.7
C80-PB_00620			
J6-PB_00591			
J11-PB_00592			
J23-PB_00589			
J27-PB_00591			
J31-PB_00591			
I4-PB_00640			
SK119-PB_00576			
VCU122-PB_00587			
ZBW5-PB_00631			
3			
C80-PB_01812	Inner membrane transport protein YnfM	Major facilitatior protein	97.7
B10-PB_01860			
J6-PB_01778			
J11-PB_01779			
J23-PB_01774			
J27-PB_01779			
J31-PB_01786			
I4-PB_01782			
SK119-PB_01763			
VCU122-PB_01801			
ZBW5-PB_01824			
4			
J6-PB_01899	Hypothetical protein	Hypothetical protein	
J11-PB_01899			
J23-PB_01894			
J27-PB_01901			
J31-PB_01907			
I4-PB_01898			

3.3.10.1 LrgB

One of the discrete CWA proteins was annotated as the antiholin-like protein LrgB, which was confirmed by BLAST. LrgB is part of a family of bacterial cell death effectors characterised in *S. aureus*. The antagonistic effects of the *lrg* operon and the *cid* operon were shown to modulate the activity of murein hydrolase activity and susceptibility to antibiotics in liquid culture (Groicher *et al.* 2000). Coordination of cell lysis and death under the conditions of biofilm formation was observed to be a function of Lrg. An *lrgAB* mutant with decreased *lrgAB* expression exhibited increased lysis during biofilm formation consistent with its role as an inhibitor of cell lysis. Increased cell lysis leads to an increase of extracellular DNA and in turn leads to increased adherence of the biofilm (E. E. Mann *et al.* 2009). LrgB was conserved across all *S. hominis* isolates with a identity of 99.9%. The *S. hominis* protein shared 65.7% identity with *S. aureus* LrgB. In the *S. hominis* genome, *lrgB* was located next to a gene annotated by Prokka as *lrgA* and confirmed by BLAST. The protein shared 56.5% identity with *S. aureus* LrgA.

3.3.10.2 Cds

Cds is a phosphatidate cytidylyltransferase which catalyses the synthesis of CDP-diacylglycerol from phospatidic acid (PtdOH) and cytidine triphosphate (CTP). The reaction product is then subsequently used to produce peptidoglycan. Cds was shown to be an essential protein since deletion of the cognate gene was lethal (Kuhn *et al.* 2015). Cds is conserved with an average 99.7% identity across all 11 isolates of *S. hominis* sharing an average 84.7% identity with *S.aureus* CdsA.

3.3.10.3 Inner membrane transport protein YnfK

Annotation by Prokka revealed the presence of YnfK, a hypothetical transport protein of *E. coli*. BLAST showed that the protein was part of the major facilitator family of transport proteins. *S. aureus* is reported to contain aproximately 20 multi drug resistance efflux pumps, and the most abundant protein family among them is the major facilitator family of proteins (DeMarco

et al. 2007). The YnfK inner membrane transport protein was conserved across all isolates of *S. hominis* with an average identity of 97.7%. The *S. hominis* protein shared an average of 22.9% identity with the *E. coli* transporter YnfK and between 11.2 and 20.2% identity to characterised *S. aureus* MFS proteins (Table 3.4). The *S. aureus* MFS transporters which shared the highest identity with the *S. hominis* protein were NorA with 18.1% identity and Tet38 with 20.2% identity. NorA is proposed to have a broad substrate specificity and as such is able to transport hydrophilic quinolones and barbicides (H. Yoshida *et al.* 1990). Tet38 is known to be chromosomally encoded in *S. aureus* and has been shown to be important in skin colonisation in a BALB/c mouse model. Furthermore Tet38 expression was upregulated in the presence of subinhibitory concentrations of fatty acids (Truong-Bolduc *et al.* 2013).

Table 3.4 Analysis of the *S. hominis* multi facilitator superfamily proteins. sequence identity between the gene products of the well-characterised *S. aureus* MFS proteins and the *S. hominis* secondary CWA protein with sequence identity determined by BLAST to the multi facilitator superfamily of proteins.

***S. aureus* MFS protein**	**Reference**	**Average sequence identity with *S. hominis* MFS protein**
NorA	(Ubukata et al. 1989)	18.1
QacA/B	(Tennent et al. 1989)	16/17
MepA	(Kaatz et al. 2005)	14.5
MdeA	(J. Huang et al. 2004)	11.2
NorB	(Ding et al. 2008)	16.0
NorC	(Truong-Bolduc et al. 2006)	15.7
LmrS	(Floyd et al. 2010)	14.5
Tet38	(Truong-Bolduc et al. 2013)	21.2

3.3.11 Putative SERAMS

Analysis of putative SERAMs in the *S. hominis* genome revealed the presence of two regions in all 11 genomes with sequence similarity to the conserved staphylococcal SERAM Ebh (Table 3.5). Ebh is encoded by only one gene, *ebh,* in *S. aureus* strains 8325, COL and USA300 whereas it is encoded by 2 separate genes, *ebhA* and *ebhB* in strains N315 and Mu50. The separation of *ebh* into two separate genes is proposed to have occurred as a result of a frame shift mutation (Clarke *et al.* 2002). Furthermore, a truncated form of *ebh* is found in *S. aureus* strain Newman which lacks the transmembrane domain of the protein and has a length of 7021 bp (Cheng *et al.* 2014). EbhA is the C-terminal portion of Ebh, a giant extracellular matrix binding protein. The *ebhA* gene of *S. aureus* is homologous to the Emb of *Streptococcus defectivus* (Manganelli & van de Rijn 1999) (Clarke *et al.* 2002), Emb has in turn been shown to be responsible for extracellular matrix binding of the species in chemical mutagenesis studies (Tart & van de Rijn 1993).

In the *S. hominis* strains, SK119 and VCU122 region 1 was 6.5 kb in size, less than the equivalent region in the other strains where it was between 7.1-7.8 Kb. Region 1 was present as a single open reading frame in strains SK119, ZBW5 and the J31 PacBio assembly. Further analysis showed that this protein contained an LPXTG-like cell wall sorting signal. Furthermore, this reading frame has an average identity of 27.6% with *S. defectivus* Emb. Multiple protein sequence analysis also showed 15.6% identity with *S. hominis* EbhA, 15.49% identity with *S. hominis* EbhB, 18.4% identity with the *S. epidermidis* RP62A Ebh homologue and 16.02% identity with the *S. warneri* Ebh homologue.

Although this *S. hominis* protein with similarity to EbhA was found in the search for the LPXTG-motif, it is possible that this binds to the bacterial cell with an alternative cell surface binding mechanism as is the case for Ebh. Ebh and its *S. epidermidis* homologue, Embp, are known to be lacking an LPXTG binding motif and interact with the cell in a sortase-independent manner.

Table 3.5 The protein ID and length of the genomic loci with similarity to Ebh in *11 S. hominis* genomes. Total lengths were calculated by summing the individual lengths of the open reading frames. Protein IDs in bold indicate proteins identified in the LPXTG analysis.

Ebh similarity region 1

Genome	Protein ID	Overlapping reading frames\separation of reading frames	Total length kb
B10-PB	**01102**, 01103		7.2
C80-PB	**01051**, 01052		7.1
J6-PB	**01024**, 01025		7.4
J11-PB	**01025**, 01026		7.2
J23-PB	**01021**, 01022		7.3
J27-PB	**01023**, 01024	01023, 01024	7.7
J31-PB	**01026**, 01027		7.2
I4-PB	**01022**, 01023		7.3
SK119-PB	**01009**		6.5
VCU122-PB	**01028***, 01029, 01030, 01031	01029, 01030	6.5
ZBW5-PB	**01062**		7.6
J31 pacbio assembly	**01043**		7.8

Ebh similarity region 2a and 2b

Underlined protein IDs in region 2 denote proteins in the sub region 2a.

Genome	Protein ID	Overlapping reading frames	Total length kb
B10-PB	**01564**, 01587	(reading frames separated by 16,545 bp)	10.5
C80-PB	**01518**, 01545	(reading frames separated by 17,019 bp)	10.6
J6-PB	**01500**, 01501, 01503, 01504, 01505*	01503, 01504, 01505	6.4
J11-PB	**01503**, 01504, 01505, 01506, 01507*	01505, 01506, 01507	14.2
J23-PB	**01499**, 01500, 01501, 01502, 01503		13.8
J27-PB	**01502**, 01503, 01504, 01505, 01506, 01507, 01508	01505, 01506	13.9
J31-PB	**01510**, 01511, 01512, 01513*	01510, 01511, 01512, 01513	6.0
I4-PB	01506, 01507, 01508, 01509, 01510, 01511		13.6
SK119-PB	**01487** 01510	(reading frames separated by 17,517 bp)	10.6
VCU122-PB	**01515**, 01539	(reading frames separated by 15, 357 bp)	10.5
ZBW5-PB	**01545**, 01566	(reading frames separated by 14,298 bp)	10.6
J31 pacbio assembly	**01534**		15.1

* the rest of the open reading frame was indicated through lack of stop codons, but no protein coding sequence predicted.

Ebh region 1 in seven of the eleven isolates consists of 2 adjacent reading frames of ~4.1 kb and ~3.1kb with similarity to Ebh proteins. With three *S. hominis* isolates these genes were fused into one reading frame, and in the remaining isolate the gene was fragmented into four open reading frames (Table 3.5). In the isolates where region 1 was present as a single open reading frame the locus was found as part of the signal peptide analysis of the LPXTG motif-containing proteins. Further investigation of the open reading frames of the remaining 8 isolates revealed a signal peptide in the N terminal-associated open reading frame of the region. Disparity between the Illumina and PacBio assembly of the J31 isolate, along with the overlapping reading frames of J27 and VCU122 is evidence that the separated reading frames may in fact be a sequencing artefact, which the longer PacBio reads have been able to resolve the locus more effectively than the Illumina sequencing method. Alternatively, there may be divergence at this locus within *S. hominis* as is observed in *S. aureus.*

The LPXTG-motif containing protein in region 1 was found by BLAST analysis to have similarity with Mrp or SasC in the isolates ZBW5, VCU122, SK119, K8 and C80. Mrp is known to be a protein homologous to Ebh in *S. aureus* and implicated in meticillin resistance through modification of cell wall peptidoglycans (Wu & De Lencastre 1999). SasC was also discovered in *S. aureus* as part of a search for LPXTG-motif containing proteins and was found to have an N-terminal domain important in cell aggregation (Schroeder *et al.* 2009). SasC also exacerbated the biofilm forming capabilities of *S. carnosus* upon plasmid-encoded expression of the surface protein. Mrp and SasC share two domains, DUF1542 and a motif found in various architectures (FIVAR motif) both found in other cell surface proteins, significantly Ebh and its *S. epidermidis* homologue Embp (Schroeder *et al.* 2009).

Region 2 is split into two sub regions, 2a and 2b (Table 3.5). Region 2a consists of an ORF containing an LPXTG-motif. In 7 of the 11 isolated this locus was adjacent to 2-6 additional open reading frames with similarity to Ebh. In the remaining 4 isolates the LPXTG protein was separated by between 14.3 and

17.0 kb from Ebh similarity region 2b. The PacBio assembly of the J31 genome was different from each of these groups with both region 2a and 2b being located as a single ORF.

There are several possible explanations that might account for this potential sequence discrepancy. The first is that there is indeed sequence divergence at this locus and two different loci occur. All open reading frames with similarity to Ebh at this locus may have been a single open reading frame in the evolutionary past and mutation events such as insertion of mobile genetic elements might have resulted in the locus becoming fragmented into several ORFs. Insertion of genetic material following an early gene fragmentation event may have resulted in the group of isolates where region 2a and 2b became separated. Good evidence against this possibility is, again, the disparity at this locus between the Illumina and PacBio assemblies of the J31 genome. The PacBio assembly indicates that regions 2a and 2b are fused whereas the Illumina assembly suggests a fragmented locus. As PacBio reads are longer than illumina reads (10kb vs 250bp) fewer reads are required to span a genomic region such as this. Due to the declining quality of sequence data at the ends of reads it is possible that there are sequencing errors within a potentially very large open reading frame traversed by multiple reads such as this one, which could affect gene calling and result in the appearance of multiple shorter open reading frames. Assembly errors could result in the spatial separation of region 2a and 2b in the isolates B10, C80, VCU122 and ZBW5, or alternatively the gene fusion in the PacBio assembly of J31. Validation by Sanger sequencing of these particular genomic regions could be used to clarify the genome structure.

Multiple sequence alignments were generated with the protein sequences of the J31 PacBio assembly of region 2, *S. aureus* strain Newman EbhA and EbhB, the putative *S. warneri* Ebh homologue and *S. epidermidis* Ebh homologue. Region 2 shares 30% identity with the *S. warneri* putative Ebh homologue, 34% identity with *S. epidermidis* RP62A Ebh homologue and 29.6% identity with *S. aureus* Newman EbhB. Region 2 also showed 18.1% identity with *S. defectivus* Emp

which is consistent with the 18.54% identity with *S. epidermidis* Ebh and the 19.12% identity found in both Newman and *S. warneri*.

Region 2 from the J31 PacBio assembly was analysed for the presence of transmembrane domains. None were found which is consistent with the *S. warneri* Ebh homologue which was also found not to contain transmembrane domains. *S. aureus* Newman was also found to carry a truncated Ebh without transmembrane domains, however it was localised to the cell surface despite this truncation (Cheng *et al.* 2014). Localisation of Ebh in *S. aureus* was revealed to follow the expected pattern for proteins containing a YSIRK-motif signal peptide (Cheng *et al.* 2014). The putative *S. hominis* Ebh homologue similarly contains such a signal peptide. Furthermore a SasC/Mrp/FmtB cellular aggregation domain was also observed supporting findings that Ebh has a role in cell aggregation and biofilm formation. It was proposed that *S. epidermidis* Embp, the Ebh homologue, is involved primarily in tolerance to the osmotic pressures encountered on the skin and that its role in cell aggregation may be through cell-to-cell adhesion (Linnes *et al.* 2013).

Although Ebh proteins represent some of the largest open reading frames in the *S. hominis* genome their length falls far short of *S. aureus* Ebh and *S. epidermidis* Embp ORFs which are around 30 kb in length with strain dependent variation observed. There is no open reading frame of this length in any of the *S. hominis* genomes.

Despite questions about the genome organisation at these loci the presence of proteins with similarity to Ebh suggest that *S. hominis* secretes proteins capable of binding to the host extracellular matrix. This would be expected as it is thought that a range of adhesins and MSCRAMMs, such as wall teichoic acid, IsdA and ClfB, are required at different stages of epithelial colonisation by *S. aureus* (Burian *et al.* 2010), although this theory has not been explored in other species of staphylococci.

3.3.12 YSIRK sequence motif

The YSIRK motif is located within the signal peptide of some LPXTG containing proteins and localises these proteins near the cell division septum (Carlsson *et al.* 2006). This results in a ring pattern of distribution around the cell (Cheng *et al.* 2014). In *S. aureus* it is known that expression of surface protein display (*spdABC*) genes is required for the trafficking of these proteins to the correct cell wall compartment (Frankel *et al.* 2010). Surface proteins lacking the YSIRK motif by contrast are localised to the cell poles (Schneewind & Missiakas 2012).

In *S. hominis* this YSIRK motif was conserved and was present in ShsB and the N terminus of both Ebh regions 1 and 2.

3.3.13 Ica locus and IS*256* insertion sequence

The presence of *icaA* and IS*256* are proposed to be markers of *S. epidermidis* invasiveness due to their involvement in biofilm production (Kozitskaya *et al.* 2004). Chromosomally encoded IS*256* in conjunction with the *ica* genes is important, particularly as IS*256* is thought to alter gene expression of the *ica* locus resulting in loss of the biofilm-positive phenotype. Reversible insertion and excision of IS*256* into either *icaA* or *icaC* results in inactivation of these genes (Ziebuhr *et al.* 1999).Their presence or absence is used to inform clinical decisions regarding treatment of blood cultures positive for CoNS, which are often regarded as contamination due to their commensal lifestyle and ubiquitous distribution on human skin. Importantly, however, diversity in the pathways of biofilm formation has been shown in *S. epidermidis* with only 60% of biofilm-forming isolates testing positive for the presence of *ica* genes (Conlan *et al.* 2012). There is also evidence to support biofilm production in *S. hominis* (Mendoza-Olazarán *et al.* 2013), (Christensen *et al.* 1983), (Jones *et al.* 1992). The increasing involvement of *S. hominis* in clinical disease, its potential for forming a biofilm and the reported lack of *icaA* gene suggests a need for a wider strategy to assess invasiveness in CoNS involving a panel of markers which represent the more complex story of biofilm formation in the staphylococci. To examine further the genetic potential for biofilm formation in *S. hominis* the

available genomes were examined for both the *ica* gene locus as well as other genetic factors implicated with biofilm formation in staphylococci.

The *ica* gene locus was not observed in any of the *S. hominis* in this study. This is in accordance with reported data which also failed to identify *icaABCD* genes in *S. hominis* (de Silva *et al.* 2002). IS*256* was observed in the ZBW5 multidrug resistant isolate, however without the presence of the *ica* genes, biofilm phenotype phase variation is unlikely to occur via IS*256* in this strain. Despite the absence of the *ica* locus, all seven isolates sequenced in this study were found to have biofilm forming capabilities (Figure 3.7) indicating that this capability is independent of PIA/PNAG synthesis.

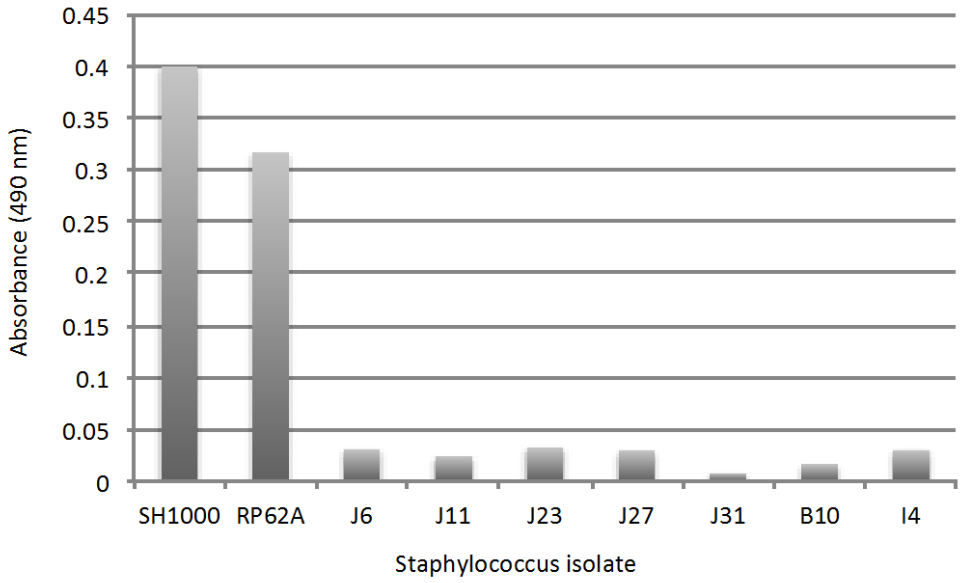


Figure 3.7 Biofilm forming capabilities of the *S. hominis* isolates sequenced in this study. *S. aureus* SH100 and *S. epidermidis* RP62A were used as positive controls as they are both known to produce a strong biofilm. Absorbance of all test wells were normalised against a methyl blue treated control well containing only BHI supplemented with glucose, without the presence of bacteria.

Other staphylococcal genes known to be involved in adherence and accumulation of biofilms include *S. epidermidis* Embp, the accumulation factor Bhp, and the murein hydrolase Atl. Bioinformatic analysis using psi-BLAST did not reveal the presence of a Bhp homologue with its characteristic Ig repeat regions among the surface proteins (Cucarella *et al.* 2001) (Lasa & PenadEs 2006) encoded in any of the *S. hominis* genomes. As described in detail previously, two regions showing similarity to Ebh, the *S. aureus* homologue of Embp, were found in all 11 isolates of *S. hominis*. Furthermore, putative CWA proteins which may have the capability to bind host ECM and cells were found in the *S. hominis* genomes. This suggests that *S. hominis* has the potential to initiate attachment to host proteins, the first stage of *in vivo* biofilm formation.

The biofilms formed by the biofilm-positive strains in this study were observed predominantly at the air-liquid interface with biofilm formation evident to a much lesser extent at the liquid-solid interface. Air-liquid interface biofilms are produced in the significant clinical pathogens *Pseudomonas aeruginosa* (Friedman & Kolter 2004) and *Salmonella typhimurium* (Römling *et al.* 2000) among others. This mechanism differs from the *S. epidermidis* and *S. aureus* strains tested which form biofilms predominantly at the liquid-solid interface. It is worth noting that a range of growth media, the presence of plasma proteins and environmental conditions were not tested with *S. hominis* and there remains the possibility that the Ica-independent biofilm formation of this species might be dependent on particular environmental cues.

The ability to form a biofilm is not only considered to be a predictor of invasiveness, it has also been suggested that biofilm production may influence the ability of a bacterial species to colonise alternative niches. This is particularly clinically relevant since biofilms provide bacterial species with the opportunity to seed new sites with infective doses of cells (Hall-Stoodley & Stoodley 2005).

3.3.14 Antibiotic resistance profile, associated transposons and insertion sequences

There is evidence to suggest that the antibacterial resistance profiles of the resident skin flora may have important diagnostic worth in the hospital setting. Obolski *et al* (2014) found the resistance profiles of CoNS contaminating blood cultures to be effective predictors of the resistance profiles, as well as specific antibiotic resistance, of the infecting organism (Obolski *et al.* 2014).

Predicted bacitracin resistance was conserved across all 11 isolates of *S. hominis* in this study with *bacA* universally present (Table 3.6); this gene was found to be chromosomally encoded in all isolates. BacA is an undecaprenyl pyrophosphate phosphatase and dephosphorylates undecaprenyl pyrophosphate, or UPP (C_{55}-PP), which is the precursor for undecaprenyl phosphate or UP (C_{55}-P), a lipid integral to synthesis of peptidoglycan and lipopolysaccharide in Gram-negative bacteria. It is a carrier lipid which facilitates transport of hydrophilic precursors through the hydrophobic cell membrane where they are externally polymerised. Dephosphorylation of UPP occurs after each round of teichoic acid synthesis to make the lipid carrier UP available for the transport of further hydrophilic intermediates (Ghachi *et al.* 2004). Bacitracin is a cyclic polypeptide antibiotic whose bactericidal capabilities come from its ability to sequester the UP precursor, UPP, to prevent it from functioning as a lipid carrier thus disrupting wall teichoic acid synthesis. Approximately 80% of the reservoir of undecaprenol in the staphylococcal cell is present in the form of the free alcohol UP-OH (Higashi *et al.* 1970). It was suggested that this constitutes a pool of UP undecaprenol available to restock the supplies of UP should this be required (Ghachi *et al.* 2004). It has been shown, however controversially (Ghachi *et al.* 2004), that BacA possessed isoprenol kinase activity when plasmid encoded in *E. coli* (Cain *et al.* 1993) and it was also suggested that BacA may have undecaprenol kinase capabilities in *S. aureus* (Chalker *et al.* 2000) and functions to phosphorylate UP-OH and replenish the pool of UP under the condition of its sequestration by bacitracin (Yoshida *et al.* 2011).

Table 3.6 *S. hominis* antibiotic resistance profiles. The genes responsible for resistance to the antibiotics are listed are shown for each strain.											
Strain	**B10**	**C80**	**J6**	**J11**	**J23**	**J27**	**J31**	**I4**	**SK119**	**VCU122**	**ZBW5**
Bacitracin	*baca*	*baca*	*baca*	*baca*	*baca*	*baca*	*baca*	*baca*	*baca*	*baca*	*baca*
Penicillin	*bl2a-pc*		*bl2a-pc*	*bl2a-pc*	*bl2a-pc*	*bl2a-pc*	*bl2a-pc*	*bl2a-pc*			*bl2a-pc*
MLS*	*msra*	*msra*							*msra*	*msra*	*ermc*
Streptogramin a									*vgaa*		
Macrolide	*mphc*								*mphc*	*mphc*	
Fusaric acid		*fusb*									
Tetracycline										*tetK*	*tetk*
Bleomycin											*ble*
Lincomycin											*lnua*
Beta lactams											*meca*
Methicillin											*mecr1*
Isepamicin											*aac6ie*
Netilmicin											*aac6ie*
Tobramycin											*aac6ie*
Amikacin											*aac6ie*
Sisomicin											*aac6ie*
Dibekacin											*aac6ie*
Tobramycin											*aadd*
Kanamycin											*aadd*
*MLS resistance is resistance to macrolides, lincosomides and streptogramin A											

Penicillin resistance was predicted in 8 of the 11 isolates through the presence of a type A beta-lactamase (Table 3.6). The beta-lactamase found in the S. hominis isolates was *blaZ*, known to confer resistance only to penicillin. Studies have shown that BlaI, the repressor of *blaZ* and MecR1, which in turn is the repressor of *mecA*, are homologous proteins which are capable of interacting to form heterodimers. It is proposed that these heterodimers bind with reduced affinity to the *mecA* promoter which results in reduced repression of *mecA* and the full expression of the meticillin resistance phenotype (Arêde *et al.* 2013). This is particularly significant in this study since *mecA, mecR1 and blaI* were all present in the genome of *S. hominis* isolate ZBW5.

Aminoglycoside resistance is conferred by the gene *aacA-aphD* and a copy was predicted in the genome of ZBW5 (Table 3.6). This gene is carried on the staphylococcal transposon Tn*4001* found on composite phage as well as the SCC*mec* transposable element (Rouch *et al.* 1987) (Kinnevey *et al.* 2014). On this transposon the *aacA-aphD* resistance determinant is also flanked by the IS*256*L and IS*256*R sequences. The gene *aacA-aphD* encodes an aminoglycoside modifying enzyme (AME) and is also harboured on the transposon Tn*4100* in *S. aureus,* which is similar to Tn*4001* however it contains a similar insertion sequence IS*257* in place of IS*256* (Culebras & Martínez 1999) and Tn*4031* in *S. epidermidis*. In *S. hominis* strain ZBW5 however *aacA-aphD* was not found to be flanked by IS*256*, however a single copy of IS*256* was found to be present in the genomic region predicted to be an intact phage at the coordinates 2253755-2328024 (Table 3.9). The amino glycoside resistance gene falls within this phage region, indicating that it is phage encoded.

In addition to the presence of *aacA-aphD* (coordinates 2296616-2299055), PBP2A and meticillin resistance regulatory elements (coordinates 2255726-2257854) were also observed in the same region of the genome along with several genes considered to be markers of the SCC*mec* element. A putative cassette chromosome recombinase with 93.7% identity to *S. aureus* ccrC2 was found downstream of the putative SCC*mec* element. The IS*431* insertion sequence is found on all types of SCC*mec* elements identified (International

Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC) 2009) and a copy of this sequence was found in the vicinity of the mec element in *S. hominis* ZBW5. The insertion sequence IS*256* is associated with Tn*4001* found on SCC*mec* type IV (Chlebowicz *et al.* 2010) and this IS element was found between the mec element and *aacA-aphD*. There is also evidence to show that mecA has been isolated from phage DNA from environmental samples (Colomer-Lluch *et al.* 2011) and to a lesser extent human faecal samples (Quirós *et al.* 2014). The presence of an SCC*mec* element similar to the type IV cassette could explain the genome region described here which contains both phage proteins and genes associated with SCC*mec*. Alternatively this may be a poorly resolved area of the genome sequence which contains multiple mobile elements. It is known that these elements are difficult to resolve where some sequencing technologies are used (Steinig *et al.* 2015).

This analysis does however present strong evidence that aminoglycoside resistance and meticillin resistance are both carried on mobile genetic elements in *S. hominis*. This provides additional evidence that *S. hominis* might donate antibiotic resistance determinants to other staphylococcal species, and also potentially acquire them through horizontal gene transfer (Strasheim *et al.* 2013).

Another important resistance phenotype is resistance to macrolides, lincosomides and streptogramin B (MLS). MLS resistance was predicted in 5 of the 11 isolates of *S. hominis* through two different pathways, *ermC* in strain ZBW5 and *msrA* in strains B10, C80, SK119 and VCU122 (Table 3.6). The gene responsible for MLS resistance was found immediately adjacent to those responsible for macrolide resistance in strain B10. These genes were located with an adjacent transposase gene immediately up and downstream indicating that they may be associated with a mobile genetic element such as a transposon. In strain SK119, these genes are also found in an area of the genome where transposases and plasmid-associated genes are also present. This is also true of the gene, *vgaA,* responsible for streptogramin resistance in this strain. In strain VCU122, MLS, macrolide and tetracycline resistance determinants are all

located immediately adjacent to each other. A transposase is also found adjacent to these resistance genes and they are found near to plasmid-associated genes. Similarly in strain C80, MLS resistance determinants reside in an area of the genome where plasmid and transposon-associated genes are present. As previously discussed, sequences of mobile genetic elements such as plasmids and phage are difficult to assemble and therefore genes associated with these MGEs often cluster at the end of genomes. This appears to be the case for the *S. hominis* genomes in this study, with the exclusion of the PacBio assembly of *S. hominis* strain J31. The presence of MLS and streptogramin A resistance genes in this unresolved area of the genome indicates that these genes may also be carried on a mobile genetic element.

MLS antibiotics are increasingly being prescribed as an alternative to meticillin as a response to meticillin resistance. An increased incidence of MLS prescription has likely selected for increasing levels of MLS resistance. MLS resistance phenotypes are complex and an understanding of their genetic basis is vital to correctly inform antibiotic therapy choices. MLS resistance attributable to *ermC,* as in the case of strain ZBW5, renders the use of first line antibiotics erythromycin and clindamycin ineffective. Resistance conferred by *msrA* however does not result in clindamycin resistance and so leaves this first line treatment available (Steward *et al.* 2005). The isolate SK119 was also predicted to contain an additional efflux pump to *msrA* which confers resistance to streptogramin A. Although the proteins encoded by the two loci are similar they do display divergent drug specificities (Chesneau *et al.* 2005) thought to be mediated by the C-terminus of the protein (Jacquet *et al.* 2008).

The fusidic acid resistance gene *fusB* was present in only 1 isolate, C80 (Table 3.6), where the gene is located near to plasmid and transposon-associated genes. FusB confers fusidic acid resistance in a highly unusual manner for antibiotic resistance by binding to elongation factor G and stalling the ribosome thus halting protein synthesis (Verbist 1990). FusB binds to the elongation factor via a zinc binding domain at a C-terminal site, the fusidic acid binding site is located at the N-terminus. Due to this spatial disparity in binding sites it was

hypothesised that steric hindrance of fusidic acid binding by FusB was unlikely. FusB causes destabilisation of the ribosome:elongation factor:GDP complex in the presence of fusidic acid which allows translation and therefore protein synthesis to continue (Cox *et al.* 2012).

3.3.15 Mobile genetic elements

3.3.15.1 Plasmids

Experimental data suggest that *S. hominis* J31 contains at least three plasmids. This is confirmed by PacBio sequence data with four of the five contigs obtained predominantly showing BLAST hits to staphylococcal plasmids. Further evidence to support that these contigs represent *S. hominis* plasmids is the presence of replication proteins on these contigs and indications from the initial PacBio sequence analysis that these contigs represent circular pieces of DNA (section 2.3.3). Each of the putative plasmid contigs are also indicated to be circular due to the annotation of similar gene fragments at both ends of the contigs (Table 3.7a-d).

Table 3.7a The genes associated with plasmid contig 1. BLAST revealed similarity to *Staphylococcus warneri* SG1 plasmid clone pvSw5 genomic sequence. Query coverage- 20%, E-value- 0, Identity- 99%

Gene name	Gene product	Plasmid annotation
czcD_1	Cadmium 2C cobalt and zinc/H(+)-K(+) antiporter	plasmids_00004
czcD_2	Cadmium 2C cobalt and zinc/H(+)-K(+) antiporter	plasmids_00005
czcD_3	Cadmium 2C cobalt and zinc/H(+)-K(+) antiporter	plasmids_00006
	Glutaredoxin 3	plasmids_00007
csoR_1	Copper-sensitive operon repressor	plasmids_00008
copA_1	Copper-exporting P-type ATPase A	plasmids_00009
copZ_1	Copper chaperone CopZ	plasmids_00010
	Hef nuclease	plasmids_00013
mobA_1	DNA strand transferase	plasmids_00017
	Integrase core domain protein	plasmids_00020
	YolD-like protein	plasmids_00023
	Putative transposase DNA-binding domain protein	plasmids_00025
soj	Sporulation initiation inhibitor protein Soj	plasmids_00027
	anaerobic benzoate catabolism transcriptional regulator	plasmids_00030
	Replication initiation factor	plasmids_00034
spsB	Signal peptidase IB	plasmids_00035
msbA	Lipid A export ATP-binding/permease protein MsbA	plasmids_00042
	ABC-2 family transporter protein	plasmids_00043
	Pentapeptide repeats (8 copies)	plasmids_00045
yidC	Membrane protein YidC	plasmids_00047
bin3_1	Putative transposon Tn*552* DNA-invertase bin3	plasmids_00048
	ABC-type enterochelin transport system 2C ATPase component	plasmids_00050
	Integrase core domain protein	plasmids_00053
ywrO	General stress protein 14	plasmids_00055
zwf	Glucose-6-phosphate 1-dehydrogenase	plasmids_00057
gnd	6-phosphogluconate dehydrogenase 2C decarboxylating	plasmids_00058
pgi	Glucose-6-phosphate isomerase	plasmids_00059
	3-hexulose-6-phosphate synthase	plasmids_00060
hxlB	3-hexulose-6-phosphate isomerase	plasmids_00061
hxlR	HTH-type transcriptional activator hxlR	plasmids_00062
	Integrase core domain protein	plasmids_00063
immR	HTH-type transcriptional regulator immR	plasmids_00065
	Integrase core domain protein	plasmids_00066
czcD_4	Cadmium 2C cobalt and zinc/H(+)-K(+) antiporter	plasmids_00068
	Glutaredoxin 3	plasmids_00069
csoR_2	Copper-sensitive operon repressor	plasmids_00070
copA_2	Copper-exporting P-type ATPase A	plasmids_00071
copZ_2	Copper chaperone CopZ	plasmids_00072
ydaF_2	Putative ribosomal N-acetyltransferase YdaF	plasmids_00074
	Putative DnaQ family exonuclease/DinG family helicase	plasmids_00076
	Hef nuclease	plasmids_00077

Table 3.7b Genes associated with plasmid contig 2. BLAST revealed similarity to *Staphylococcus epidermidis* ATCC 12228 plasmid pSE-12228-04, complete sequence. Query coverage- 28%, E-value- 0, Identity- 95%

Gene name	Gene product	Plasmid annotation
bin3_2	Putative transposon Tn552 DNA-invertase bin3	plasmids_00078
	Putative ATP-binding protein involved in virulence	plasmids_00079
bmrA_1	Multidrug resistance ABC transporter ATP-binding/permease protein BmrA	plasmids_00082
	Conjugal transfer protein TrbA	plasmids_00083
mobA_2	DNA strand transferase	plasmids_00089
	Polar chromosome segregation protein	plasmids_00093
lrpC	HTH-type transcriptional regulator lrpC	plasmids_00095
	Putative integral membrane protein	plasmids_00097
ecsA_1	ABC-type transporter ATP-binding protein EcsA	plasmids_00099
	ABC-type transport system involved in multi-copper enzyme maturation 2C permease component	plasmids_00100
bin3_3	Putative transposon Tn*552* DNA-invertase bin3	plasmids_00102
	putative ATP-binding protein involved in virulence	plasmids_00103
bmrA_2	Multidrug resistance ABC transporter ATP-binding/permease protein BmrA	plasmids_00106
	Conjugal transfer protein TrbA	plasmids_00107

Table 3.7c Genes associated with plasmid contig 3. BLAST revealed similarity to *S. saprophyticus* subsp. saprophyticus MS1146 plasmid pSSAP1 complete genome. Query coverage-16%, E- value- 0, Identity- 98%

Gene name	Gene product	Plasmid annotation
drrA_1	Daunorubicin/doxorubicin resistance ATP-binding protein DrrA	plasmids_00119
	Divergent AAA domain protein	plasmids_00120
nisC	Nisin biosynthesis protein nisC	plasmids_00127
	Thiopeptide-type bacteriocin biosynthesis domain protein	plasmids_00128
apr	Subtilisin	plasmids_00129
pepA	Lantibiotic Pep5 precursor	plasmids_00130
fabG	3-oxoacyl-[acyl-carrier-protein] reductase FabG	plasmids_00132
ecsA_2	ABC-type transporter ATP-binding protein EcsA	plasmids_00133
	ABC-type transport system involved in multi-copper enzyme maturation 2C permease component	plasmids_00134
bin3_4	Putative transposon Tn*552* DNA-invertase bin3	plasmids_00135
ydgH	Putative membrane protein ydgH	plasmids_00136
opuAC	Glycine betaine-binding protein precursor	plasmids_00137
	Putative dihydroxyacetone kinase regulator	plasmids_00138
	Glyoxalase-like domain protein	plasmids_00139
	MepB protein	plasmids_00140
	Transcriptional repressor DicA	plasmids_00142
drrA_2	Daunorubicin/doxorubicin resistance ATP-binding protein DrrA	plasmids_00147
	Divergent AAA domain protein	plasmids_00148
	Type 2 lantibiotic biosynthesis protein LanM	plasmids_00156
	Type 2 lantibiotic biosynthesis protein LanM	plasmids_00157
	Thiopeptide-type bacteriocin biosynthesis domain protein	plasmids_00159
	Thiopeptide-type bacteriocin biosynthesis domain protein	plasmids_00160

Table 3.7d Genes associated with plasmid contig 4. BLAST revealed similarity to *S. epidermidis* CH plasmid pSepCH complete sequence, Query coverage-100%, E-value- 0.0, Identity 99%

Gene name	Gene product	Plasmid annotation
	Replication protein	plasmids_00163
qacC_1	Quaternary ammonium determinant C	plasmids_00164
	Replication protein	plasmids_00166
qacC_2	Quaternary ammonium determinant C	plasmids_00167
	Replication protein	plasmids_00170

BLAST hits (99% identity, 100% query coverage) of plasmid contig 4 to the *S. epidermidis* plasmid pSepCH and the presence of *qacC* suggest the presence of a *qacC* group plasmid (Table 3.7d). Plasmid contig 3 carries putative lantibiotic bacteriocin biosynthesis genes and Blast revealed high sequence identity to the epicidin gene cluster of plasmid pCH01 (Heidrich *et al.* 1998) and the *S. warneri* plasmid pPI-1 known to encode the nukacin biosynthesis gene cluster (Aso *et al.* 2005) (Table 3.7c). The *nuk* operon genes of the *S. warneri* plasmid were not found in the *S. hominis* J31 genome, however. *S. warneri* strain ISK-1 harbours a *qacC* plamid, pPI-2(Aso *et al.* 2005). *S. epidermidis* strain CH has the *qacC* plasmid pSepCH metioned above plus the epicidin encoding plasmid pCH01 discussed above. The putative lantibiotic biosynthesis cluster in *S. hominis* J31 has a conserved gene synteny when compared to the epicidin biosynthesis gene cluster of *S. epidermidis* (Figure 3.8). Gene product identity is between ~80-93%. The similarity observed between the homologous *S. epidermidis* lantibiotics epicidin and pep5 was 68-75% (Heidrich *et al.* 1998). The *S. hominis* lantibiotic biosynthesis gene products have 55-58% identity to the pep5 operon gene products. Upstream of the *eci* operon in *S. epidermidis* BN 280 lies the transposon Tn*4003* associated with the insertion sequence IS*257* (Heidrich *et al.* 1998). In *S. hominis* J31 the tranposase for Tn*552* is present. Tn*552* is however also associated with IS*257* (Sidhu *et al.* 2002). Furthermore the plasmid gene *drrA* has BLAST similarity to a bacitracin ABC transporter of *S. simulans*. The presence of pSepCH and pCH01 in *S. epidermidis* CH, pPI-1 and pPI-2 in *S. warneri* ISK-1 plus the presence of a putative lantibiotic biosynthesis plasmid and *qacC* plasmid in *S. hominis* J31 could suggest that these two plasmids often occur together.

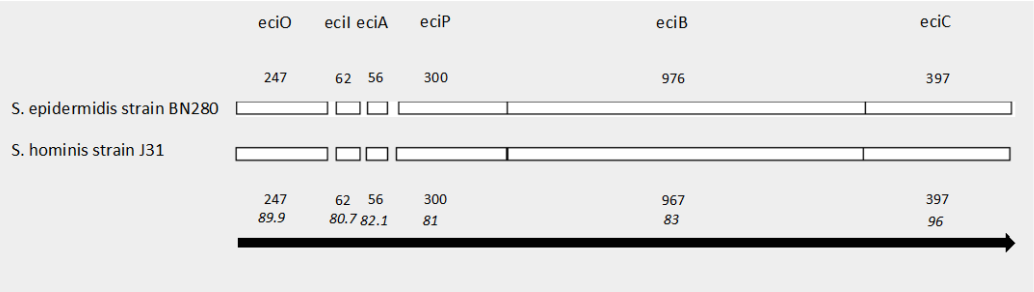


Figure 3.8 Lantibiotic biosynthesis operon. Representation of the synteny of the epicidin gene cluster in *S. epidermidis* BN280 and the lantibiotic biosynthesis gene cluster of *S. hominis* J31. The numbers indicate the number of amino acids comprising the genes and the arrow denotes the direction of transcription. The numbers in italics indicate the % identity between the residues in each gene product.

S. hominis was described as a bacteriocin producer (Santos Nascimento *et al.* 2005) however the gene cluster found in the strain J31 in this study contrasts with earlier studies. Wilaipun et al identified a homologue of the *S. warneri* encoded lantibiotic nukacin ISK-1 which was designated nukacin KQU-131 (Wilaipun *et al.* 2008). Another study detailed the sequence of yet another lantibiotic encoded by *S. hominis* dubbed hominicin (Kim *et al.* 2010). Hominicin was found to be a type I lantibiotic of 21 amino acids in length, without similarity to other antimicrobial peptides and which does not contain thioether bridges in its structure (Kim *et al.* 2010). The two analogous peptides pep5 and epicidin are both around 30 amino acids in length and both contain a similar pattern of thioether bridges (Heilmann 2011). Given the high degree of sequence identity between the epicidin biosynthesis pathway genes and the homologous operon observed in *S. hominis* it may be extrapolated that a similarly structured peptide would be expected to result from it. It is documented that other species of staphylococci, including *S. epidermidis* and *S. aureus*, can encode multiple lantibiotic bacteriocins (Santos Nascimento *et al.* 2005). Perhaps then the *S. hominis* pan-genome encodes a repertoire of different bacteriocins.

Lantibiotics are synthesised by the ribosome as two precursors; an N-terminal leader and a C- terminal propeptide which is modified by lanthionine synthase enzymes. LanC is the lanthionine synthase responsible for the post translational modification of type A(I) lantibiotics and is the synthase encoded by the *S. hominis* genomes in this study. Nukacin ISK is a type A(II) lantibiotic encoded by *S. warneri* and its constitutive pre-peptides are post translationally modified by LanM lanthionine synthase (Nagao *et al.* 2005).

The epicidin gene cluster does not have an associated TCS or ABC transporter, however expression in *S. carnosus* of only the epicidin gene cluster is sufficient to result in expression of the lantibiotic epicidin. It is hypothesised that a chromosomally-encoded transporter fulfils the function of exporting the lantibiotic in *S. epidermidis*. In *S. hominis* the presence of a putative bacteriocin ABC transporter encoded on the same plasmid as the lantibiotic synthesis operon suggests that this may not be the case in *S. hominis* strain J31.

Genes involved in metal ion resistance were also found to be encoded on plasmid contig 1 in *S. hominis* (Table 3.7a). Divalent cations such as zinc, iron and nickel are required for the growth and replication of bacteria and play catalytic (Sabala *et al.* 2014) as well as structural roles (Krishna *et al.* 2003) within proteins at nanomolar concentrations. Heavy metal ions are often added to feed stocks to reduce gastrointestinal diseases in livestock and as a result heavy metal resistance is widespread among livestock-associated strains of bacteria (Cavaco *et al.* 2010). Consequently heavy metal resistance mechanisms have evolved and are often found encoded by plasmids in addition to being found on the chromosome. *czc* genes are found on the chromosome in *S. aureus* (Poston & Hee 1991) (Kuroda *et al.* 1999) and the genomes of the *S. hominis* isolates in this study. Additional *czc* and *czr* genes are also found on plasmid contig 1 which is consistent with the previous observation of cadmium/zinc^{2+} resistance via a different mechanism on the pI258 plasmid of *S. aureus* (Ye *et al.* 2005). Heavy metal resistance has often been found to coexist with resistance to certain antibiotics and often on the same plasmid, this may not the case in this study since J31 contains only two probable antibiotic resistance factors one of

which is chromosomally encoded, the other appears phage-associated. There are however several putative multidrug ABC transporters present on the plasmid contigs that may represent uncharacterised pathways of antibiotic or biocide resistance in staphylococci.

3.3.15.2 Conservation of plasmid genes among *S. hominis* strains

To better understand how these plasmids were conserved across *S. hominis* strains an OrthoMCL analysis was performed to cluster the plasmid genes with any orthologous genes across the rest of the isolates. 171 proposed plasmid genes from the 4 putative plasmid contigs were grouped into 113 clusters of orthologous genes. The reduced number of clusters in comparison to genes would be expected in this analysis, not only due to the presence of families of similar genes such as those encoding replication proteins on multiple plasmids, but also due to the sequencing artefact of duplication of genes at the beginning and ends of the plasmid contigs. The presence or absence of these plasmid clusters of genes was then examined in the *S. hominis* genomes (Figure 3.9). All putative plasmid-encoding contigs were conserved across the similar isolates J6, J11, J23, J27, J31 and also K8 with more than 80% of gene clusters present in each isolate, with the exception of the small contig 4 which contained only 4 clusters.

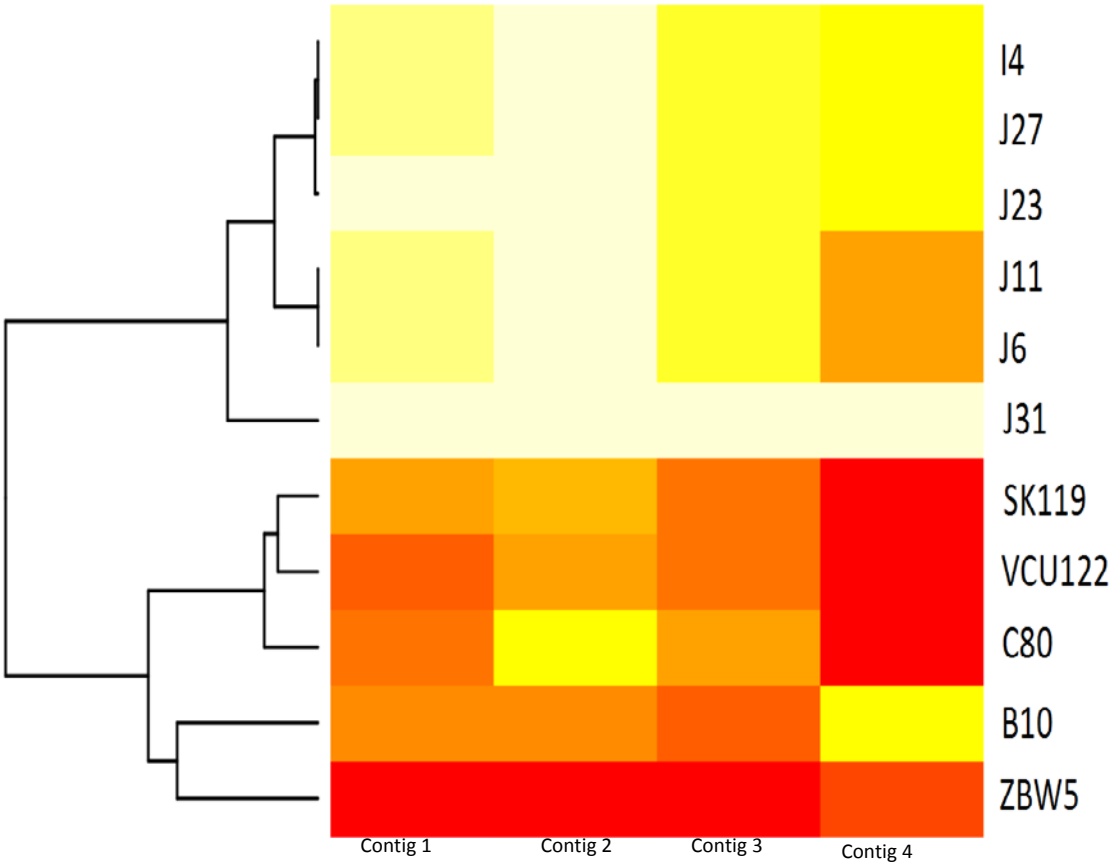


Figure 3.9 Conservation of plasmid associated genes The percentage of gene clusters from each putative plasmid contig determined by OrthoMCL present in the 11 *S. hominis* strains. Low percentage of plasmid gene clusters is indicated by white, a high percentage is indicated in red.

Of these sequences, contig 2 is proposed to encode a plasmid with a possible drug resistance efflux pump and was the most widely conserved with 8 strains sharing more than 50% of the genes clusters found on this contig (Figure 3.9).

Contig 3 is proposed to be a lantibiotic biosynthesis plasmid and is conserved only among the most similar isolates with the Liverpool "J" and K8 isolates posessing the *nisC* gene and sharing over 80% of the gene clusters. Strain C80 shares 50% of the gene clusters of plasmid contig 3 but does not contain the lantibiotic biosynthesis operon.

3.3.15.3 Transposable elements

Transposable elements and insertion sequences are ubiquitous features of prokaryotic genomes. They are credited as a source of expanded genome plasticity due to their involvement in recombination between sections of genomes and their facile and frequent mobilisation between species (Mahillon & Chandler 1998). One meta-analysis, however, disputed this claim and found no correlation between the number of IS sequences and the frequency of recombination across a range of species (Pál & Papp 2013). Transposons contain genes directing their insertion and excision and this is not discriminate, occuring on the chromosome, plasmids or even within other transposable elements. These elements can undergo homologous recombination with other transposable elements thus forming novel structures (Roberts *et al.* 2008). Insertion sequences are short sections of DNA, which carry only the genes required for their mobilisation. Movement of transposable elements and insertion sequences into a new genomic location can be traumatic for the local architecture resulting inactivation or upregulation of genes in the vicinity. In general, insertion events are considered to be detrimental to the host genome, a consideration corroborated by the finding that larger genomes contain the highest frequency of insertion element and the lowest density of essential genes thus reducing the risk of a fatal insertion event (Touchon & Rocha 2007). They can, however produce beneficial mutations that increase fitness which is one theory as to why selection maintains them within a genome despite their huge potential for deleterious effects (Fehér *et al.* 2012).

In *S. hominis*, as with most other species of bacteria, the presence of several families of insertion sequences was detected (Table 3.8). The insertion sequence families IS1182 and IS6 were conserved across all isolates in this study. The insertion sequence IS257, an IS6 family insertion sequence also known as IS431, is conserved across all *S. hominis* isolates in this study. This insertion element is frequently associated with antibiotic resistance genes and carries a promoter region able to drive expression of drug resistance genes such as *tetA* more strongly than the native promoter in *S. aureus* (Simpson *et al.* 2000). Among human-associated staphylococci IS257-like insertion sequences of the IS6 family such as IS257, IS431 and ISSau10 were revealed to play a role in the integration of a multitude of smaller mobile genetic elements into larger mobile genetic elements or into the chromosome itself (Schwarz *et al.* 2011). Plasmid encoded *ermC,* for example, was shown to integrate into the chromosome in an IS257-dependent manner (Diep *et al.* 2006). Collectively, the IS6 family were described to be closely linked with the integration of resistance genes into the staphylococcal genome (Schwarz *et al.* 2014).

IS256, as previously discussed, is a marker for clinical invasiveness in *S. epidermidis* due to its association with the *icaADBC* gene locus and multidrug resistance (Yao *et al.* 2005) (Montanaro *et al.* 2007). This insertion sequence was identified in the multidrug resistant isolate, *S. hominis* ZBW5. The IS3 family and the IS30 family of insertion sequences were also present in the isolates B10, J31, SK119 VCU122 and ZBW5 (Table 3.8). The Insertion sequences present in the *S. hominis* isolates have origins in a range of staphylococcal species as well as being present in other genera of bacteria highlighting the capacity of IS elements to be agents of genetic diversity and genome plasticity and their effective lateral transfer capability.

Table 3.8 Summary of the insertion sequences present across the *S. hominis* isolates		
Genome	**IS family**	**IS family origin**
B10	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
	IS6	*S. aureus*
	IS3	*S. aureus*
J6	IS6	*S. aureus*
	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
J11	IS6	*S. aureus*
	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
J23	IS6	*S. aureus*
	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
J27	IS6	*S. aureus*
	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
J31	IS30	*S. aureus*
	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
	IS3	*S. aureus*
	IS6	*S. aureus*
SK119	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
	IS3	*S. aureus*
	IS6	*S. aureus*
VCU122	IS1182	*S. aureus, S. epidermidis, S. haemolyticus*
	IS3	*S. aureus*
	IS6	*S. aureus*
ZBW5	ISL3	*S. haemolyticus*
	IS256	*S. aureus, E. faecealis*
	IS6	*S.aureus*
	IS1182	*S. epidermidis*

3.3.15.4 Bacteriophages in the *S. hominis* genome

Bacteriophages were identified using the PHAST phage finder program which assigns putative phage regions as intact, questionable or incomplete prophages according to criteria, including the number of coding sequences in the region, and the presence of characteristic phage genes such as tail and tape measure proteins (section 3.2.9). Several intact, questionable and incomplete prophage (Table 3.9) were found across the isolates of *S. hominis.*

Table 3.9 Summary of the prophage present in the *S. hominis* isolates. Incomplete phages were those with a completeness score below 60, questionable phage were those with a completeness score between 60 and 90. Complete phages were those with a score above 90.

Isolate genome	Genome position	Region length	Phage completeness	Designation	Possible phage
B10	83326-131512	48186	140	Intact	*Staphylococcus* phage StB20 (NC_019915)
	295758-315091	19333	70	Questionable	*Clostridium* phage phiMMP04 (NC_019442)
	1506131-1532195	26064	88	Questionable	*Staphylococcus* phage PT1028 (NC_007045)
C80	43650-64723	21073	30	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	269721-308532	38811	50	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	512225-536746	24521	150	Intact	*Lactobacillus* prophage Lj965 (NC_005355)
	1486948-1522307	35359	81	Questionable	*Staphylococcus* phage PT1028 (NC_007045)
J31 (PacBio assembly)	389028-402406	133378	20	Incomplete	*Staphylococcus* phage phiRS7 (NC_022914)
	1362133-1380156	18023	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1781926-1810085	28159	40	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
	2217429-2238965	21536	40	Incomplete	Cyanophage Syn30 (NC_021072)
J6	374212-387475	13263	20	Incomplete	*Staphylococcus* phage phiRS7 (NC_022914)
	1330793-1348816	18023	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1749977-1778136	28162	40	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
J11	374213-387492	13279	20	Incomplete	*Staphylococcus* phage phiRS7 (NC_022914)
	1330589-1348612	18023	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1749634-1777793	28159	30	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
J23	374048-387311	13263	20	Incomplete	*Staphylococcus* phage phiRS7 (NC_022914)

	1330889-1348912	18023	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1749653-1777812	28159	30	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
J27	373945-387213	13268	20	Incomplete	*Staphylococcus* phage phiRS7 (NC_022914)
	1331029-1349052	18023	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1750132-1778291	28159	30	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
I4	374220-387635		20	Incomplete	*Staphylococcus* phage phiRS7 (NC_02291)
	1331466-1349489		20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1749754-1777913		40	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
	2201825-2242040		109	Intact	*Enterococcus* phage phiX174 (NC_001422)
SK119	1323456-1337222	13766	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1459791-1496741	36950	88	Questionable	*Staphylococcus* phage PT1028 (NC_007045)
VCU119	1472455-1490085	17630	20	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
ZBW5	502486-517216	14730	150	Intact	*Staphylococcus* phage CNPH82 (NC_008722)
	1341869-1367471	25602	10	Incomplete	*Staphylococcus* phage vB_SepiS-phiIPLA7 (NC_018284)
	1494188-1515035	20847	30	Incomplete	*Staphylococcus* phage PT1028 (NC_007045)
	1787353-1806571	19398	30	Incomplete	Invertebrate iridescent virus 30 (NC_023611)
	2253755-2328024	74269	140	Intact	*Staphylococcus* phage StB20 (NC_019915)

An incomplete phage with homology to the staphylococcal phage PT1028 was found at a chromosomally conserved location across all of the isolates. In those isolates where Ebh similarity regions were separated by between ~14 and ~17 kb (B10, C80, SK119, VCU122, ZBW5 Table 3.5) the PT1028 incomplete phage was found to be in the intervening region (Figure 3.10 A). In the isolates found to have an intact Ebh similarity region 2 (J6, J11, J23, J27, J31, K8, Table 3.5) the PT1028 incomplete phage was found upstream of this region (Figure 3.10 B). The length of the questionable PT1028 phage, however do not fit within the base range separating region 2a and 2b for the strains B10, C80 SK119 and ZBW5 indicating that the boundaries of the questionable and incomplete phage may not be accurate. However, this data does suggest that the separation of Ebh regions 2a and 2b in the strains B10, C80, SK119, VCU122 and ZBW5 may have been a result of a phage insertion, and that the phage genome may have subsequently decayed thus the designation as a questionable phage.

The intact phages were examined in detail, and the gene content investigated (Table S1). B10 phage 1 and ZBW5 phage 5 are intact prophages with homology to the staphylococcal phage STB20. Nucleotide sequence alignment of the B10 and ZBW5 phages revealed 62% and 64% identity with STB20 respectively, and 62% identity to one another. As previously mentioned *aacA-aphD* was found to be within the annotation of ZBW5 intact phage 5 however, despite the high degree of sequence identity between the B10 phage 1 and ZBW5 phage 5, *aacA-aphD* is absent from the

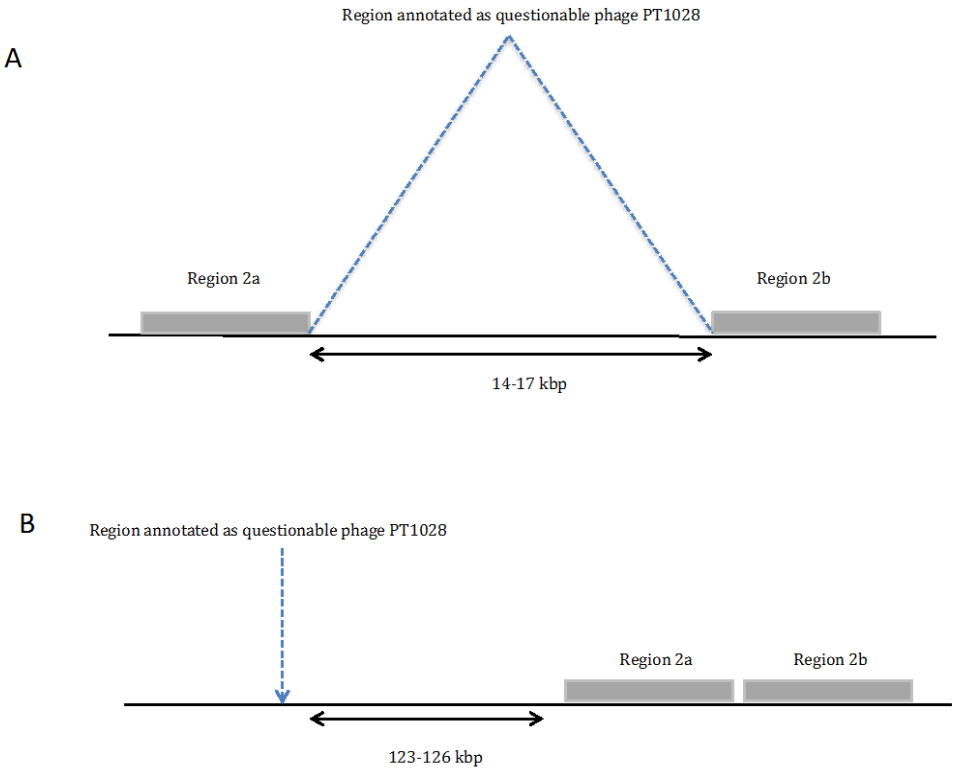


Figure 3.10 Schematic representation of the two possible structures of *ebh* region 2. Structure A is present in isolates B10, C80, VCU122 and ZBW5. Structure B is present in isolates J6, J11, J23, J27, J31, I4 and SK119

annotation of B10 phage 1. This is consistent with the analysis of antibiotic resistance genes (Table 3.6).

ZBW5 phage 5 also showed 68% sequence identity to StB27 (Accession number JN700519). Mauve alignment showed a region conserved between ZBW5 phage 5 and STB27 (Figure 3.11). Investigation of the gene content of this region using ACT revealed the presence of essential phage proteins such as the tail tape measure protein, and other structural proteins. The ZBW5 phage 5 also has an extended length of 74,269 bp in comparison to the 40071 bp of the phage STB27 (Figure 3.12).

The genes present in these additional bases include cargo genes such as the *mecR1* gene for meticillin resistance. The *mecR1* gene is often present in conjunction with the ACME element, which contains an arginine deiminase

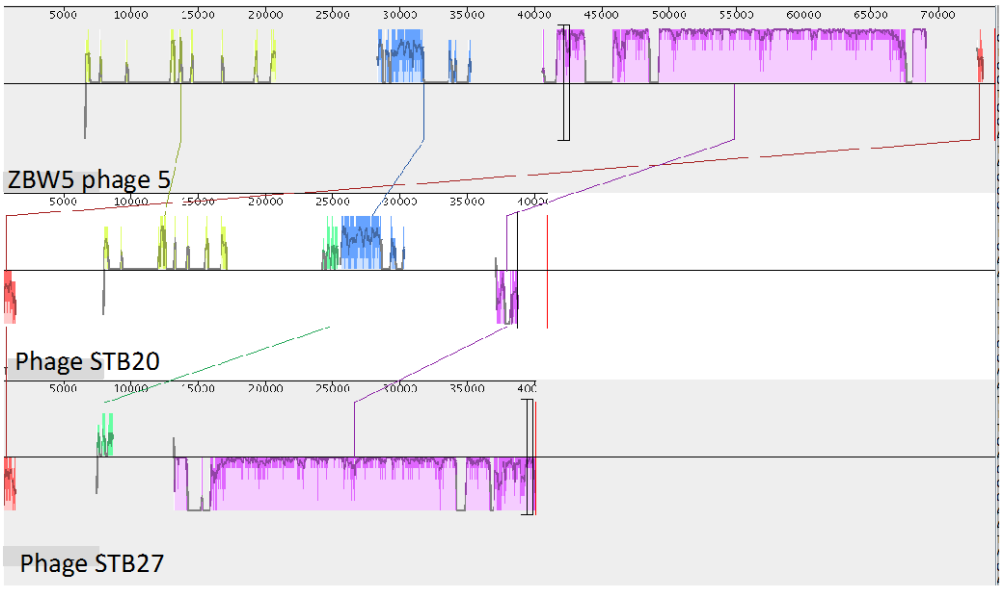


Figure 3.11 Mauve alignments of ZBW5 phage 5 with STB20 and STB27. The ZBW5 genome region annotated as the intact phage 5 by PHAST was aligned to the complete sequences of the phages STB20 and STB27.

gene. 67.9% of methicilin-resistant *S. epidermidis* strains carry ACME (Barbier et al. 2011), however only one copy of the *arcA* gene encoding an arginine deiminase was found in the ZBW5 genome, and this gene was not within the annotation of phage 5. The SCC*mec* element is mobilised *mecR1* in conjunction with the ACME element which in turn introduces an additional copy of *arcA* to the genome. As previously discussed (section 3.3.14) the *mecR1* gene is present in ZBW5 on a genomic element similar to *SCCmec* IV. A study by Scharn *et al* revealed that smaller SCC*mec* elements, such as type IV elements, can be mobilised by transduction, and that rearrangements and deletions in the SCC*mec* element can occur as a result (Scharn et al. 2013).

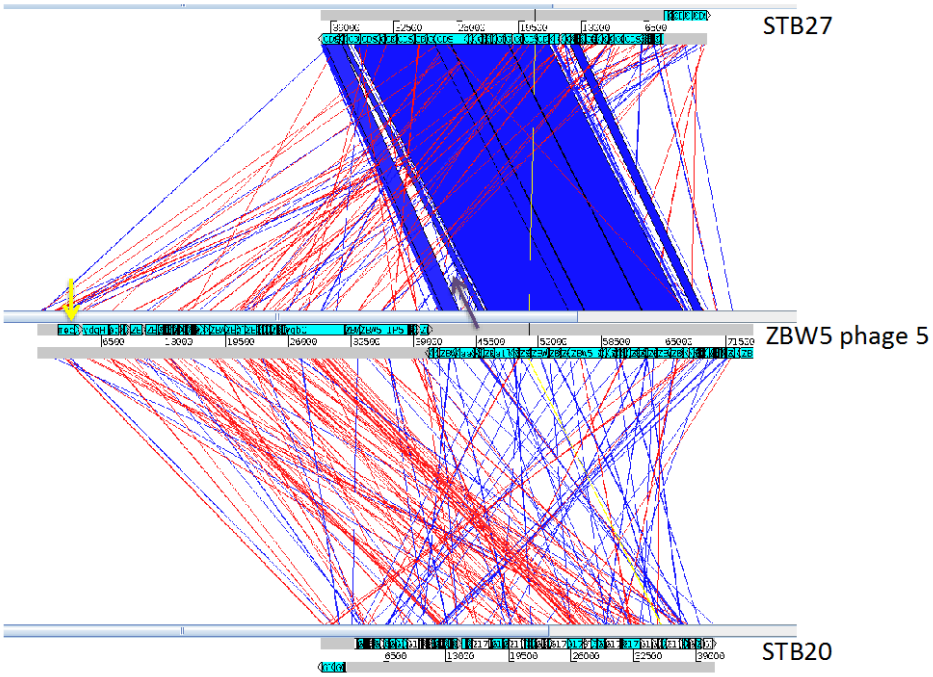


Figure 3.12 Artemis genome alignments between ZBW5 phage 5, STB27 and STB20. The yellow arrow indicates the position of the gene *mecR1*, the purple arrow indicates the position of the *aacA-aphD* gene.

Although not found within the annotation of this phage (Table S1) the loci responsible for tobramycin and kanamycin resistance, bleomycin resistance and streptrogramin a resistance in the ZWB5 multidrug resistant isolate were all found at the boundary of intact prophage region 5 (Figure 3.13) . The *attR* site, however is annotated a further 5,875 bases upstream of this (Table S1). This is further evidence that this region is not well resolved as *attL* and *attR* sites would be expected to flank the boundaries of a phage location.

Figure 3.13 Boundary of the annotated ZBW5 phage 5. The nucleotide highlighted yellow marks the bioinformatically annotated boundary of the genome region annotated as phage 5. The genes responsible for bleomycin (*ble*), streptogramin A (*ermC*) and tobramycin and kanamycin (*knt*) resistance are shown adjacent to the phage region.

3.4 Conclusion

This research study aimed to take advantage of developments with next generation sequencing technologies to interrogate *S. hominis* as a species. The data derived due to the increased resolution of newer technologies such as PacBio sequencing is able to increase the resolution and accuracy to better understand potential gene traffic between genomes. Specifically, it can better resolve structures of phage and plasmids than earlier technologies such as 454 and Illumina (Jørgensen *et al.* 2014). The addition of several more *S. hominis* genomes to those publically available has allowed the investigation of the *S. hominis* pan-genome and a wider understanding of the complement of genes present within the species. The *S. hominis* pan-genome was found to be open and contain ~1879 genes and be expanding at a rate of 41 news genes per each genome added. It is expected that the size of the pan-genome will fall as new genomes are sequenced and analysed. Greater than 60% of the pan-genome comprises accessory genes, yet only 17% of the genome of each strain is accessory genome. The large pool of accessory genes indicates that *S. hominis* may occupy a range of niches and overcome the specific obstacles to survival found there such as diverse community composition requiring adaptability.

The accessory genome is often associated with the presence of mobile genetic elements and for this reason emphasis was placed on characterising the repertoire of plasmids, phage, insertion sequence and transposons in the *S. hominis* genomes. All of the 11 genomes carried mobile genetic elements, some of which are associated with virulence in *S. aureus,* for example aminoglycoside and meticillin resistance. A putative SCC*mec* element was found in the genome of ZBW5, adding further evidence that CoNS have donated this MGE to *S. aureus* where it has acted as a virulence factor. The observed presence in the *S. hominis* isolates of biofilm forming capabilities, carotenoid production, a suite of CWA proteins, including putative MSCRAMM-like proteins and SERAMS, some of which are classical *S. aureus* virulence factors, invites further investigation of the roles of these genes in distinct backgrounds. Better understanding of their

role as survival factors may offer new perspectives on their role in staphylococcal virulence.

The diversity of antibiotic resistance determinants observed in this study, and their association with mobile genetic elements suggests that on skin *S. hominis* represents an important source of transferable antibiotic resistance. The expanded antibiotic resistance profile of the multidrug resistance isolate ZBW5 was possibly accounted for in the main by a single phage with homology to the StB20 staphylococcal phage. Spread of this phage to successful clones, as was seen in the case of the introduction of the PVL phage into the USA300 epidemic clone (Thurlow *et al.* 2012), could result in an increase in infection caused by *S. hominis*.

Not only has the mobilome been found to be a source of an expanded repertoire of genes, but also a potential for divergent profile of phenotypes. The ST1028 incomplete prophage was found to insert into two alternative genomic locations. One of these locations was between two fragments of a gene shown to have similarity to the giant ECM binding protein Ebh, which has been shown to be sufficient for maintaining a biofilm in *S. epidermidis* (Christner *et al.* 2010). The "J" isolates of *S. hominis* were all shown in this study to have the capability to form an air interface biofilm. These isolates also had the ST1028 prophage inserted upstream of the gene similar to Ebh.

Chapter 4

Staphylococcus Interspecies Comparative Genomics

4.1 Introduction

4.1.1 Interspecies comparative genomics

The field of comparative genomics is primarily centred upon comparing multiple DNA sequences and there have been numerous algorithms developed to meet the needs of different types of comparative genomic analyses. The field took roots in the late 1970s with the completion of the first two whole genomes, the phages ϕX174 and MS2, however with limited available sequence data and two largely unrelated sequences this was a stilted beginning (Koonin & Galperin 2013). One of the early published comparative genomic studies, a comparison of the Epstein-Barr virus and varicella-zoster virus, dealt with many of the themes integral to comparative genomics such as genome architecture, conservation and prediction of genes (Davison & Taylor 1987).

The field evolved and perhaps the best known and most ubiquitously used alignment tool, BLAST, was developed to allow rapid searching of the growing databases of protein and nucleotide sequences on a desktop computer (Altschul *et al.* 1990). From this beginning there followed BLASTZ, the first algorithm geared towards whole genome alignments with reasonable computational requirements and timescales (Mullikin 2014) (Schwartz *et al.* 2003). This algorithm was adapted for the purpose of comparing larger mammalian genomes from the web based PipMaker, a workbench which allowed comparisons of complete or incomplete genomes, with the aim of coding sequence and regulatory element prediction in mammalian and bacterial genomes (Schwartz *et al.* 2000). Today there is an array multiple genome alignment software publically available, such as Mauve (Darling *et al.* 2004), the Artemis Comparison Tool (ACT) (Carver *et al.* 2005) and the Integrative Genomics Viewer (IGV)(Robinson *et al.* 2011), which cater variously to the different requirements, questions and approaches explored in comparative genomics studies.

These tools facilitate mining the wealth of genome sequence data publically available in the wake of high throughput next-generation sequencing, which has

led comparative genomic studies to become a commonplace means of both asking specific questions and addressing broad themes. For example, Burgess-Herbert & Euling used comparative genomics for the latter purpose to assess the effects of extrapolating the potential impact of certain chemicals from one species to another, with particular emphasis on extrapolating the effects of pharmaceuticals from model organism to humans (Burgess-Herbert & Euling 2013). Their approach is also pertinent, however, when the human microbiome is considered. Understanding the interspecies comparative genomics of the pathogenic *S. aureus* relative to beneficial resident staphylococci, such as *S. hominis*, *S. haemolyticus* and *S. epidermidis* could help to extrapolate the effects of interventions across species. This kind of approach may help to develop interventions to clear *S. aureus* from the skin or manipulate its microflora without affecting the desirable members of the community.

Interspecies comparative genomics of other microbial communities have been carried out with the hope of gathering information that might allow the manipulation of these coummunities for the benefits of human health and prevention of disease. Kaoutari *et al* for example used comparative genomics to describe the diversity of carbohydrate active enzymes in the genomes of gut resident microbial species, noting that these microbial enzymes aid in the digestion of some of the complex polysaccharides in our diet, thus showing a healthy gut microbiome is integral to human health (Kaoutari *et al.* 2013). Comparative genomics has also been used to assess the integrity and safety of the food chain; Papadimitrou *et al* used comparative genomics to assess the risk of the *Streptococcus macedonicus* in fermented dairy products, given that it is part of the *Streptococcus bovis/Streptococcus equinus* complex implicated in endocarditis and colon cancer in humans (Papadimitriou *et al.* 2014).

4.1.2 Staphylococcal whole genome sequencing

The genus *Staphylococcus* comprises 51 species with 27 subspecies. The species of this genus have a broad host range with specialists colonising non-human primates (Tong *et al.* 2015), dolphins (Varaldo & Kilpper-Bälz 1988) and fermented food products (Probst *et al.* 1998). The genus continues to expand

with next generation sequencing technologies aiding the refinement of speciation for closely related staphylococci, such as *S. argenteus* and *S. schweitzeri* which are indistinguishable from each other and *S. aureus* at their 16S locus (Tong *et al.* 2015). At the time of writing over 4400 strains of *S. aureus* had been sequenced with over 4200 of these genomes publically available in a range of formats from scaffolds to complete genomes, including the important USA300 community-acquired MRSA strain (Diep *et al*, 2006).

The prolific sequencing of this notorious pathogen has increased understanding of many aspects of its genomic content, such as the acquisition of vancomycin resistance phenotypes which eliminate the availability of this last line antibiotic from the treatment of MRSA (Kos *et al.* 2012). More recently it was used to investigate the efficacy of employing whole genome sequencing as a means of determining these resistance phenotypes (Lee *et al.* 2015). Clinically relevant timescales for high-throughput whole genome sequencing have allowed the interrogation of the genomes of infection isolates and aided understanding of outbreaks of MRSA and its transmission in real time (Köser *et al.* 2012). Large scale sequencing studies have also begun to unravel the basis for the broad host range of *S. aureus* through their power to estimate timelines of host jumps and the genetic mechanisms for these events (Spoor *et al.* 2013) (Lowder *et al.* 2009) (Lowder & Fitzgerald 2010) (Viana *et al.* 2015).

Complete genomes have also been finished and published for 8 other staphylococcal species, including 9 strains of *S. epidermidis,* which represents the most common pathogen associated with infection of implanted medical devices. In total, the genomes of 44 species of the *Staphylococcus* genus have been sequenced, with 27 of these being available publically (Zakour *et al*, 2008 & data compiled from NCBI and GOLD online database). Although individual isolates, or small cohorts of isolates have been sequenced for most of the species in the genus it is still unclear to what extent these species interact and share genetic information with one another. These interactions are important in the context of human health and disease since several species of staphylococci,

S. epidermidis, *S. hominis*, *S. haemolyticus* among them, are known to be dominant members of the human skin microflora.

4.1.3 Features of staphylococcal genomes

To understand the interactions between staphylococcal genomes we must first understand the structure of the genomes themselves. Sequencing of multiple species of staphylococci elucidated broad similarities among the genomes. There is a conservation of genome size and synteny of hypothesised core genes in *S. haemolyticus*, *S. epidermidis* and *S. aureus* which was proposed following the sequencing of the first *S. haemolyticus* genome (Takeuchi *et al.* 2005). The presence of an *oriC* environ thought to represent the source of much of the divergence between staphylococcal species was proposed by Takeuchi et al in 2005. The described *orfX* region is near to the *oriC* origin of replication and this region serves as an integration site for multiple staphylococcal cassette chromosomes (SCCs). This region of the genomes is often observed to contain strain specific genes and genes associated with niche adaptation acquired through horizontal gene transfer (Hiramatsu *et al.* 2014) (Harrison *et al.* 2013).

Genome plasticity appears to be important to the success of staphylococci within their niche. This was revealed by the extensive sequencing of *S. aureus* and *S. epidermidis,* whose pan genomes comprise 37% and 56% strain specific genes respectively (Kelly 2013). In addition to SCCs, other mobile genetic elements such as plasmids, phages and transposons were identified across the staphylococci (Takeuchi *et al.* 2005) (Kuroda *et al.* 2005) and contribute to diversity; most frequently these elements contribute virulence factors (Moon *et al.* 2015) and generate resistance profiles to antibiotics and heavy metals, as highlighted in the study of the *S. hominis* genome in chapter 3 of this thesis. Non-pathogenic species of staphylococci were shown to contain fewer MGEs; the food grade species *S. carnosus* for example, was shown through comparative genomic studies to contain fewer MGEs than other members of the genus (Rosenstein & Goetz 2010). Not only does the non-pathogenic *S. carnosus* contain a comparatively low amount of MGEs it also has a smaller amount of repeat sequences, which may suggest fewer recombination events and thus

reduced genomic diversification (Aras *et al,* 2003). The low frequency of repeat sequences in the *S. carnosus* genome suggests it is less plastic than that of species such as *S. haemolyticus*, *S. epidermidis* and *S. aureus*. This supposed lack of genome plasticity of *S. carnosus* may be explained by its selection from a restricted niche within a starter culture (Rosenstein & Friedrich 2010) or spontaneously fermented foods (Anon 2008). Despite its limited plasticity and generally regarded as safe (GRAS) status, *S. carnosus* does contain genes with homology to *S. aureus* virulence genes. One possible explanation of this may be that genes classified as virulence factors and contributing to disease in pathogenic staphylococci might be involved in environmental survival mechanisms in other species. The non-pathogenic bacterium *Wolinella succinogenes*, which has a primary niche in the bovine rumen, shares multiple homologous genes with *Campylobacter jejuni* and *Helicobacter pylori* (Baar *et al.* 2003). Strikingly it harbours genes having homology and synteny with pVir, an important *Campylobacter* virulence plasmid (Bacon *et al.* 2000). This observation led to the suggestion that the virulence factors were involved in maintaining symbiosis between *W. succinogenes* and its host (Baar *et al.* 2003).

The arginine catabolic mobile element (ACME) and the SCC*mec* element represent evidence of interactions between staphylococci and the exchange of genetic material. It was proposed that the ACME element spread through horizontal gene transfer from *S. epidermidis* to *S. aureus*. This inheritance has been well-characterised and it was revealed that the recombinase required for recombination of the *SCCmec* element into the genome is hijacked by ACME (Miragaia *et al* , 2009).

Although several studies suggest that carriage of ACME in disease isolates of *S. aureus* does not affect virulence (Montgomery *et al*, 2009), a role with respect to colonisation has been speculated since ACME carriage does influence fitness with respect to skin colonisation (Diep *et al*, 2006). A role as a fitness determinant as opposed to a virulence factor is also supported by the presence of ACME in the genomes of other commensal staphylococci, such as *S. haemolyticus (*Pi *et al*, 2009*)*. Further evidence for ACME increasing survival on

the skin can be gleaned from the presence of an arginine deiminase pathway in *Streptococcus pyogenes* that promotes survival in a low pH environment (Holden *et al*, 2006).

A recent comparative genomic analysis of a large cohort of *S. aureus* and *S. epidermidis* isolates revealed that the two species rarely undergo homologous recombination with one another, but that there is a high degree of commixing mobile genetic elements between the two species. Recombination within each species appears to differ with isolates of *S. epidermidis* recombining more frequently than isolates of *S. aureus*.

4.1.4 Staphylococcal two component systems

Two component systems (TCSs) allow many prokaryotes to interact with their environment and react to the presence of stimuli. These two component systems comprise: a sensor histidine kinase which typically spans the cell membrane and interacts with the external environment; interaction with its target molecules resulting in conditional autophosphorylation; a response regulator, which is typically an intracellular protein that interacts with the histidine kinase sensor and is phosphorylated. The role of a two component system is to change the behaviour of a cell to compensate for changing environmental conditions, therefore it is to be expected that the internal regulator is often a DNA-binding protein which can effect altered gene expression (Stock *et al.* 2000).

TCSs are present across bacterial genomes to varying degrees and absent from certain species, such as *Mycoplasma genitalium*. For particular bacteria these systems comprise relatively large proportions of the genome, for example *Synechocystis sp* of which 2.5% of the total genome is accounted for by TCSs (Kolar 2012). The *S. aureus* genome contains 16 TCSs with target genes from an array of functional groups. Although TCSs are widespread across prokaryotes, there is one particular family specific to the Firmicutes, those containing an intramembrane-sensing histidine kinase (IM-HK). This family of HKs are smaller than those found in other species of bacteria and have a small, 25 amino

acid linker region between their two 400 amino acid transmembrane helices. Due to their minimal extracellular presence they were thought to sense signals occurring within the cell membrane (Mascher 2006). Subsequent research has recently suggested alternative mechanisms of action (Mascher 2014).

This relatively small Firmicutes family of histidine kinases can be divided into two subgroups. The first, HPK7, is defined by an additional transmembrane protein required to effect the downstream response. A well-studied member of this family of HKs is LiaS which is part of the LiaFSR three component system of *B. subtilis*. In the staphylococci, a well-studied example is the VraFSR system (Mascher 2006). This IM-HK is proposed to indirectly sense antibiotics by responding to the resulting cell membrane damage (Rietkötter *et al.* 2008). The second subgroup of IM-HK is HPK3i, which is defined by the presence of an associated ABC transporter. Research has shown that the histidine kinase itself does not play a direct role in sensing the stimulus which activates this class of two component system. Rather, it is binding of the substrate to the cognate ABC transporter (Rietkötter *et al.* 2008) and a possible physical interaction between the TCS and ABC transporter (Falord *et al.* 2012) that activates the TCS to upregulate the ABC transporter and consequently remove the substrate from the cell. This second family of IM-HK TCSs has therefore been shown to respond directly to the presence of antibiotic. Examples of this subgroup of IM-HK TCS are the GraXSR system which is present across the staphylococci and the well-studied BceRS system of *B. subtillis*.

The GraRS TCS of *S. aureus* has been examined in detail. GraRS is part of a global regulation network responsible for resistance to cationic antimicrobial peptides (CAMPs) (section 3.3.9.9). GraXSR regulates the expression of DltABCD and MprF which both alter the *S. aureus* surface charge to evade electrostatic interaction-mediated interaction with CAMPs (Falord *et al.* 2011). Furthermore, the autolysin *SsaA* that is also involved in cell wall homeostasis is upregulated by GraRS. Transcriptionally, *ssaA* is regulated by the essential TCS WalKR (Delaune *et al.* 2011). In addition to these loci VraSR expression is controlled by GraXSR (Meehl *et al.* 2007). The GraRS regulated protein SdrH belongs to the

serine aspartate repeat (SDR) family of proteins in *S. aureus*; cell wall-associated proteins which contain serine-aspartate tandem repeats located at the C-terminal of the protein. Several members of this SDR protein family interact with the host extracellular matrix, for example SdrG of *S. epidermidis* which binds to fibrinogen (Ponnuraj *et al.* 2003). Differential regulation of multiple protein families demonstrates the complexity of TCS regulatory networks and highlights the challenges involved in deciphering their interconnected roles.

An orthologous TCS to GraRS was described in *S. aureus* and designated BraRS and NsaRS by two different groups concurrently (Hiron *et al.* 2011) (Blake *et al.* 2011). Blake *et al* discovered the role of the uncharacterised protein SAOUHSC_02955 in lantibiotic resistance through an evolution based approach where resistance to nisin was selected by serial passage in sub-MIC concentrations of the lantibiotic. This approach, and subsequent genome sequencing of the evolved isolate revealed two SNPs, one of which was associated with the gain of function and was located in the putative sensor histidine kinase SAOUHSC_02955, thus it was designated NsaSR or nisin susceptibility-associated sensor regulator (Blake *et al.* 2011).

Hiron *et al* revealed that the same TCS was characteristic of the Bce-like family of IM-HKs which are involved in cationic antimicrobial peptide resistance, and that analysis of gene mutants showed that lack of the TCS resulted in increased susceptibility to bacitracin. The TCS was thus designated BraSR for bacitracin resistance-associated sensor regulator. Susceptibility of the knockout mutant to a panel of antibiotics was investigated, however only a decrease in MIC for bacitracin and the lantibiotic nisin was observed (Hiron *et al.* 2011). BraRS binding sites were discovered upstream of *braDE* and *vraDE* , two ABC transporter genes induced in the presence of bacitracin, but not in a *braRS* mutant. The BraDE and VraDE transporters were shown to play different roles in antibiotic resistance. BraDE contributes to the detection of nisin and bacitracin and subsequent signal transduction via BraRS, whereas VraDE is more directly involved in detoxification (Hiron *et al.* 2011).

A later study by Kolar *et al* showed that *braRS* (*nsaRS)* transcription was upregulated by the presence of multiple antibiotics such as ampicillin, phosphomycin and nisin. An allelic replacement mutant of *braS* (*nsaS)* altered the transcription of 245 genes in a transcriptomic analysis (Kolar *et al.* 2011). This study revealed multiple effects of loss of BraRS (NsaRS) function in *S. aureus,* such as effects on biofilm formation, regulation of cellular transport and responses to anaerobic conditions. The size of the reported regulon may be an overestimate due simply to the overlapping circuitry reporting stimuli at the cell surface and a more detailed study is certainly required to map the regulon more precisely.

The potential complexity of the BraRS regulon indicates that, although multiple TCSs in Firmicutes detect damage and changes in the cell membrane, there are still others with more diverse functions. The well-studied Agr quorum sensing system involves AgrC, as the histidine kinase, and AgrA, as the DNA binding response regulator (Koenig *et al.* 2004). Still other TCSs deal with the toxicity of metal ions such as HssRS which regulates the expression of the efflux pump HrtAB (Stauff *et al.* 2007). Although the TCSs of *S. aureus* are well studied and the mechanisms of action and global regulons are being unravelled, it is unclear to what extent these TCSs are distributed throughout the genus.

4.1.5 Chapter Aims

In recent times, efforts have been made to understand the genetic aspects of the staphylococci which make them successful as infective agents. As described, *S. epidermidis* and *S. aureus* represent two major, but distinct human pathogens, consequently work has been focused on the genomes of these two species to uncover mechanisms relating to disease. In this chapter the study sought to use interspecies comparative genomics to investigate how *S. hominis* relates to other members of the *Staphylococcus* genus. Furthermore the study sought to investigate the ways in which members of this genus broadly differ and resemble one another in their mechanisms to thrive in their niche and within their communities. Particular emphasis was given to exploring mechanisms which might pertain to being successful skin residents.

4.2 Methods

4.2.2 Analysis of orthologous gene content across the staphylococci

A representative genome from all staphylococci sequenced at the time of analysis was either retrieved from the NCBI FTP (ftp://ftp.ncbi.nlm.nih.gov/) site or sequenced as part of this study. Complete genomes were used as far as possible, however where only draft genomes were available NCBI scaffolds were reordered against an appropriate reference using the perl script written by Dr Roy Chaudry (CGR). The base set of strains used in the analysis is described in Table 4.1. The genomes were annotated using PROKKA (version 1.5.2) to ensure coherent annotation. OrthoMCL with the parameters stated above was used to cluster the orthologous proteins.

Table 4.1 List of base strains for the OrthoMCL analysis. All genomes were retrieved from NCBI or sequenced as part of this work, the draft or complete nature of the genome is indicated.

Species	**Strain**	**Accession**	**Draft/complete**
S. arlettae	CVD059	Uid175126	Draft
S. aureus	Newman	Uid58839	Complete
S. capitis	SK14	Uid55415	Draft
S. carnosus	TM300	Uid59401	Complete
S. cohnii	(this work)	PRJEB10525	Draft
S. delphini	8086	Uid199664	Draft
S. epidermidis	ATCC_12228	Uid57861	Complete
S. equorum	Mu2	Uid169178	Draft
S. haemolyticus	(this work)	PRJEB10525	Draft
S. hominis	J6 (this work)	PRJEB10524	Draft
S. intermedius	NCTC_11048	Uid199665	Draft
S. lentus	F1142	Uid200144	Draft
S. pettenkoferi	VCU139	Uid180074	Draft
S. lugdunensis	HKU09	Uid46233	Complete
S. pseudintermedius	HKU10	Uid62125	Complete
S. saprophyticus	ATCC_15305	Uid58411	Complete
S. simiae	CCM_7213	Uid77893	Draft
S. simulans	(this work)	PRJEB10525	Draft
S. vitulinus	F1028	Uid200114	Draft
S. warneri	SGI	Uid187059	Complete
S. xylosus	(this work)	PRJEB10525	Draft

The python script *Table_alt_cluster-1.py* was used to create a table describing the presence or absence of each OrthoMCL cluster within each genome. This table was then converted to a data matrix using the statistical package R and a heatmap was generated from the matrix. To control for gross strain specific effects on the clustering this step was repeated after substituting with the strains described in Table 4.2 for an alternative species strain from Table 4.1. All permutations were carried forward and used in subsequent steps of the analysis.

Table 4.2 List of alternative strains. Strain substitution was used to rule out strain specific effects in species as a driver for the OrthoMCL clusters and their genetic basis.

Species	**Strain**	**Accession**	**Draft/complete**
S. aureus	USA300	Uid58555	Complete
S. epidermidis	RP62A	Uid57663	Complete
S. hominis	J11 (this work)	PRJEB10524	Draft
S. pseudintermedius	ED99	Uid162109	Complete
S. saprophyticus	CCM_833 (this work)	PRJEB10567	Draft

4.2.3 Identifying the drivers of OrthoMCL group formation

The R library randomForest was used to investigate the genetic basis directing the formation of each of the OrthoMCL groups. The output from the script *Table_alt_cluster-1.py* was edited to designate each species from the analysis as either belonging to the cluster being investigated or not belonging to the cluster. This information was then used as the input data for the randomForest analysis. The output analysis was summarised using the variable importance plot function and the mean square error resulting from the permutation of each variable and was used as the measure of importance. Clusters were mapped back to the genome and the annotation and protein sequence for a species representative from each cluster were retrieved. Protein sequences were then used to search the NCBI BLAST database to confirm the PROKKA genome annotation. Outputs from substituted strains for each cluster were compared and the genes conserved in the variable importance analysis for all combinations of strains within a cluster were collated and carried forward for subsequent analysis.

Artemis (version 14.0.0)was used for genome interrogation (Rutherford *et al.* 2000) and BLAST and CLustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) were used to perform sequence alignments.

4.2.4 Investigating the diversity of loci associated with CAMP resistance in the staphylococci

The BraR, BraS, GraR and GraS protein sequences were retrieved from the relevant genomes in the analysis and protein sequence alignments were carried out using ClustalW2.

All neighbour-joining trees were generated using Geneious global alignment with the Blosum62 cost matrix followed by the geneious tree builder and a consensus tree based on 1000 bootstrap repitions (Geneious version 8.1.7).

4.3 Results and Discussion

4.3.1 Analysis of orthologous gene content across the staphylococci

The orthologous gene content of one genome from each of the available genomes of sequenced staphylococcal species was analysed using OrthoMCL, which groups orthologous genes (homologues separated by speciation) from across the different species into clusters. The number of shared orthologous clusters between the different species' genomes was represented as a heatmap (Figure 4.1). The output from this analysis revealed the assembly of three groups of species with high numbers of shared orthologous clusters. The cladogram accompanying the heatmap also reinforced these three groups which were therefore labelled as A, B and C. In addition to these three groups, there were also several pairs of species which showed a high degree of shared orthologous clusters of genes, and branched together in the cladogram. These pairs of strains were *S. aureus* and *S. simiae, S. simulans* and *S. carnosus,* and *S. lentus* and *S. vitulinus.* The effect of substituting alternative strains of individual species and rerunning the OrthoMCL analysis was investigated and did not affect the three groups of staphylococci (Supplementary figures). The similarity and grouping of the species *S. simulans* and *S. carnosus* in the cladogram is to be expected as *S. carnosus* was originally identified as *S. simulans*. The percentage DNA similarity between the type strain of *S. simulans* was found to be more similar to *S. carnosus* (32-39%) than to any other species in the genus; this is observed with the homologous gene content of the two species which cluster together in the OrthoMCL heatmap, and group together in the cladogram (Figure 4.1). The phenotypic differentiation of *S. carnosus* as a species from *S. simulans* is based upon the ability of *S. carnosus* to produce acetoin, and its inability to ferment sucrose, lactose and maltose, whereas *S. simulans* can ferment these sugars. Futhermore, *S. carnosus* can be differentiated from all other staphylococci by its strongly positive growth in high salt concentrations (Schleifer & Fischer 1982).

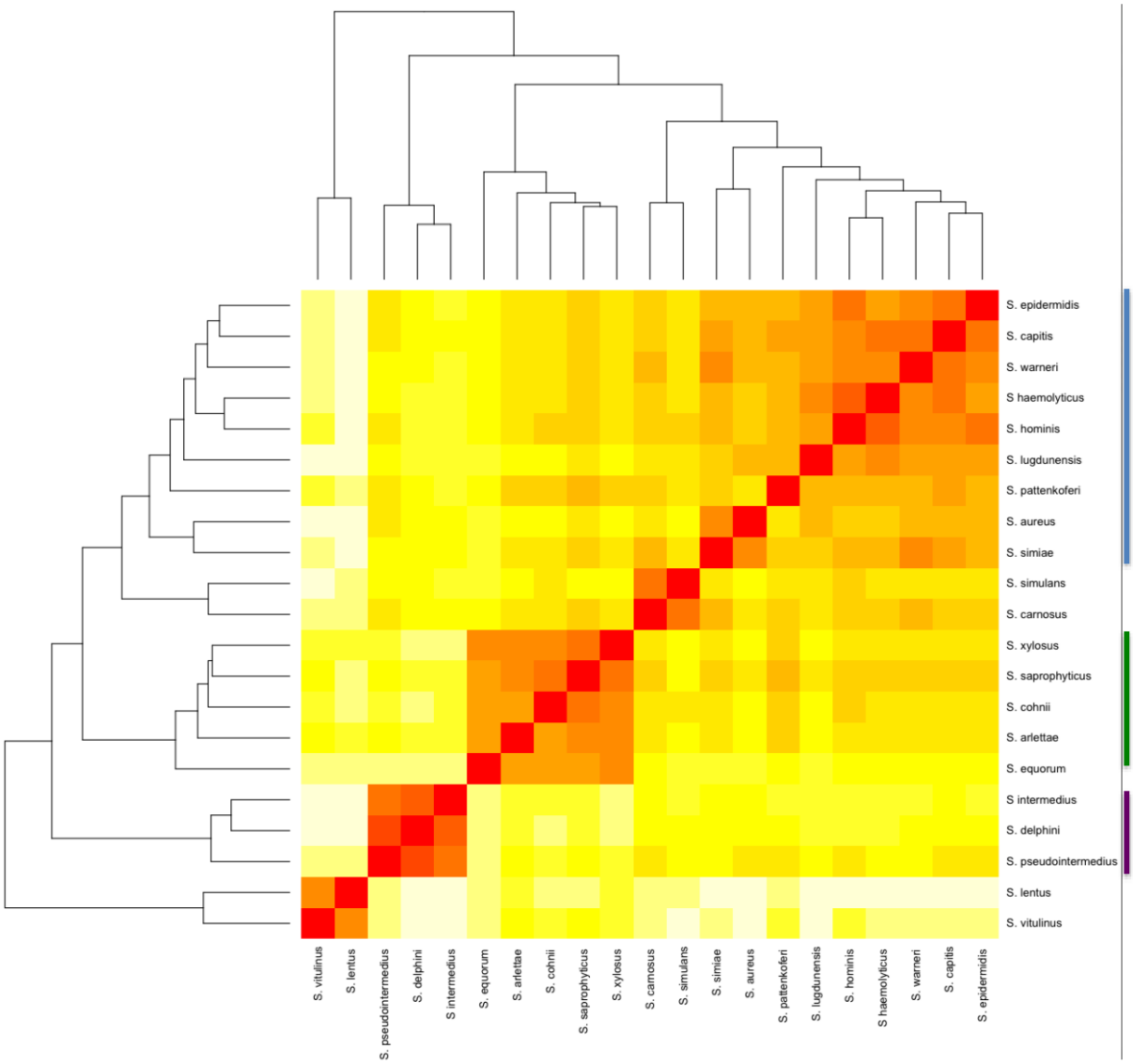


Figure 4.1 Heat map representation of the presence of clustered orthologous proteins across the staphylococci. Presence is indicated on a colour scale from red (high) to white (low). A high number of shared clusters of orthologous proteins is indicated by red colouration and few shared clusters is indicated by white. The three major groups of species observed in the analysis are highlighted by coloured lines; Group A is indicated in blue, B is indicated in green and C is indicated in purple.

S. lentus and *S. vitulinus* would also be expected to cluster together in the OrthoMCL analysis as these species comprise two of the four species in the *Staphylococcus sciuri* group, the other two species being *S. sciuri* and *S. pulvereri* (Stepanovic *et al.* 2003). These species are considered to be primarily animal species, however colonisation in humans and isolation from clinical samples has been observed (Nagase *et al.* 2002). The *S. sciuri* group are differentiated from other staphylococci as they are positive for cytochrome c oxidase activity. This aspect of their genome makes them the most closely related group of staphylococci to the micrococci (Kwok & Chow 2003).

The largest, although least well-defined, group was comprised of *S. epidermidis*, *S. capitis*, *S. warneri*, *S. haemolyticus*, *S. hominis*, *S. lugdunensis*, *S. pettenkoferi*, *S. aureus* and *S. simiae*. This set was designated as group A (Figure 4.1). It was noted that group A contained the species that most commonly occur on the human skin (Kloos 1980). The likelihood of a strain-dependent effect structuring the output was investigated by substituting *S. epidermidis*, *S. hominis* and *S. aureus* with alternative strains (Table 4.2). The same structure of groups was observed in each of the instances of switching strains indicating that there was a robustness to this output independent of strain choice (Supplementary figures). Within group A, *S. aureus* and *S. simiae* were most similar to one another and branch together in the cladogram. *S. aureus* and *S. simiae* belong to the the *S. aureus* group of staphylococci, which includes only these two species. Investigation of *Staphylococcus* phylogeny using only core genes revealed the clustering of *S. aureus* and *S. simiae* (Suzuki *et al.* 2012). In the analysis by Suzuki *et al,* however, this grouping broke down upon hierarchical clustering based on whole genome content (Suzuki *et al.* 2012), whereas in this present study branching of both *S. aureus* and *S. simiae* was observed based on orthologous gene content of the whole genome. It is to be noted that the methodologies and strain sets used by the two studies differed. It was decided to include the species *S. aureus* and *S. simiae* within group A for subsequent comparative analyses, since by taking into account both the cladogram and the heatmap these two species are more related to the other species within group A than those outside of it. Downstream analysis, however, involved investigating

the genetic basis driving the groups of species. Inclusion of *S. aureus* in group A, which also contains those frequent members of the human skin microbiome, such as *S. epidermidis, S. hominis* and *S. haemolyticus*, could reveal important similarities between *S. aureus* and the skin resident species; exclusion of *S. aureus* from group A would result in these similarities being missed. Equally, inclusion of *S. aureus* and *S. simiae* in group A would result in differences from the skin resident species *S. epidermidis, S. hominis* and *S. haemolyticus* being missed. Therefore group A was also divided into two subgroups, A_1 and A_2. A_1 contained the species in group A without the inclusion of *S. aureus* and *S. simiae.* A_2 contained included only these two. Downstream analysis was carried out first with only subgroup A_1 and then expanded to include subgroup A_2; the analysis of both subgroups together is discussed simply as group A.

A second, smaller group of species from the cladogram/heatmap was comprised of *S. equorum*, *S. arlettae*, *S. cohnii*, *S. saprophyticus* and *S. xylosus*; this was designated as group B. A common lifestyle identified in the group B species is their association with animal and meat products plus the presence of novobiocin resistance (Devriese *et al.* 1985) (Bannerman *et al.* 1991) (Schleifer *et al.* 1984). Furthermore, commonalities in the compositions of their cell walls was observed (Schleifer *et al.* 1984).

A third species group was comprised of *S. pseudintermedius*, *S. delphini* and *S. intermedius* and was labelled as group C. This set of organisms has previously been designated the *S. intermedius* group, or SIG, and is known to cause infection in companion animals and equids, with the emerging antibiotic resistance in this group of organisms representing a clinical veterinary concern (Stull *et al.* 2014). Routine speciation within this group is difficult due to the high degree of sequence identity at their 16S locus (Slettemeås *et al.* 2010).

To investigate the genetic basis directing the formation of these groups of species (A, B and C) the randomForest package was used in R. This analysis employs the randomForests algorithm of Leo Breiman for classification (Breiman 2001). This algorithm ranks the importance of variables, in this case

OrthoMCL clusters, for their contribution to the presence of groups A, B and C based on a forest of trees generated from these variables. A variable importance plot of each random forest was generated which plots the scaled importance of the variables in order of importance from the top of the plot to the bottom. The measure of importance used here was %IncMSE, which refers to the mean squared error resulting from the permutation of the variables. Permuting an important variable will result in a large mean squared error, with decreasing importance of variables resulting in decreasing mean squared error. Variables are found on the y-axis and their importance score on the x-axis, an example plot is included (Figure 4.2).

The OrthoMCL clusters representing each variable were then examined and mapped back to the genome of a publically available reference for each cluster and the PROKKA annotation of each protein coding sequence was verified using BLAST. This was repeated for each group and for each of the strains used to verify the presence of the three clusters (Table 4.2). The genes conserved in the variable importance analysis between substituted strains are discussed in the following sections.

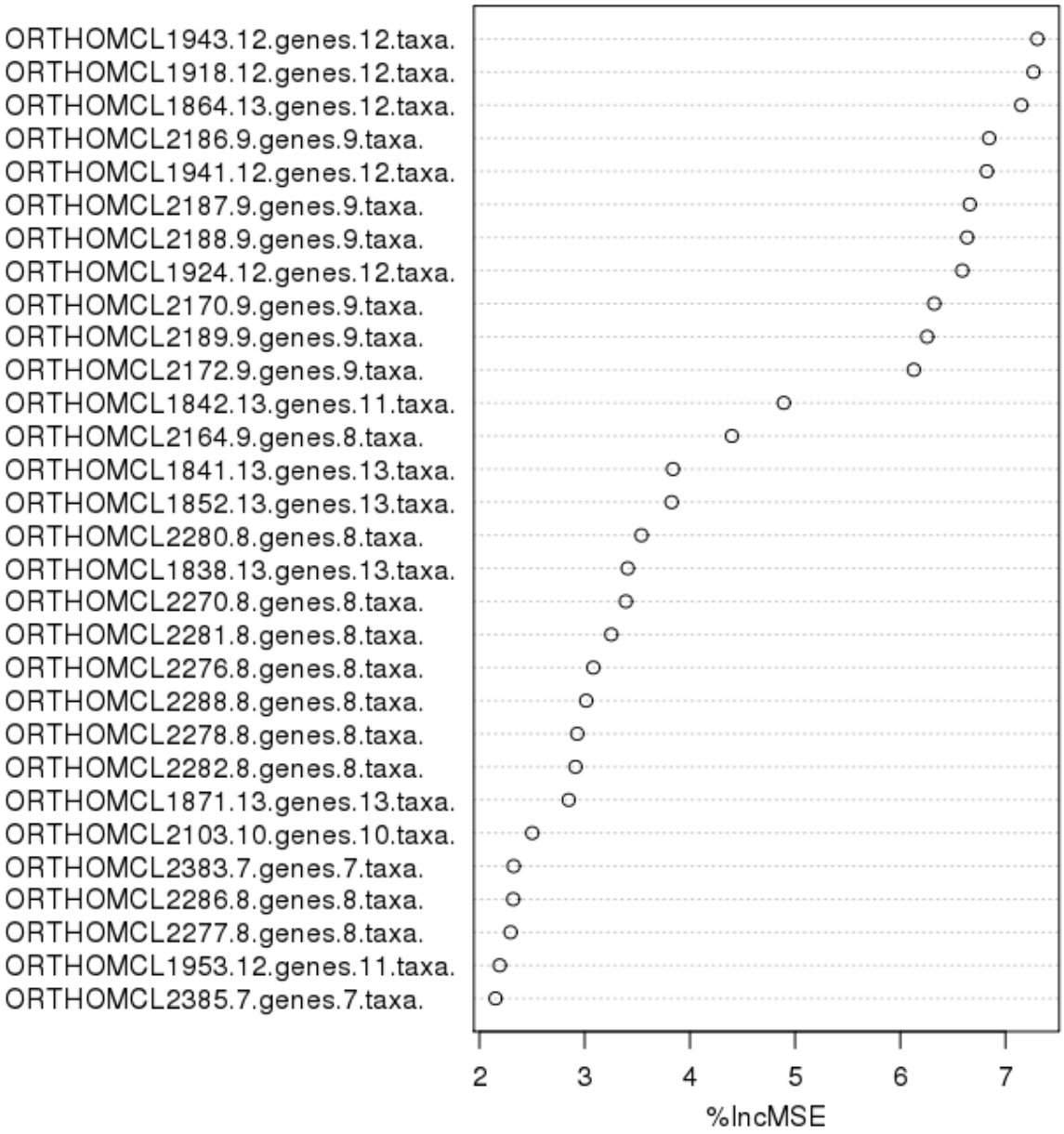


Figure 4.2 Variable importance plot output from investigation of the genetic basis for group A. This variable importance analysis is based upon a randomForest analysis with 1000 trees for the base strain set. The 30 most important variables from the randomForest analysis are found on the y-axis, their importance is found on the x-axis. Increasing importance is described as increasing mean squared error resulting from the permutation of the variable in the randomForest analysis.

4.3.1.1 Clusters driving formation of group A

Four of the clusters whose presence is important in defining group A are proteins in sequential genomic locations (epi_02134, epi_02135, epi_02136, epi_02137). Two of these clusters represent a TCS sensor/regulator with similarity to GraSR (Table 4.3a). PROKKA annotated these proteins as sensor histidine kinase (GraS) and glycopeptide resistance associated protein R (GraR). The genomes of the organisms in group A were revealed to contain 2 loci annotated by PROKKA as GraSR, except in the case of *S. pettenkoferi* which contained only one locus annotated as GraRS. Bioinformatic analysis revealed that in the included *S. aureus* genomes the locus identified by the randomForest analysis shared ~100% sequence identity to SA2417/SA2418 of *S. aureus* HG001, which was shown by Hiron *et al* to encode a TCS associated with resistance to nisin and bacitracin (Hiron *et al.* 2011). Further investigation of this locus revealed that it was present only in the genomes of group A species of staphylococci, and absent from the rest of the genomes included in the OrthoMCL analysis. This locus is explored further in section 4.3.3.1

The remaining two adjacent clusters (epi_02134, epi_02135) positively associated with formation of group A constitute an ABC transporter composed of two subunits with 98% and 99% sequence identity, respectively to the SA2415 and SA2416 ABC transporter proteins, designated as BraD/BraE (Hiron *et al.* 2011) was also observed. These proteins were annotated by PROKKA as FtsX-like permease family protein and macrolide export ATP binding/permease protein MacB. This locus is absent in staphylococci outside of group A. The ABC transporter genes were reported to be co-transcribed as an operon separately from *braRS*. Furthermore, the ABC transporter BraDE is induced in the presence of bacitracin and nisin and this inducible expression is controlled by BraRS (Hiron *et al.* 2011).

The absence of two clusters which represent multidrug ABC transporters (sap_00398, sap_00399) is also important in defining group A. Multidrug ABC transporters mobilise a range of cytotoxic molecules across the cell membrane. One well-characterised example is the ABC transporter SAV1866 from *S. aureus*

which is used as a model in the study of multidrug ABC transporters (Dawson & Locher 2006) (Velamakanni *et al.* 2008). ABC transporters can also act as sensors in some families of TCS (Rietkötter *et al.* 2008). Together with the absence of these clusters of ABC transporter proteins the presence of a separate cluster representing a different ABC transporter in group A contributes to the variable importance analysis (epi_00394). This cluster was found by BLAST similarity to be a bacitracin ABC transporter . The importance of a differential repertoire of ABC transporters associated with survival from antimicrobials may indicate the importance to speciation of adaptation and responses to the arsenal of antibiotics produced by the members of the communities in their different niches in order to compete for resources.

The contribution of the absence of succinate semialdehyde dehydrogenase NADP+ (sap_00201) from group A staphylococci may indicate differences in the metabolism of glutamate and responses to certain environmental stressors in organisms inside of group A and those outside of group A. Glutamate is involved in multiple metabolic processes across the tree of life. In bacteria the enzyme glutamate dehydrogenase catabolises glutamate, which plays a role in acid tolerance. Succinate semialdehyde dehydrogenase is involved in catabolism of γ-aminobutyrate, a product of glutamate dehydrogenase activity (Feehily & Karatzas 2013). This pathway is proposed to have a role in oxidative stress or glutamate metabolism. Redundancy in succinate semialdehyde dehydrogenases across species of bacteria has been shown; *E. coli* for example encodes two such enzymes, whereas *Pseudomonas syringae* encodes three isozymes (Donnelly & Cooper 1981) (Buell *et al.* 2003). Different succinate semialdehyde dehydrogenases can have different requirements for the cofactor NAD(P) (Zhu *et al.* 2010).

The absence of a cluster representing UDP-glucose 4-epimerase (sap_00643) from group A organisms is supported by known species differences in carbohydrate metabolism. UDP-glucose 4-epimerase is an enzyme of the Leloir pathway for metabolising lactose and galactose. This

Table 4.3a Proteins driving formation of group A. PROKKA annotation was found by mapping clusters from the variable importance analysis to the S. epidermidis genome in the case of "present" clusters and the S. saprophyticus genome in the case of the "absent clusters". The PROKKA locus tag is indicated in brackets. BLAST similarity was found by searching the protein against the NCBI BLAST database.

PROKKA annotation	**BLAST similarity**
Presence	
FtsX like permease family protein (epi_00394)	Bacitracin ABC transporter
Hypothetical protein (epi_00430)	Hypothetical protein
Hypothetical protein (epi_00538)	Hypothetical protein
Putative distant relative of cell wall associated hydrolases (epi_00542)	Hypothetical protein
Hypothetical protein (epi_01643)	Hypothetical protein
Staphylococcal accessory regulator U (epi_01831)	Hypothetical protein
Regulatory protein soxS (epi_01832)	AraC family transcriptional regulator
Hypothetical protein (epi_01980)	Abortive infective protein
Hypothetical protein (epi_02055)	Hypothetical protein
Hypothetical protein (epi_02098))	Cell wall surface anchor protein
Hypothetical protein (epi_02108)	Hypothetical protein
FtsX like permease family protein (epi_02134)	ABC transporter permease
Macrolide export ATP binding/permease protein MacB (epi_02135)	bacteriocin ABC transporter ATP-binding protein
Sensor histidine kinase GraS (epi_02136)	Sensor histidine kinase
Glycopeptide resistance associated protein R (epi_02137)	PhoB family transcriptional regulator
Absence	
Succinate semialdehyde dehydrogenase NADP + (sap_00201)	Succinate-semialdehyde dehydrogenase
Putative membrane protein putative toxin regulator (sap_00203)	PTS sugar transporter subunit IIC
Putative multidrug resistance ABC transporter ATP binding/permease protein YheI (sap_00398)	multidrug ABC transporter ATP-binding protein
Putative multidrug resistance ABC transporter ATP binding/permease protein YheH (sap_00399)	Multidrug ABC transporter ATP-binding protein
UDP-glucose 4-epimerase (sap_00643)	UDP-glucose 4-epimerase
L-lactate utilization operon repressor (sap_00760)	transcriptional regulator
Glutamate aspartate carrier protein (sap_01003)	sodium:dicarboxylate symporter
L-serine dehydratase beta chain (sap_01607)	serine dehydratase subunit beta
Hypothetical protein (sap_02162)	serine dehydratase subunit beta

pathway is catalysed by β-galactosides. It is known that *S. aureus* cannot utilise this pathway, instead it metabolises lactose and galactose via a different mechanism catalysed by phospho-β-galactosidase (Bissett & Anderson n.d.).

4.3.1.2 Clusters driving formation of group A_1.

The absence of the OrthoMCL clusters which encode genes of the *fad* operon is perhaps the most striking difference between group A_1 and the rest of the species analysed (Table 4.3b). The *fad* genes, *fadA, fadD,* in *S. aureus* encode the proteins of the fatty acid degradation pathway. These genes were down-regulated in response to linoleic acid (Kenny *et al.* 2009). The indicated importance of the of the absence of the fatty acid metabolism pathway (sap_02495, sap_02494, sap_02493, sap_02491) from group A_1 species indicates that fatty acid metabolism occurs differently in the two groups. Notably this operon is also absent from *S. simulans,* which is closely related to *S. carnosus* that does encode the fatty acid metabolism pathway. *S. epidermidis* was found to carry a gene annotated as *fadD* by PROKKA (epi_00312). This gene encodes a long chain fatty acid CoA ligase, as does *S. saprophyticus fadD (*sap_02494*).* The protein sequences of the two genes however show only 20% sequence identity with one another. The *S. epidermidis fadD* is found next to a gene encoding an Acetyl-CoA acetyltransferase, in *S. saprophyticus* the adjacent gene is an Acyl-CoA dehydrogenase, short-chain specific (sap_02493).

Differences in the histidine biosynthesis operon were also important in defining the subgroup A_1. The histidine operon contains eight genes and is similar in structure and function to that of *S. typhimurium* (KLOSS & PATTEE 1965). Absence of the genes sap_00436, sap_00437, sap_00438, sap_00439, sap_00440, sap_00441, sap_00442, sap_00443 and sap_00444 was shown to be important. Upon investigation, *S. epidermidis* was revealed to encode 7 of the 8 genes of the histidine operon; only *hisC* was missing. A homologue of the gene sap_00444 encoding a regulator was also present. The *S. epidermidis* histidine biosynthesis operon was also included in the OrthoMCL clusters, with 8 of the 9 genes of the non-group A_1 species.

Table 4.3b Proteins driving formation of group A_1. PROKKA annotation was found by mapping clusters from the variable importance analysis to the S. epidermidis genome in the case of "present" clusters and the S. saprophyticus genome in the case of the "absent clusters". The PROKKA locus tag is indicated in brackets. BLAST similarity was found by searching the protein against the NCBI BLAST database.	
PROKKA annotation	BLAST similarity
Presence	
Hypothetical protein (epi_02098)	Cell wall surface anchor protein
FtsX like permease family protein (epi_02134)	ABC transporter permease
Hypothetical protein (epi_01465)	Hypothetical protein
Hypothetical protein (epi_00207)	Hypothetical protein
Hypothetical protein (epi_00829)	Hypothetical protein
Hypothetical protein (epi_02277)	Hypothetical protein
Hypothetical protein (epi_01534)	Hypothetical protein
Staphylococcal secretory antigen ssaA2 precursor (heamolyticus_00353)	Hypothetical protein
Absence	
Na + /dicarboxylate symporter (sap_00900)	Na + /dicarboxylate symporter
Putative ring cleaving dioxygenase mhqO (sap_00404)	Glyoxalase
Oxidoreductase YdhF (sap_02092)	Aldo/keto reductase
Phosphoribosyl ATP pyrophosphatase (sap_00436)	phosphoribosyl-AMP cyclohydrolase
Imidazole glycerol phosphate synthase subunit HisF (sap_00437)	imidazole glycerol phosphate synthase
1-5 phosphoribosyl-5 [5 phosphoribosylamino methylideneamino] imidazole 4 carboxamide isomerase (sap_00438)	1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase
Imidazole glycerol phosphate synthase subunit HisH 2 (sap_00439)	imidazole glycerol phosphate synthase
Imidazoleglycerol phosphate dehydratase (sap_00440)	imidazoleglycerol-phosphate dehydratase
Histidinol dehydrogenase (sap_00442)	histidinol dehydrogenase
ATP phosphoribosyltransferase (sap_00443)	ATP phosphoribosyltransferase catalytic subunit
ATP phosphoribosyltransferase regulatory subunit (sap_00444)	ATP phosphoribosyltransferase
Histidinol phosphate aminotransferase (sap_00441)	histidinol-phosphate aminotransferase
Butyrate-acetoacetate CoA transferase subunit A (sap_02495)	acetyl-CoA transferase
Long chain fatty acid-CoA ligase (sap_02494)	long-chain fatty acid--CoA ligase
Acyl CoA dehydrogenase-short chain specific (sap_02493)	glutaryl-CoA dehydrogenase
3 ketoacyl CoA thiolase (sap_02491)	acetyl-CoA acetyltransferase
Formate channel 1 (sap_00605)	formate/nitrite transporter

Investigation of the *S. haemolyticus* genome revealed that the histidine biosynthesis operon was not present in the genome. This indicates that this inability to produce histidine may be important in defining group A_1. Despite this, *S. epidermidis* may not produce histidine due to the lack of the *hisC* gene from the histidine operon.

The presence of the staphylococcal secretory antigen was important to defining group A_1. The *S. epidermidis* SsaA protein was not however present in the OrthoMCL cluster with the rest of the group A_1 species. Investigation of the genome of *S. saprophyticus* demonstrated the presence of an *ssaA* gene, however this was evidently sufficiently diverged from that of the group A_1 species, excluding *S. epidermidis*, that they did not cluster together in the OrthMCL analysis. SsaA was isolated from *S. epidermidis* strain NCTC 11047 and shares homology with SceB from *S. carnosus*. The protein is antigenic and patients with *S. epidermidis* endocarditis showed elevated levels of antibodies to the protein (Lang *et al.* 2000).

The presence of the gene product designated as BraD in the analysis above is important in defining group A_1 though *S. simiae* and *S. aureus* are both also present in this OrthoMCL cluster. The inclusion of these species in the group-defining clusters is also the case for a hypothetical protein with BLAST similarity to a cell wall surface anchor protein (epi_02098). The presence of group A_2 species in important "presence" clusters such as these, and conversely the presence of group A_2 species in important "absence" clusters indicated that groups A_1 and A_2 can not easily be separated. These clusters which hinder separation of the two clusters are highlighted in bold (Table 4.3b). As a consequence of this basis for the formation of group A, it will be discussed as a whole henceforth.

4.3.1.3 Clusters driving formation of group B

The presence of an OrthoMCL cluster representing a transcriptional repressor contributes to defining group B (Table 4.4, xylosus_02454). This transcriptional regulator was annotated by PROKKA as a lactose operon repressor in *S. xylosus* and catabolite control protein A in *S. saprophyticus*. The *ccpA* gene was first discovered in *B. subtillis* and its product is known to regulate the utilisation of alternative sugars in the absence of glucose (Faires *et al.* 1999). The lactose operon repressor inhibits the expression of genes of the *lac* operon that encodes proteins involved in lactose metabolism. Loss of DNA binding of the Lac repressor in the presence of lactose allows transcription of the lac operon.

An orthogolous cluster corresponding to farnesyl diphosphate:farnesyl transerase (xylosus_00489), otherwise known as squalene synthase (SQS), is involved in defining group B. SQS is an alternative pathway for the biosynthesis of carotenoid pigments in certain bacteria, such as species of *Methylomonas*, a methanotrophic Gram-negative bacterium (Tao *et al.* 2005). In *S. aureus* the dehydrosqualene synthase CrtM converts farnesyl diphosphate (FPP) to dehydrosqualene (Furubayashi *et al.* 2014) replacing this SQS activity in carotenoid production. In Methylomonas, no *crtM* locus is detected, instead *sqs* is present and encodes an enzyme that exhibits low dehydrosqualene synthase activity (Tao *et al.* 2005). Furubayashi *et al* determined that in a synthetic pathway *E. coli* SQS and the *S. aureus* carotenoid desaturase CrtN function together to synthesise the carotenoid pigment staphyloxanthin (Furubayashi *et al.* 2014). The putative SQS from *S. saprophyticus* has 13.24% identity to *S. aureus* CrtM and 28% identity with squalene synthase from *Methylomonas*. Despite the presence of the alternative squalene synthase, the group B staphylococci *S. xylosus* and *S. arlettae* do encode CrtM in their genomes, along with the *crtPQN* genes. In contrast to the *S. hominis* genomes discussed however, an acetylesterase (crtO) is present in the *S. xylosus* and *S. arlettae* genomes. This glysocyl-4,4'-diaponeurosporenoate acetyltransferase shows 42% and 39% identity to

Table 4.4 Proteins important in defining group B. PROKKA annotation was found by mapping clusters from the variable importance analysis to the S. xylosus genome in the case of "present" clusters and the S. epidermidis genome in the case of the "absent clusters". The PROKKA locus tag is indicated in brackets. BLAST similarity was found by searching the protein against the NCBI BLAST database.

PROKKA annotation	**BLAST similarity**
Presence	
Lactose operon repressor (xylosus_02454)	Transcriptional regulator
Hypothetical protein (xylosus_02429)	Hypothetical protein
Hypothetical protein (xylosus_02133)	Hypothetical protein
Hypothetical protein (xylosus_02128)	Hypothetical protein
Hypothetical protein (xylosus_02121)	Hypothetical protein
Hypothetical protein (xylosus_01818)	Hypothetical protein
3 demethyl-ubiquinone 3 methyltransferase (xylosus_01429)	SAM-dependent methyltransferase
2 dehydropantoate 2 reductase (xylosus_01388)	Transcriptinal regulator
Glucokinase (xylosus_01379)	Transcriptional regulator
Na +/dicarboxylate symporter (xylosus_01326)	Anion transporter
Hypothetical protein (xylosus_01105)	Transporter protein
HTH type transcriptional repressor glcR (xylosus_01104)	Transcriptional regulator
Hypothetical protein (xylosus_01043)	Membrane protein
Hypothetical protein (xylosus_00825)	Hypothetical protein
Osmoprotectant binding protein (xylosus_00598)	Glycine/betaine ABC transporter substrate-binding protein
Hypothetical protein (xylosus_00536)	Hypothetical protein
Farnesyl diphosphate:farnesyl transferase (xylosus_00489)	Phytoene synthase
Elastin binding protein ebpS (xylosus_00325)	Elastin-binding protein Ebps
Hypothetical protein (xylosus_00257)	Hypothetical protein
Hypothetical protein (xylosus_00177)	Hypothetical protein
Putative BCR/YitT family (xylosus_00036)	Hypothetical protein
Absence	
Hypothetical protein (epi_02082)	Hypothetical

USA300 crtO in *S. xylosus* and *S. arlettae* respectively. It is unknown if the carotenoid operon is functional in these species.

2-dehydropantoate 2-reductase (xylosus_01388), an oxidoreductase involved in the synthesis of pantothenate, contributes to the variable importance analysis and associates with group B species. Pantothenate is a precursor for the biosynthesis of coenzyme A. There is evidence of intraspecific variation in pantothenate biosynthesis, including at loci encoding 2-dehydropantoate 2-reductase (Aakra *et al.* 2007). According to KEGG pathway annotation 2-dehydropantoate 2-reductase is involved in the biosynthesis of secondary metabolites.

Other clusters that contribute to the definition of cluster B include an osmoprotectant (glycine/betaine ABC transporter substrate, xylosus_00598) binding protein. Osmoprotectant transporters can uptake substrates such as glycine-betaine from the environment to maintain cytoplasmic osmolarity. The presence of a protein which binds these substrates indicates a protein with a role in managing the osmotic stress response (Kuroda *et al.* 2004). In addition, a protein with sequence identity to the elastin binding protein EbpS (xylosus_00325), an integral membrane protein present in several staphylococcal species, was present in the variable importance analysis. EbpS is discussed in detail in chapter 3 (section 3.1.4.3). The presence of EbpS as a separate OrthoMCL cluster in the group B organisms suggests that this protein is sufficiently diverged in these organisms that it might no longer represent an orthologue of EbpS in *S. aureus*. Finally, the presence of another hypothetical protein with BLAST similarity to a membrane protein was also found to be importance in the definition of group B.

4.3.1.4 Clusters driving formation of group C

A putative glycosyltransferase with similarity to TagX (sap_02141)was absent from group C comprising SIG species: *S. pseudintermedius*, *S. delphini* and *S.intermedius* (Table 4.5). Certain glycosyl-transferases incorporate N-acetylglucosamine into surface molecules, such as wall teichoic acids. TagX

however was shown not to be involved in the glycosylation of wall teichoic acids (Winstel *et al.* 2014). Glycosyltransferases are also involved in glycosylation and therefore reducing the immunogenicity of other surface proteins (Hazenbos *et al.* 2013); this could be the function associated with TagX.

The presence of a second cluster which represents a D-Ala-teichoic acid biosynthesis protein (inter_00088) with 45% identity to *Streptococcus sanguis* DltX was shown to be important in defining group C. This may imply that the SIG staphylococci have an alternative pathway for modifying their cell wall teichoic acids with D-alanine. It is known that there is redundancy in the *tag* genes of staphylococci, even on an intraspecific basis (Winstel *et al.* 2014), therefore it may be possible that there are also differences in cell wall modifications via other pathways.

The presence of a macrolide export ABC transporter protein (inter_02006) is important in defining group C. A BLAST search revealed similarity to a bacteriocin ABC transporter. Bacteriocins are narrow spectrum antibiotics produced by bacteria to inhibit the growth of competitors (Riley & Wertz 2002). The presence of a bacteriocin transporter protein in the importance analysis suggests that SIG species have evolved resistance to the toxic antimicrobials secreted by the community in their niche as canine and equid colonisers. The fact that it was shown to be important in defining group C shows that it is distinct enough from the bacteriocin transporter involved in defining group A (epi_00394) that it is in a separate OrthoMCL cluster.

Two clusters found to represent glutamyl specific endopeptidases (inter_00008, inter_00009) with identity to *S. aureus* V8 serine protease were also found to be important in defining group C. V8 protease has similarity to *S. aureus* exfoliative toxins (Ladhani 2003) and has been shown to cause epidermal barrier dysfunction in nude mice (Hirasawa *et al.* 2009).

Table 4.5 Proteins important in defining group C. PROKKA annotation was found by mapping clusters from the variable importance analysis to the S. intermedius genome in the case of "present" clusters and the S. saprophyticus genome in the case of the "absent clusters". The PROKKA locus tag is indicated in brackets. BLAST similarity was found by searching the protein against the NCBI BLAST database.

PROKKA annotation	**BLAST similarity**
Presence	
V8 like Glu specific endopeptidase (inter_00008)	Glutamyl-endopeptidase
V8 like Glu specific endopeptidase (inter_00009)	Glutamyl-endopeptidase
D Ala teichoic acid biosynthesis protein (inter_00088)	D-Ala-teichoic acid biosynthesis protein
Yip1 domain protein (inter_00369)	Hypothetical protein
Hypothetical protein (inter_00749)	Possible membrane protein
DoxX (inter_01836)	DoxX family protein
Macrolide export ATP binding/permease protein MacB (inter_02006)	Amino acid ABC transporter permease
Hypothetical protein (inter_02256)	Hypothetical protein
Hypothetical protein (inter_02629)	Hypothetical protein
Absence	
Aspartokinase 2 (sap_01421)	Aspartokinase ii
Putative glycosyltransferase tagX (sap_02141)	Teichoic acid biosynthesis protein

4.3.2 Diversity across the staphylococci at loci associated with CAMP survival

4.3.2.1 Intramembrane histidine kinase sensor : response regulator systems

As described above, the presence of a two component sensor regulator was shown to be important in the definition of group A, and that this TCS in the strains of *S. aureus* analysed had ~ 100% sequence identity to BraRS. The two component system identified in the random forest analysis as being important in defining group A will therefore be referred to as BraRS from this point.

Due to the presence of two homologous TCSs in the group A species, the sequence identity of GraRS, the conserved TCS, was investigated across groups A-C in this analysis. A high degree of identity, ≥77% (Table 4.6a), between the protein sequences of the GraR regulator was seen within the three groups with the highest identity of 97.9% among the organisms of group C. The identity between the groups fell to 66.7% indicating some degree of divergence in the GraR protein.

A neighbour joining tree was generated from an alignment of the GraR protein sequences (Figure 4.3). The same clustering observed in the heatmap generated from the OrthoMCL analysis (Figure 4.1) was observed, with the labels A, B and C corresponding to the respective heat map groups. Notably *S. pettenkoferi* was observed to be an outlier in this analysis.

The sequence identity of the GraS sensor histidine kinase locus within the three groups was also investigated. This protein was shown to be less conserved than the regulator within each group A and B as the identity within these groups was 60.4% and 66%, respectively (Table 4.6a). GraS was more highly conserved in members of group C, with an identity of 88.2%. In comparison, the sequence identity between the groups A-C was 48.2% and lower than that observed for the GraR regulator protein.

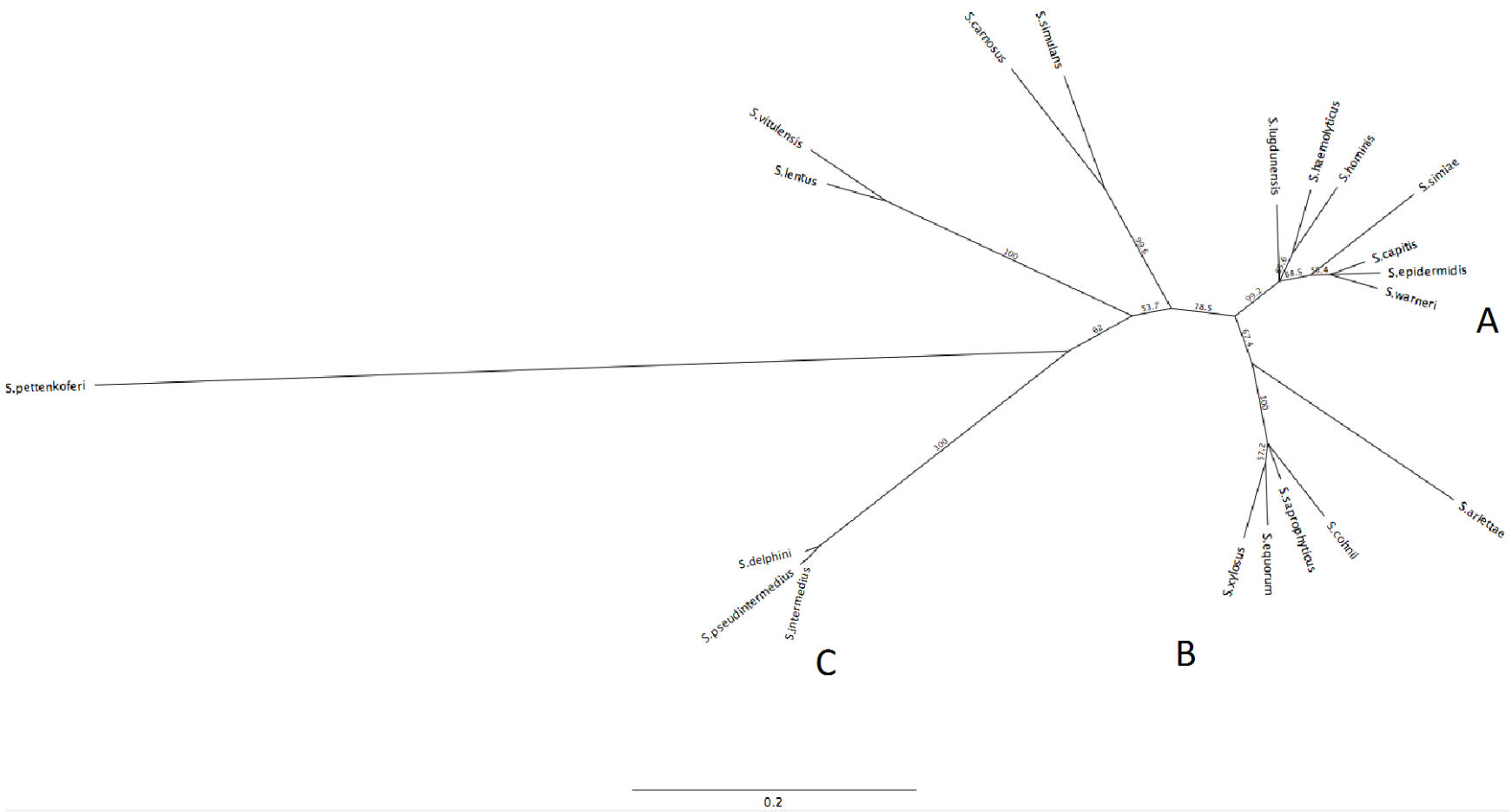


Figure 4.3 Neighbour-joining tree of the GraR protein. The basis of the tree is the alignment of the GraR protein sequence from all staphylococci included in the OrthoMCL analysis. A indicates group A strains from the OrthoMCL analysis heatmap, similarly B and C indicate groups B and C from the same analysis. Bootstrapping using 1000 replicates was performed.

A neighbour-joining tree was generated from an alignment of the protein sequences of GraS (Figure 4.4). Again, consistent with GraR, a clustering pattern in the heatmap generated from the OrthoMCL analysis was observed (Figure 4.1) with the exception of an increased tree branch length for *S. lentus* and *S. vitulinus* and indicating increased divergence of the GraS sensor histidine kinase across the different species. *S. pettenkoferi* was again observed as an outlier by not clustering with group A in this analysis.

The sequence identity of the BraR regulator protein sequence among the group A organisms was investigated (Table 4.6b). Due to the fact that the GraRS proteins from *S. pettenkoferi* did not cluster with the rest of the group A organisms in the neighbour-joining trees, and that the GraRS annotated locus in this species was present in the OrthoMCL clusters representing BraRS in the rest of group A, this locus was included in the BraRS analysis. It was shown that the identity in the BraR protein sequence within group A was 77.1%. Identity of this protein with GraR in group B and C species was 39.6% and identity with GraR in group A was 44.34%.

Table 4.6a Percentage identity in the GraRS TCS. Average % identity in the GraR and GraS within groups A, B and C as well as between all the groups was calculated from multiple sequence alignments of all protein sequences.

Two component system	**identity within group A**	**Identity within group B**	**Identity within group C**	**Identity between all groups**
GraR	77.3/87.8*	84	97.9	66.7
GraS	60.4/69.4*	66	88.2	48.2

Table 4.6b Percentage Identity in the BraRS TCS. Average % Identity in the BraR and BraS within group A as well as % identity to GraRS was calculated from multiple sequence alignments of all protein sequences.

	Identity within group A	**Identity to group B & C graRS**	**Identity to group A Gra**
BraR	77.1	39.6	44.34
BraS	62.9	26.4	29.8

* % Identity within group A was also calculated excluding the outlier S. pettenkoferi

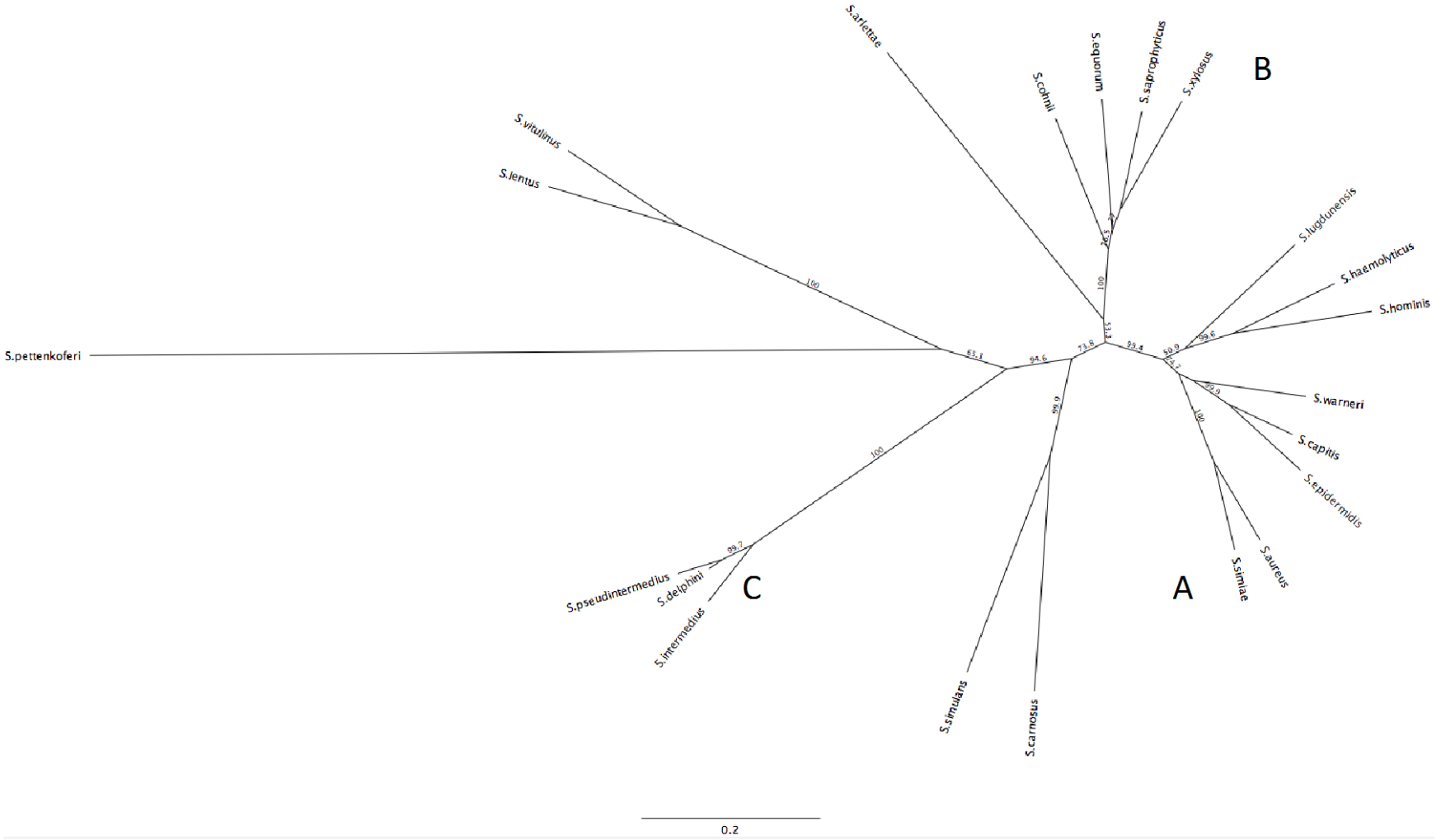


Figure 4.4 Neighbour-joining tree of the GraS protein. The basis of the tree is the alignment of the GraS protein sequence from all staphylococci included in the OrthoMCL analysis. A indicates group A strains from the OrthoMCL analysis heatmap, similarly B and C indicate groups B and C from the same analysis. Bootstrapping using 1000 replicates was performed.

A neighbour-joining tree based on the alignment of the BraR protein was generated (Figure 4.5). The clusters present in the heatmap generated from the OrthoMCL analysis were observed and clear divergence between the orthologues BraR and GraR was identified.

Sequence identity of the BraS protein sequence within group A was found to be 62.9% (Table 4.6b). Identity of BraS to GraS in groups B and C and group A was 26.4% and 29.8% respectively indicating increased divergence from the GraS homologue compared to the regulator protein. A neighbour-joining tree was generated based on the alignment of the BraS protein sequence from group A and the GraS protein sequences from groups B and C (Figure 4.6). Groups B and C clustered together however the clustering of group A broke down with increased branch lengths in this cluster representing increased divergence of the BraS protein sequence compared to the GraS protein sequence. *S. lentus* and *S. vitulinus* were also observed to cluster with the group A genomes indicating identity between their GraS protein sequence and the BraS protein sequence. Despite this, the sequences of the *S. lentus* and *S. vitulinus* GraS protein share only an average 26.5% identity with the group A BraS proteins.

It is clear that the process of responding to cationic antimicrobial peptides in the staphylococci is complex, highlighted by the level of divergence observed between species in the two TCSs GraRS and BraRS. This divergence indicates that the TCSs are under strong selection pressures, which is unsurprising as cationic peptides are constantly present in their host environment. The first point of interest to note is the presence of only one of the two TCSs in *S. pettenkoferi*. Due to the fact that the TCS shows higher percentage identity to BraRS and *S. pettenkoferi* was observed as an outlier in the neighbour-joining tree based on the GraRS protein sequences (Figures 4.6, 4.7), it appears that this singular TCS is BraRS. This raises questions about the evolution of BraRS in the group A organisms.

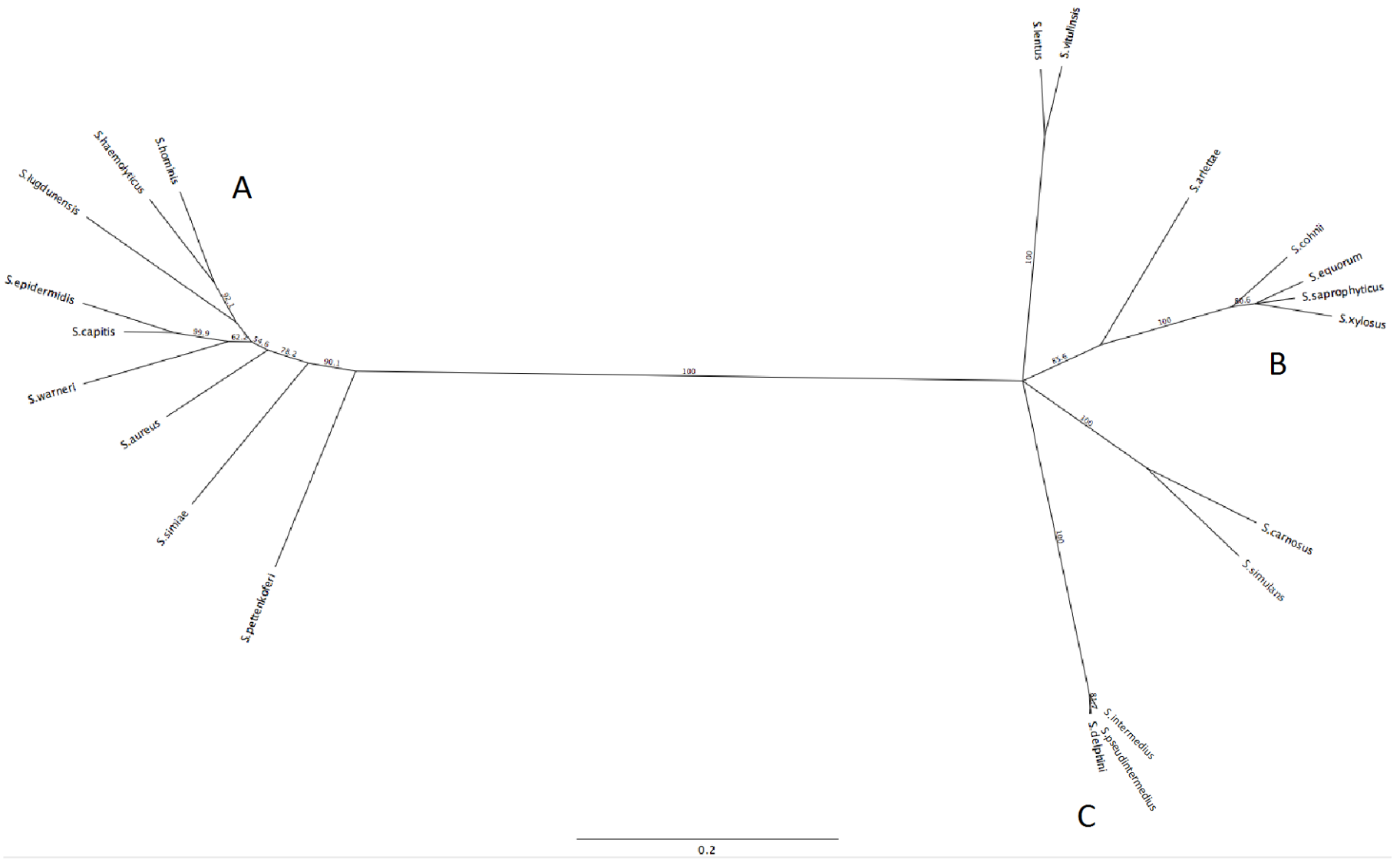


Figure 4.5 Neighbour joining tree of graR and the orthologous BraR proteins. The basis of the tree is the alignment of the BraR protein sequence from group A organisms with the GraR protein sequence from non-group A organisms from the OrthoMCL analysis. The labelled A, B and C refer to the groups A, B and C defined in the same analysis. Bootstrapping using 1000 replicates was performed.

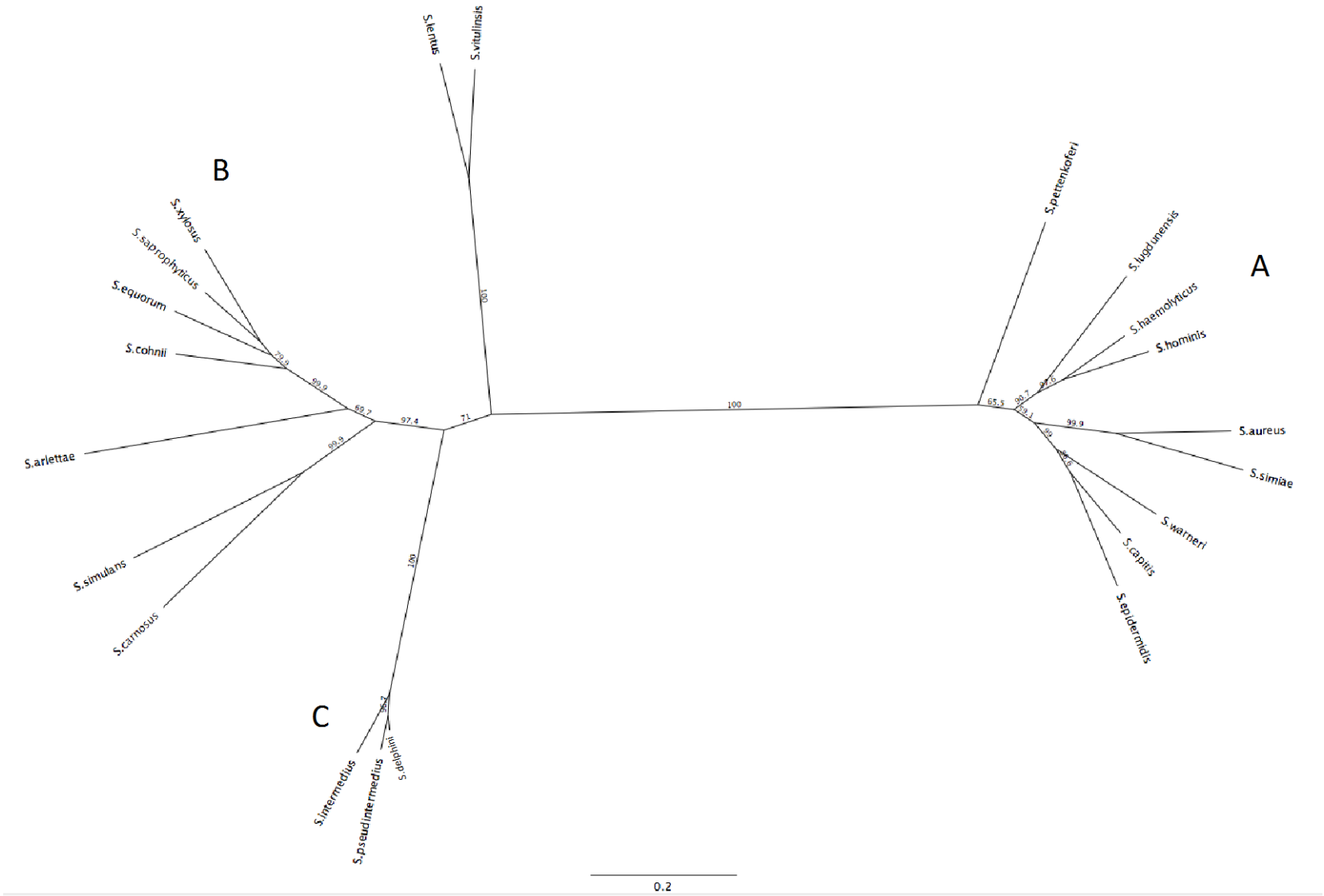


Figure 4.6 Neighbour joining tree of GraS and the orthologous BraS proteins. The basis of the tree is the alignment of the BraS protein sequence from group A organisms with the GraS protein sequence from non-group A organisms from the OrthoMCL analysis.. The labelled A, B and C refer to the groups A, B and C defined in the same analysis. Bootstrapping using 1000 repetitions was performed.

One explanation could be that a gene duplication of GraRS in one of the group A species and then subsequent divergence of the BraRS system from GraRS; this duplicated system may have then been spread through the rest of group A by horizontal gene transfer. *S. pettenkoferi* presents a challenge to this paralogue hypothesis, however, due to the presence of only BraRS. There are two possibilities which might explain the presence of only one of the TCSs; *S. pettenkoferi* could have either lost *graRS* following acquisition of *braRS*, or *S. pettenkoferi* never acquired *braRS,* but rather its *braRS* evolved from an ancestral *graRS*. The other group A organisms may then have acquired *braRS* from *S. pettenkoferi* as an additional TCS to *graRS*.

Regardless of the origins of the two TCSs, the presence of sequence variation both among and between GraRS and BraRS suggests that these systems evolved to confer resistance to the variety of cationic antimicrobial peptides with different structures and mechanisms of action present in the different niches to which the staphylococci are specialised. It is known that mutations in the TCSs alter resistance to a variety of antibiotics, particularly mutations in the gene encoding the sensor histidine kinase protein of both systems. It was shown that a mutation in *graS* leads to intermediate vancomycin resistance in MRSA (Howden *et al.* 2008), for example. Furthermore, Blake et al found a point mutation in *braS* to be responsible for increased resistance to the lantibiotic nisin (Blake *et al.* 2011).

The role of GraRS and BraRS in species of staphylococci inside and outside of group A is explored in further detail in chapter 5.

4.3.2.2 Associated ABC transporters

It was discussed that BceS-like intra-membrane histidine kinases do not act directly to sense the presence of antimicrobial peptides, but rather act as signal transfer molecules which activate the response regulator in the presence of the antimicrobial peptide substrate bound to the associated ABC transporter(Mascher 2014). Both GraRS and BraRS are members of this family of TCSs. Due to the importance of the ABC transporter to the functioning of

these TCSs the protein sequence of the ABC transporters associated with both BraRS and GraRS were retrieved from the genomes and investigated. It was evident that the proposed ABC transporter associated with GraRS was not present in groups B genomes. *S. pettenkoferi* was only found to carry BraRS, however no ABC transporter was observed in the expected location adjacent to BraRS, or in the surrounding genomic region.

The ABC transporter associated with GraRS is VraFG (Meehl *et al.* 2007). A protein sequence alignment of the ABC transporter adjacent to the GraRS locus with the protein sequences of VraFG from *S. haemolyticus* strain JCSC1435 and *S. aureus* strain Z172 was carried out. Intraspecies variation was low with the *S. aureus* VraF from this analysis sharing 93% identity with *S. aureus* Z172, and *S. haemolyticus* VraF from this analysis sharing 99.6% identity with *S. haemolyticus* JCSC1435. An average of 68.3% identity was observed at the putative *vraF* locus among all species (Table 4.7). Identity within the groups was higher with group A organisms sharing 79% identity, group B organisms sharing 85.3% identity and group C sharing 96.8% identity. Noticeably, *vraF* was not observed to be adjacent to GraRS in group B organisms. Instead *vraF*, and indeed *vraG* are located elsewhere on the chromosome.

Table 4.7 % sequence identity in the VraFG ABC transporter. Average % identity in the VraF and VraG within groups A and C as well as between the groups was calculated from multiple sequence alignments of all protein sequences.

Two component system	Identity within group A	Identity within group B	Identity within group C	Identity between all groups
VraF	79	85.3	96.8	68.3
VraG	61.9	65.2	88	47.5

A neighbour-joining tree based on the protein sequence alignment of VraF was generated (Figure 4.7). Groups A, B and C clustered together. The length of *S. hominis* VraF however is 275 amino acids compared to 253 amino acids for the protein sequence of the rest of the group A VraF. Despite this, in the protein sequence alignment *S. hominis* showed 73.7% identity with the rest of the group A VraF protein sequences.

Intraspecies variation in the VraG protein sequence was also low with the *S. aureus* VraG from this analysis sharing 90% identity with *S. aureus* Z172, and *S. haemolyticus* VraG from this analysis sharing 100% identity with *S. haemolyticus* JCSC1435. Identity between all species in the VraG protein sequence was 47.5%, and 88%, 65.2% and 61.9% identity for groups A, B and C respectively. Again, no *vraG* gene was observed adjacent to the *graRS* locus in group B species, the protein was located in the same location as the putative *vraF.* A neighbour-joining tree based on the protein sequence alignment of VraG was also generated (Figure 4.8) and Groups A, B and C clustered together (Figure 4.8).

As discussed in section 4.3.2.1 an ABC transporter was identified in group A which was proposed to be BraD/BraE. The sequence identity at this locus between the genomes in group A was investigated. The organisms of group A were shown to share 68.4% identity in their BraD protein sequences, and a much reduced 38.9% identity in their BraE protein sequences.

In the genomes where VraFG was encoded there is a high degree of protein sequence identity, indicating that the ABC transporter is well-conserved among the species. This conservation is higher for VraF than it is for VraG, for example identity of VraF between all groups is 68.3% and 47.5% for VraG. Notably BraD is also more highly conserved than BraE showing 68.4% identity compared to 38.9% identity.

The divergence of BraRS and GraRS may be as a result of the adaptation of the ABC transporter sensors to different antimicrobial peptides. As the ABC transporters evolved, the histidine kinase may have also evolved in such a way

that it was still able to receive and transfer this signal to the response regulator. The reduced divergence of the response regulators may be explained by the fact that they are the part of the system most isolated from the external environment and the changing stimuli.

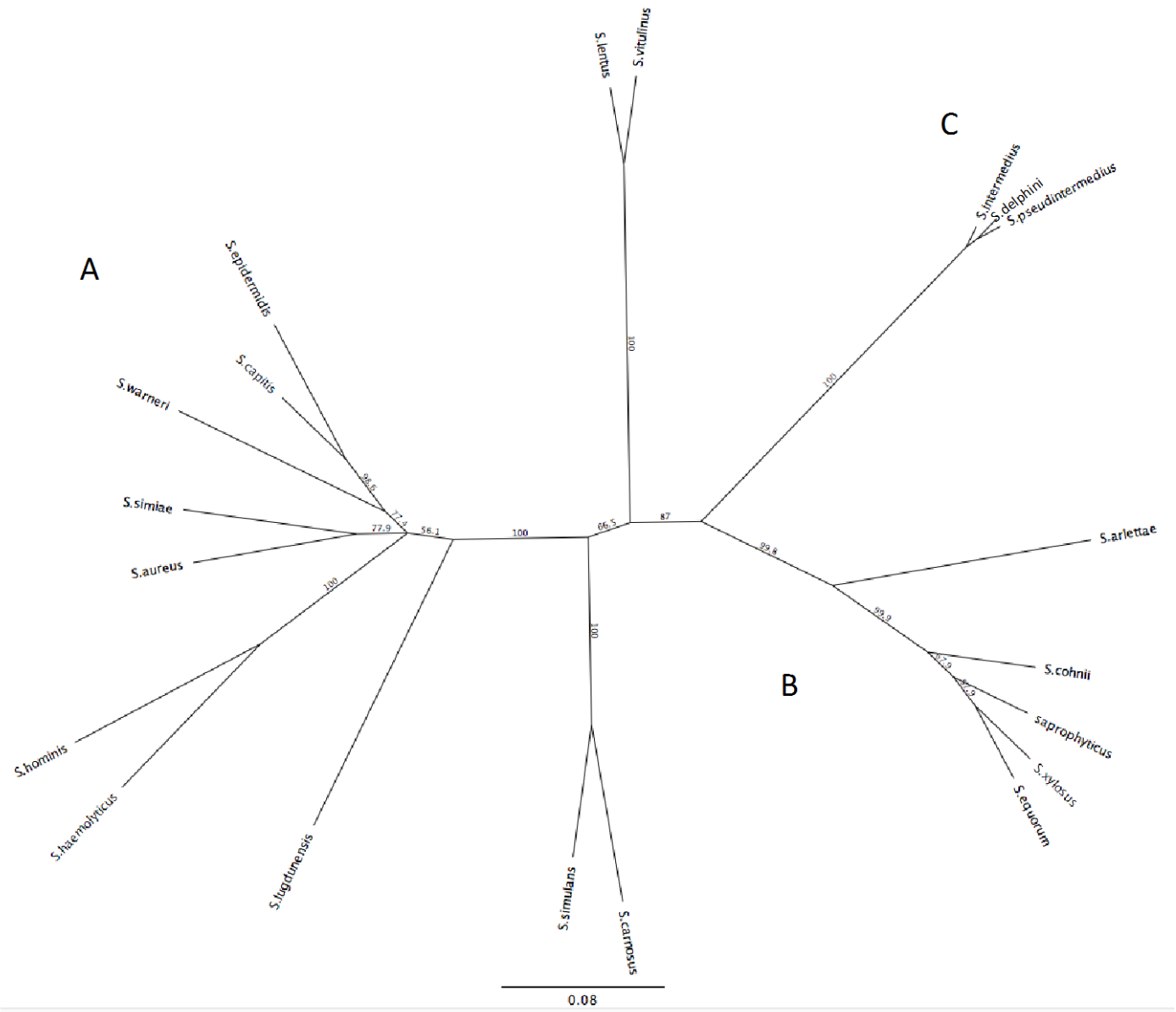


Figure 4.7 Neighbour joining tree of VraF. The basis of the tree is VraF protein sequence alignment from all genomes. The labeled A. B and C indicate groups A and C from the OrthoMCL analysis. Bootstrapping using 1000 replicates was performed**.**

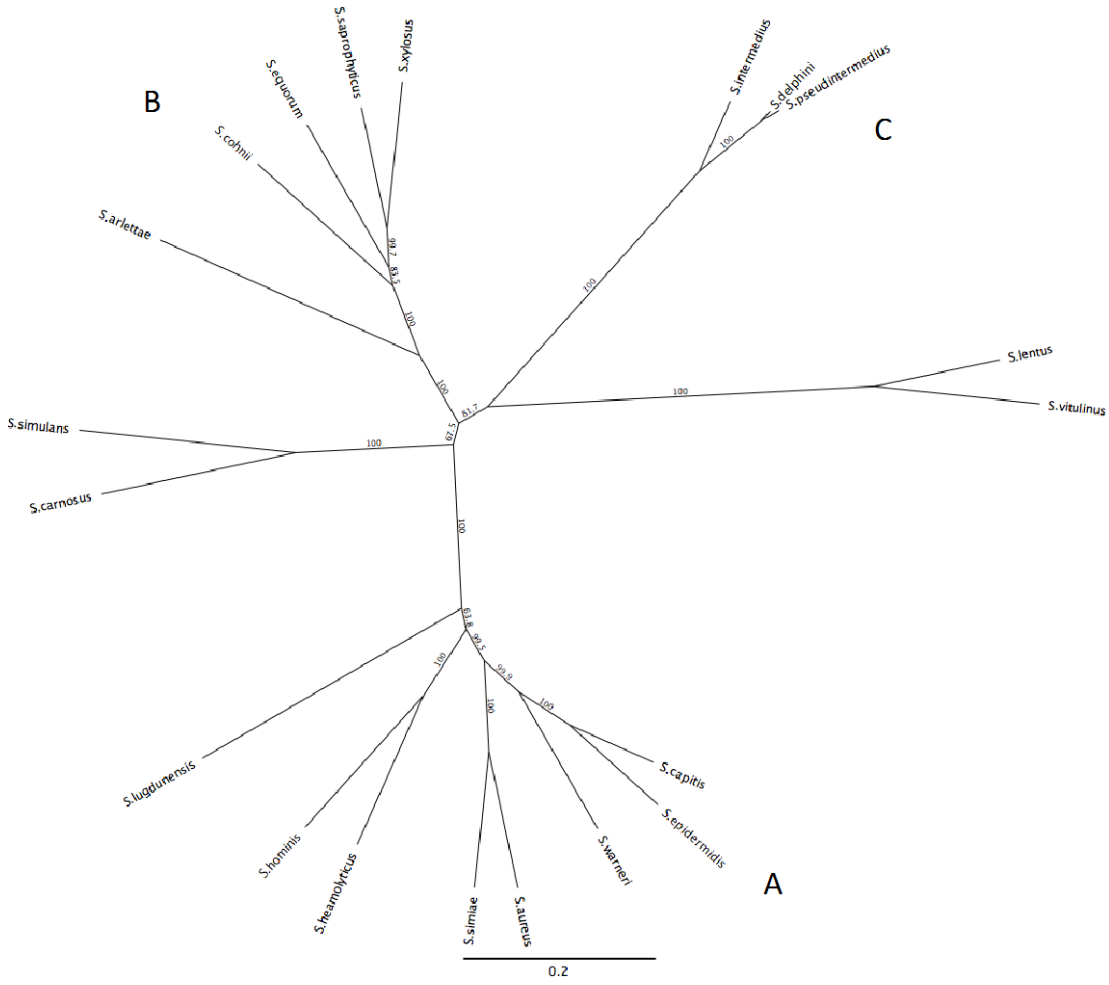


Figure 4.8 Neighbour joining tree of VraG. The basis of the tree is a VraG protein sequence alignment.. The labeled A, B and C indicate groups A, B and C from the OrthoMCL analysis. Bootstrapping using 1000 replicates was performed.

4.4 Conclusion

An investigation of the genetic basis directing formation of the three groups of staphylococcal species based upon shared orthologous clusters (Figure 4.1) highlighted several common themes. Many of the genes driving formation of the clades were associated with metabolism, such as the absence of UDP-glucose 4-epimerase from group A. Differences in resistance to antimicrobial peptides were also suggested by the different repertoirs of ABC transporters among the three groups. Other proteins have roles in mediating the cell's interaction with external stimuli such as the proposed presence of a divergent D-Ala-treichoic acid biosynthesis protein in group C and an alternative version of EbpS in group B. Finally proteins involved in survival of hostile conditions in the host niche, such as osmotic stress, were identified. The presence of genes with these functions in the importance analyses investigating the genetic basis of groups A, B and C indicate that species groups may be defined by metabolism, response to antimicrobial peptides and interaction with the host. Interestingly, differences in fatty acid metabolism were found to differentiate *S. epidermidi, S. capitis, S. warneri, S. haemolyticus, S. hominis, S. lugdunensis* and *S.pettenkoferi* from the rest of the species of Staphylococci in the analysis. Excluding *S. lugdunensis* and *S. pettenkoferi*, this group contains the staphylococci associated with the human skin microbiome. How these species interact with fatty acids on the skin may therefore impact on their successful skin colonisation.

The ability to compete for nutrients, resist the antimicrobial peptides produced by the host and other members of a community and survive the hostile conditions of the host niche are all pre-requisites for success as part of a microbial community. As each of the three groups were comprised of species isolated from similar hosts and niches (4.3.2) it makes sense that the genetic basis behind each group involves functions important in niche specialisation.

One important aspect of niche diversity is the antimicrobial peptides community members both encounter and produce. One would therefore expect diversity in the mechanisms with which the different groups resist these antimicrobial peptides due to strong selective pressures. Diversity was shown

in this analysis although the forces that directed these differences are unclear. Diversity was observed at the GraRS locus both within and between the three groups of species investigated, which is central as the common antimicrobial sensor in almost all *Staphylococcus* species.

The broad host range of *S. aureus* presents questions regarding the suggestion that genes involved in niche specialisation may be involved in creating the basis for the three groups discussed in this analysis. *S. aureus* has been isolated from a variety of hosts including cattle, chickens and rabbits. The presence of BraRS in the animal *S. aureus* strains RF122 (Bovine associated) ED98 (poultry associated) and ERS400835 (rabbit associated) was investigated. The BraRS TCS was found to be present and have 100% protein sequence identity to BraRS from human-associated strains. The associated ABC transporter similarly shared 98% identity with human-associated strains.

In the model laid out by Mascher for the mechanism of action of BceS-like IM-HK systems the antimicrobial peptide is sensed by its binding to an additional protein shown to be the ABC transporter (Rietkötter *et al.* 2008). This signal is then transferred to the TCS which functions to alter gene expression within the cell to confer resistance to the stimulus. It has been shown that increased resistance to antimicrobials and antibiotics is frequently conferred by changes in the histidine kinase (Cui *et al.* 2009) (Blake *et al.* 2011). Furthermore, the largest amount of protein sequence variation in the two TCSs discussed in this chapter is observed in the histidine kinase. This protein is proposed to function as a signal transfer protein between the external signal input from the ABC transporter and the response regulator inside the cell (Mascher 2014).

When the cognate ABC transporters for BraRS and GraRS were investigated the largest amount of protein sequence diversity was observed among VraG and BraE. These proteins function as ABC transporter membrane permeases. It may be expected that these proteins exhibit less sequence identity than the ATP-binding domains VraF and BraD as conservation of the ABC domain in these permeases is important in defining them as ABC transporters (Higgins 2001).

The clustering pattern of the concatenated housekeeping protein sequences is observed for all TCSs and ABC transporters investigated (Figures 4.7-4.12). The neighbour joining tree of the concatenated protein sequences of a panel of housekeeping genes showed that *S. aureus* and *S. simiae* cluster with group A.

Finally, the clustering pattern within group A observed in the the GraS (Figure 4.4) TCS neighbour joining trees does not follow that of the well-conserved housekeeping gene panel (Figure 4.9) in that GraS of *S. lugdunensis* does not cluster closely with that of *S. hominis* and *S. haemolyticus* as is the case for the housekeeping genes as well as BraRS, GraR and VraFG.

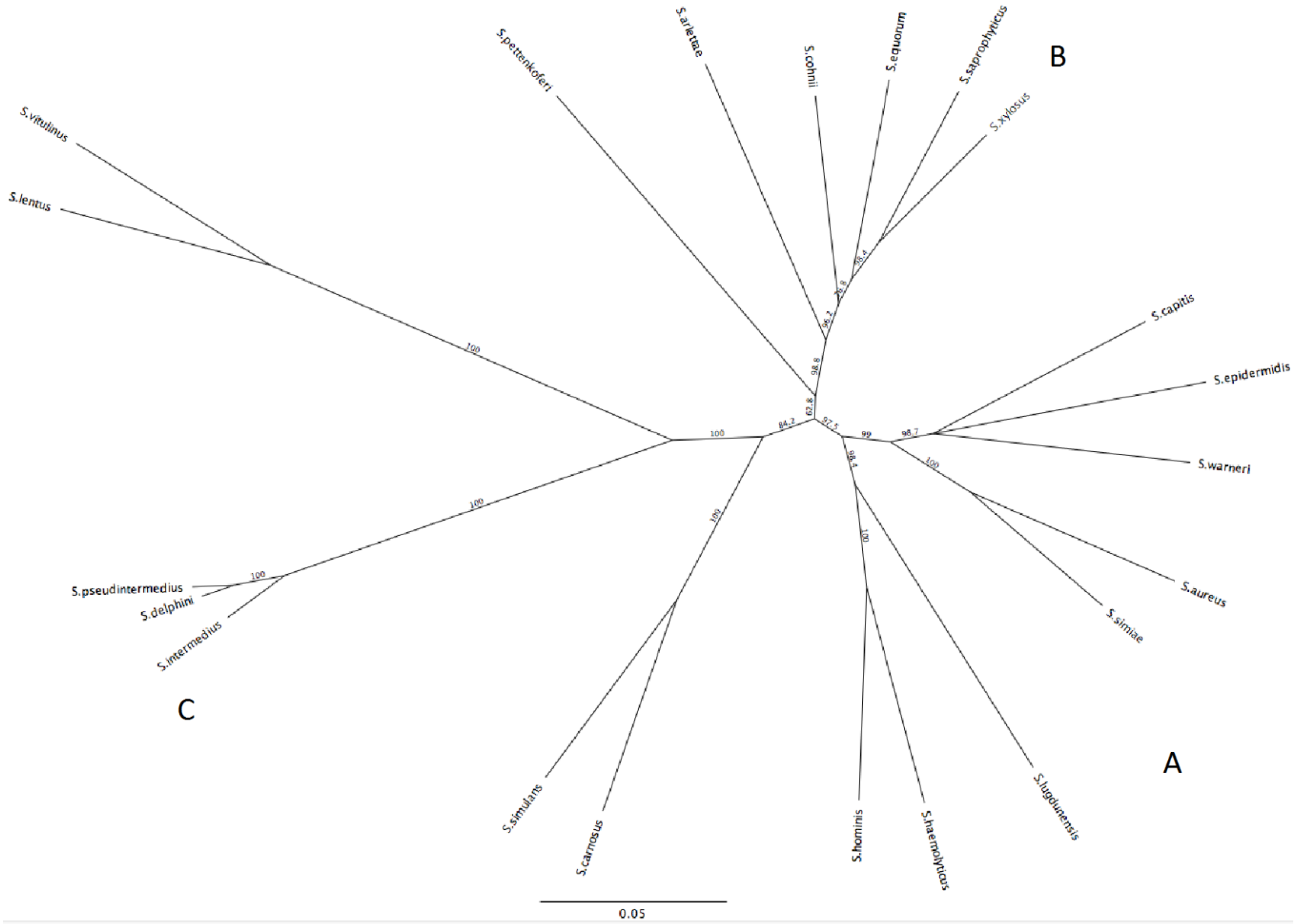


Figure 4.9 Neighbour joining tree of housekeeping genes. The basis of the tree is the concatenated protein sequences of the housekeeping genes *aroE, gmk, pta* and *tpi.* The labelled A, B and C indicate groups A, B and C from the OrthoMCL analysis. Bootstrapping using 1000 repetitions was carried out.

Chapter 5

Experimental evolution of *Staphylococcus hominis*, *Staphylococcus saprophyticus* and *Staphylococcus aureus* with type-A (i) lantibiotic nisin selection.

5.1 Introduction

5.1.1 Biosynthesis, structure and function of nisin

Nisin is a type-A(i) lantibiotic produced by *Lactococcus lactis*. This peptide antibiotic is active against a range of Gram-positive bacteria, including *S. aureus* and species of *Bacillus* and *Clostridium* (Delves-Broughton *et al.* 1996). Type A lantibiotics are so called due to the inclusion of the unusual thioether amino acids lanthionine or methyllanthionine in their structure (Bauer & Dicks 2005). Their biosynthesis also follows an unusual route in that a precursor peptide is ribosomally synthesised as a prepeptide comprised of a C-terminal propeptide and an N-terminal leader peptide. The C-terminal propeptide is post-translationally modified and the leader peptide is cleaved in the final stage of nisin biosynthesis to release the active lantibiotic. The structure of the leader peptide is important to the classification of nisin as a type A(I) lantibiotic. Leader peptides in this group are hydrophilic, contain multiple charged amino acids and are highly conserved (McAuliffe *et al.* 2001).

The nisin biosynthesis gene cluster encodes the protein activities required for the post-translational modifications for maturation of nisin. Of these, *nisB* encodes a hydrophilic protein of around 100 amino acids in length. Membrane localisation of the proteins has been shown by the presence of hydrophobic domains within these relatively large proteins and their presence in the same centrifugation fraction as membrane vesicles in *L. lactis* (Engelke *et al.* 1992). Experiments show NisB is a dehydratase which converts Ser33 in the propeptide to didehydroalanine (Dha)(Sen *et al.* 1999). Dehydration of the hydroxyl serine residue in the propeptide by the membrane-associated cyclase, NisC couples a cysteine residue to the dehydrated serine, thus forming a thioether (Li *et al.* 2006). The NisB and NisC proteins form an unstable lantibiotic synthase complex with a putative peptide translocator protein NisT. The ratio of the three proteins in the complex is one molecule of NisB to two molecules each of NisC and NisT (1:2:2) (McAuliffe *et al.* 2001).

Before nisin can exert its antimicrobial activity the leader peptide must be cleaved and the mature, active lantibiotic translocated across the cell

membrane. Cleavage of the leader peptide occurs through the proteolytic action of the peptidase NisP, a 682 amino acid extracellular protein (McAuliffe *et al.* 2001) (Siegers *et al.* 1996). Inactivation of the lantibiotic biosynthesis peptidase, *epiP*, showed that the protein is not essential to the production of a mature, functional lantibiotic peptide, suggesting that an alternative peptidase can compensate for the loss of *epiP* (Geissler *et al.* 1996). Conversely, the peptidase responsible for the leader protein cleavage of the Pep5 lantibiotic was shown to be essential for the production of mature and fully functional Pep5 (Meyer *et al.* 1995).

The NisT translocator is an ABC transporter, which exports nisin out of the cell. This transport is proposed to be mediated by the leader peptide and is reliant on the presence of active NisT. In the case of other lantibiotics, for example the aforementioned Pep5, the lantibiotic is still exported from the cell with reduced efficiency in the absence of the Pep5 translocator protein. This suggests that other ABC transporters encoded in the genome are able to compensate for the role of the lost Pep5 translocator (Meyer *et al.* 1995). This compensation is also observed with the lantibiotic epidermin (McAuliffe *et al.* 2001).

Lantibiotic biosynthesis is controlled by a TCS, NisKR, in response to changes in cell density, where nisin acts as a substrate for the sensor histidine kinase. The response regulator activates transcription of *nisA* in the presence of nisin, or NisA. Mutant nisin and nisin analogues can also induce the transcription of *nisA* (Kuipers *et al.* 1995).

The structure of nisin is characteristic of the type(I)-A lantibiotics, being composed of rigid lanthionine rings separated by areas of less rigidity. Nisin contains 5 lanthionine rings, and the N-terminal rings A and B are conserved in several lantibiotics within the same group including gallidermin (Wilson-Stanford *et al.* 2009). The basis of these lanthionine rings is the presence of thioether bridges. These thioether bridges are essential to the function of nisin since oxidation of the thioether moieties to sulfoxides results in the loss of its antimicrobial properties (Slootweg *et al.* 2014).

The mode of action of this lantibiotic involves interaction of nisin with lipid II, in the cell membrane that results in pore formation. Studies revealed that lipid II enhances the capacity of nisin to form pores in membrane model systems (Breukink *et al.* 1999), and that the presence of lipid II is required for the formation of a stable pore (Wiedemann *et al.* 2004). Furthermore, the conserved N-terminal lanthionine rings of nisin have been shown to be the lipid II binding site. In the presence of lipid II the N-terminus of nisin adopts a cage-like structure, involving the first two lanthionine rings, A and B, around the pyrophosphate group of lipid II. The interaction between nisin and lipid II is enhanced by the formation of hydrogen bonds between the pyrophosphate group and the binding cage. The presence of this pyrophosphate moiety in both lipid I and lipid II explains the interaction of nisin with both molecules (Hsu *et al.* 2004).

In addition to the action of forming pores in the bacterial membrane nisin sequesters lipid II and removes it from its functional location thus blocking cell wall synthesis. This mode of action also results from the presence of the pyrophosphate cage created by the A and B lanthionine rings, thus other lantibiotics where this motif is conserved also exibit this action (Hasper *et al.* 2006).

5.1.2 Biosynthesis, structure and function of type-A(I) lantibiotics produced by staphylococci

Lantibiotic production in the staphylococci is dominated by the commensal species. Several of these antimicrobial peptides are type-A(I) lantibiotics. The type-A(I) lantibiotics can be further divided into three structural groups; the nisin group, the pep5 group and the epidermin group (Chatterjee *et al.* 2005). *S. epidermidis* produces two such antimicrobial peptides, Pep5 and epidermin. Pep5 was isolated from *S. epidermidis* strain 5, and its biosynthesis gene cluster contains the *pepA* structural gene, *pepB, pepC* and *pepP* which encode enzymes responsible for post-translationally modifying and processing the prepeptide, *pepT* which encodes the PeP5 translocator and *pepI* which, when expressed,

confers producer immunity (Meyer *et al.* 1995). Pep5 biosynthesis is encoded by the plasmid pED503 and the biosynthesis gene clusters is arranged *pepTIAPBC* (Meyer *et al.* 1995). A key structural difference in Pep5 is the N-terminally exposed didehydroaminobutyric acid following cleavage of the leader peptide. This is an unstable moiety which is quickly deaminated to form 2-oxobutyryl (McAuliffe *et al.* 2001). Like nisin, Pep5 is able to form pores in Gram-positive cell membranes (Kordel *et al.* 1988). Pep5 also weakens the cell wall by counteracting the inhibition of lytic enzymes, which regulate cell wall turnover, by their inhibitors that include teichoic and lipoteichoic acids (McAuliffe *et al.* 2001). In contrast to nisin, however, Pep5 has low affinity for the peptidoglycan precursor lipid II (Brötz *et al.* 1998).

Epidermin is produced, modified and transported by a gene cluster similar to that of Pep5 encoded by the plasmid pTu32 in the *S. epidermidis* strain Tü3298(Augustin *et al.* 1992) . The epidermin biosynthesis gene cluster contains the additional genes *epiD* and *epiQ,* encoding an oxidative decarboxylase (Bierbaum *et al.* 1996), and a regulator protein respectively (McAuliffe *et al.* 2001) (Götz *et al.* 2014); there is also a producer immunity gene cluster. Gallidermin, synthesised by *S. gallinarum,* is similar to epidermin and Pep5. The gallidermin biosynthesis cluster is organised as that of epidermin and is borne on a large plasmid. Regulation of epidermin and gallidermin occurs through the transcription of *epiQ/gdmQ,* which is not activated by mature lantibiotic, as is the case for nisin. The *epiQ* promoter which activates transcription of the *epiP* and *epiQ* operon is not itself controlled by EpiQ, therefore epidermin/gallidermin biosynthesis is not autoregulated (Götz *et al.* 2014). Expression of the additional *epiD* gene and its homologue *gdmD,* encoded in the epidermin and gallidermin biosynthesis clusters, structurally alters the mature lantibiotic peptide to contain an S-((Z)-2-aminovinyl)-D-cysteine modification that is lacking in nisin and Pep5 due to the absence of an *epiD* homologue (Götz *et al.* 2014).

Both epidermin and gallidermin are synthesised as prepeptides and their leader peptides are cleaved by the leader peptidase. EpiP and GdmP are both serine

proteases which contain a signal peptide which controls their secretion. As is the case for nisin, the post-translocation processing of the prepeptide by the peptidase is essential for its maturation and function.

Due to the shorter length of epidermin and gallidermin compared to the well-studied nisin and Pep5, their mode of action is proposed to differ. The interactions with lipids I and II involved in murein synthesis is conserved, however epidermin and gallidermin interact differently with WTA precursors and these interactions may be the cause of bacterial killing by these two lantibiotics rather than pore formation (Hasper *et al.* 2006). Gallidermin is not able to form pores by interacting with lipid II, and in fact interaction of gallidermin with this molecule prevents pore formation by nisin *in vitro* (Müller *et al.* 2012). It is likely that the short lantibiotic peptides gallidermin and epidermin act by sequestering lipid II and preventing cell wall synthesis in this way (Hasper *et al.* 2006).

5.1.3 Other lantibiotics produced by staphylococci

The lantibiotic hominicin has been purified from *S. hominis* strain MBBL 2-9 and its structure contains those uncommon amino acids characteristic of the type-A(I) lantibiotics, however it has no thioether bridges integral to the lanthionine ring structures (Kim *et al.* 2010) of the type-A(I) lantibiotics. This implies that the mode of action of hominicin will not follow that of nisin; that is it will not interact with lipid II through a pyrophosphate cage structure.

Nukacin ISK-1, a type-A(II) lantibiotic, was isolated from *S. warneri* strain ISK-1 and its propeptide is similar to lacticin type lantibiotics. The nukacin ISK-1 biosynthesis operon appears to be comprised of a structural gene, *nukA,* a modification enzyme expressed from *nukM* and a response regulator (Sashihara *et al.* 2000). The ring structures characteristic of type-A(I) lantibiotics are also present in nukacin ISK-1 and ring A was shown to act as a lipid II binding site (Islam *et al.* 2012). Interaction of nukacin ISK-1 with the target membrane is due to electrostatic interactions mediated by three N-terminal lysine residues. Mutation of these residues was shown to attenuate the antimicrobial activity of

nukacin (Asaduzzaman *et al.* 2006). Variants of nukacin are produced by other staphylococci.

Nukacin 3299 was isolated from *S. simulans* involved in bovine mastitis in Brazil. The lantibiotic was originally known as simulancin 3299, however gene sequencing of the structural gene revealed 100% identity with *nukA* of the nukacin ISK-1 biosynthesis cluster. Similar to other lantibiotic biosynthesis clusters the nukacin 3299 biosynthesis cluster is plasmid encoded and this isolate demonstrated a wider range of host producer (Ceotto *et al.* 2010).

S. hominis also produces a nukacin variant. Nukacin KQU-131 is produced by *S. hominis* KQU-131, which was isolated from fermented fish. The lantibiotic prepeptide has 3 differing amino acid residues when compared to nukacin ISK-1: 1 in the leader peptide and 2 in the mature lantibiotic. The three lysine residues proposed to be important in membrane interaction are conserved, and the two amino acid changes in the mature lantibiotic are found in the last lanthionine ring of the structure(Wilaipun *et al.* 2008).

As well as the type-A(I) lantibiotics epidermin and Pep5, *S. epidermidis* produces several other lantibiotics; epicidin 280, epilancin K7 and epidermicin NI01. Epidermicin was the most recently characterised of the three and is a type-(II) lantibiotic. *S. epidermidis* strain 224 encodes the epidermicin NI01 biosynthesis cluster on a plasmid (Sandiford & Upton 2012). Epilancin K7 is a type-A lantibiotic, and contains the characteristic thioether bridges. Unlike the type-A(I) lantibiotics discussed, however, the peptidase associated with the epilancin K7 biosynthesis cluster does not carry an export signal, implying that it acts inside the producer cell. Epilancin K7 may therefore be exported as a mature lantibiotic peptide. The exposure of the N-terminus, cleaved of its leader peptide, to the cytoplasm could be the reason for the unusual modification of the first residue of the lantibiotic (van de Kamp *et al.* 1995). Finally, Epicidin 280 is produced by *S. epidermidis* BN 280. Both the lantibiotic peptide and its biosynthesis cluster exhibit a high degree of similarity to Pep5. Differences in structure of the biosynthesis cluster are however present as the ABC

transporter associated with high yields of Pep5 was absent, and an additional modification gene, *epiO* was found. The epicidin gene cluster was present on the larger plasmid pCH01. The lack of an associated transporter and evidence from Pep5 that deletion of the dedicated Pep5 transporter reduces, but does not abolish the production of the lantibiotic, suggests that an alternative transporter may export epicidin. Finally, strains which produce Pep5 or epicidin 280 exhibit cross immunity to the two peptides (Heidrich *et al.* 1998). This suggests that bacteria within niche communities producing similar bacteriocins might exhibit cross immunity to one another's antimicrobial peptides.

5.1.4 Producer immunity to lantibiotics

The mechanism of *L. lactis* immunity to the lantibiotic nisin is well understood. Several proteins are involved in immunity to nisin; the ABC transporter encoded by *nisFEG* and the membrane bound immunity protein NisI. This is also true of epidermin, gallidermin and nukacin immunity which also utilise this two factor resistance. Pep5, epilancin and epicidin immunity each require only one immunity peptide; PepI, EciI and ElxI respectively (Alkhatib *et al.* 2012). The ABC transporter involved in the immunity to nisin is in addition to the transporter responsible for the translocation of the peptide out of the cell. NisF is the nucleotide binding domain of the transporter, located on the cytoplasmic side of the cell membrane and the intramembrane subunits are NisE and NisG (Alkhatib *et al.* 2012). This transporter architecture is characteristic of importers however this ABC transporter expels unmodified nisin from the membrane before the occurrence of pore formation. Interaction of the ABC transporter with its substrate is mediated by the C- terminus of nisin, including the last lanthionine ring (Alkhatib *et al.* 2014). Expression of this ABC transporter occurs in the presence of nisin as a result of the phosphorylation of the histidine kinase NisK by the regulator protein NisR.

NisFEG contributes only around 20% of immunity to nisin, the remainder is contributed by the NisI immunity protein. NisI is a lipoprotein which contains an N-terminal cell wall anchor motif, however this cell wall anchoring is inefficient and around 50% of the lipoprotein is secreted into the extracellular

space (Koponen *et al.* 2004). This family of protein is poorly understood and does not share a great degree of similarity with other proteins; the hydrophobic N-terminal region is however shared with other lantibiotic immunity peptides such as PepI, which gives Pep5 producer strains of *S. epidermidis* resistance. A 21 amino acid portion of NisI provides specificity of other immunity peptides to nisin, and is responsible for most of the immunity against nisin conveyed by NisI (Takala & Saris 2006). Similarly the C-terminus of PepI is responsible for Pep5 immunity in *S. epidermidis* (Hoffmann *et al.* 2004). NisI is suggested to sequester aggregates of nisin to prevent them from interacting with the producer cell membrane. This ability to interact with nisin is not dependent on being anchored to the cell wall, therefore the NisI secreted into the extracellular space may also have a role in producer immunity (Koponen 2004) (Takala *et al.* 2004). Furthermore, NisI secreted into the extracellular space is present in a ~1:50 ratio with aggregated nisin further enhancing activity of the lantibiotic against producer strains *in vitro* (Koponen *et al.* 2004).

5.1.5 Experimental evolution and its role in elucidating antimicrobial resistance

Evolution of resistance to antibiotics by bacteria is one of the most significant problems facing medicine. Understanding the potential for resistance to new antibiotics and the mechanisms of emerging resistance to current antibiotics is required to develop effective, future proof antimicrobials. Experimental evolution is one method that has been used to investigate the potential for resistance to newly discovered antimicrobials, and also to infer information about the mechanism by which this resistance may have occurred.

Acquired antimicrobial resistance has been explored in several ways, such as using clinical isolates from antibiotic resistant infections (Mwangi *et al.* 2007) (Howden *et al.* 2011) as well as experimentally evolved isolates (Renzoni *et al.* 2011) (Song *et al.* 2013). There are advantages and disadvantages to both approaches. In an experimental evolution model the variables are tightly controlled so any variation in phenotype and genotype can be attributed to the antimicrobial under test. A model which uses clinical isogenic pairs of strains to

examine evolution of resistance instead, has the advantage of including host or community factors which might affect resistance. Such factors are synergistic effects of other antimicrobials or acquisition of resistance by horizontal gene transfer from community members. Regardless of the model used, minimum inhibitory concentration (MIC) assays are employed to identify the presence and understand the extent of antibiotic resistance developed.

Resistance to antimicrobial peptides often arises through mutations which may occur in a sequential fashion, that is one mutation may be necessary for the next to be acquired. Resistance to some groups of antibiotics, however, is the result of a single mutation, as is the case for rifamycin and quinolones (Toprak *et al.* 2012). In most laboratory systems of evolution this series of mutations is selected for by maintaining the mutant selection window, which requires the concentration of the antibiotic being used as the selection pressure to be continuously increased as mutations conferring resistance occur. The rate at which resistant mutants occur leading to the concentration of the antibiotic to be increased changes for each organism and antimicrobial (Drlica 2003).

Where resistance determinants are encoded by elements, which can be gained through horizontal gene transfer, the story is slightly different. Multidrug resistance plasmids carry resistance determinants to a broad range of antimicrobials and these plasmids represent a fitness cost for the host bacteria in the absence of the selective pressure of antibiotics. This has been investigated through laboratory evolution of mixed populations of *E. coli* comprised of organisms with and without plasmid-encoded resistance to antimicrobials introduced as a selection pressure. It was shown that the levels of antibiotics required to maintain these resistant determinants, especially where multiple selection pressures for multidrug resistance plasmids are present, are far below the levels of MIC and may be found in certain environments such as hospital settings and agriculture (Gullberg *et al.* 2014).

5.1.5 Chapter aims

The staphylococci live in complex communities as part of the skin microbiota. To compete for resources such as space and nutrients the members of these communities employ an arsenal of weapons such as the production of lantibiotic bacteriocins. Data from chapter 4 led to the hypothesis that the way in which different groups of staphylococci combat these lantibiotics may be different through the divergence of the *graRS* operon and the presence of an additional orthologous operon, *braRS* between groups of species. BraRS has previously been shown to be important in lantibiotic resistance in *S. aureus*. Since the involvement of BraRS in nisin resistance has thus far only been investigated in the pathogenic species *S. aureus* the first aim of this chapter was to determine if this involvement extended to those other commensal species of staphylococci bearing the *braRS* operon. The responses of strains of *S. hominis* and *S. aureus* to the type-A(I) lantibiotic nisin were therefore compared using experimental evolution with analysis of resulting SNPs to understand if nisin resistance in the two species occurs through similar mechanisms.

A further aim of this chapter was to use this experimental evolution model to investigate if species without BraRS evolved resistance to nisin to the same extent as those species which carry *braRS*. The resulting SNPs in the evolved populations were analysed to understand any differences in the pathways to increased nisin resistance in the two groups. To this end the evolution of the *braRS*-positive species *S. hominis* and *S. aureus* to nisin was compared with that of the *braRS*-negative species *S. saprophyticus*.

5.2 Methods

5.2.1 Preparation of nisin stock solutions

For all experiments where preparation of nisin stocks was required, master stocks were prepared at 20, 2 and 0.2 mg/mL in 10mM sodium citrate at pH 3 and stored at 4 °C and allowed to reach room temperature before use.

5.2.2 Minimum inhibitory concentration assay

A microtiter plate method was used with doubling dilutions of a nisin stock solution prepared at 2X the highest concentration required in sodium citrate and diluted to the appropriate concentration using BHI. Concentrations were then diluted 1 in 2 with bacterial suspension. The final concentrations used are found in Table 5.1. Optical densities of overnight cultures of each strain tested were adjusted to OD_{600} 0.2 ± 0.005 and 100 µL was used to inoculate the microtiter plates. The lowest concentration with an optical density ≤ to that of the initial optical density was taken to be the minimum inhibitory concentration (MIC).

With thanks to Dr Miriam Korte-Berwanger, University of Bochum, for kindly providing the *Staphylococcus saprophyticus* strains used in this study.

Table 5.1 Concentrations of nisin used in the nisin MIC assay. All concentrations given in μg/mL.			
S. aureus	*S. hominis*	*S. saprophyticus* (excluding strain 349)	*S. saprophyticus* strain 349
500	50	500	50
400	40	400	40
300	30	300	30
250	25	250	25
200	20	200	20
150	15	150	15
125	12.5	125	12.5
100	10	100	10
75	7.5	75	7.5
62.5	6.25	62.5	6.25
50	5	50	5
37.5	3.75	37.5	3.75
31.25	3.125	31.25	3.125
25	2.5	25	2.5
18.75	1.875	18.75	1.875
15.625	1.5625	15.625	1.5625
12.5	1.25	12.5	1.25
9.375	0.9375	9.375	0.9375
7.8125	0.78125	7.8125	0.78125
6.25	0.625	6.25	0.625
4.6875	0.46875	4.6875	0.46875
3.90625	0.390625	3.90625	0.390625
3.125	0.3125	3.125	0.3125
2.34375	0.234375	2.34375	0.234375
1.953125	0.1953125	1.953125	0.1953125
1.5625	0.15625	1.5625	0.15625
1.171875	0.1171875	1.171875	0.1171875
0.9765625	0.09765625	0.9765625	0.09765625
0	0	0	0

5.2.3 Experimental evolution and selection

Experimental evolution was carried out using a microtiter plate method. Plates were set up with doubling dilutions of nisin (prepared as described in section 4.2.1) in triplicate across the microtiter plate. Concentrations used for experimental evolution are shown in Table 5.2. The initial stock solution concentration of nisin used for each strain was selected based on the results of the MIC assay and produced a set of doubling dilutions which encompassed the MIC value observed. For *S. aureus* and *S. saprophyticus* this initial stock concentration was 60% 0.02 g/10mL. For *S. hominis* this was 0.002 g/10mL. Sodium citrate control plates were set up using the same method but omitting nisin so the control experiment reflected the concentration of sodium citrate in the nisin experiment. All plates were incubated static at 37 °C. The first plates were inoculated with OD_{600} 0.2 bacteria for the first passage and bacteria from the well with the highest concentration of nisin which supported growth consistent with the control after 24 or 48 hours was passaged forward to the next plate. All wells of subsequent passages were inoculated with a 1:1000 dilution of the bacteria passaged forward. This passaging was repeated to a terminal concentration of 10 000 μg mL^{-1} of nisin or for a period of 12 days. Strains were stocked in 20% glycerol after each passage. The T_0 time point was also stocked as a comparator strain.

Table 5.2 Concentrations of nisin used in the experimental evolution. All concentrations are given in μg/mL, percentages refer to a starting stock prepared from a percentage dilution of the master stock.

well	60% 0.2	100% 0.2	60% 0.02	100% 0.02	60% 0.002	100% 0.002
2	3000	5000	300	500	30	50
3	1500	2500	150	250	15	25
4	750	1250	75	125	7.5	12.5
5	375	625	37.5	62.5	3.75	6.25
6	187.5	312.5	18.75	31.25	1.875	3.125
7	93.75	156.25	9.375	15.625	0.9375	1.5625
8	46.875	78.125	4.6875	7.8125	0.46875	0.78125
9	23.4375	39.0625	2.34375	3.90625	0.234375	0.390625

Time points for sequencing were selected by using the highest concentration of nisin where each of the three independent biological replicates reached the same level of nisin resistance. Stocks from each time point were streaked on BHI agar and incubated overnight at 37 °C . Clones from each repeat were selected and cultured once more without nisin by inoculating 10 mL of BHI with the clones and incubating at 37 °C with shaking at 200 rpm overnight. Increased MICs were then confirmed by altering the MIC assay described above to include the end point MIC and clones with MICs similar to each other and the end point MIC were selected.

This process was repeated for the corresponding control time point for each of the three species.

5.2.4 Analysis of nisin selection SNPs

DNA was extracted from clones selected from each experimental and control repeats from the method described above using the protocol described (section 2.2). 30 ng of DNA from each of five selected clones was pooled in a volume of 55 μL to make up the 150 ng of DNA required to make Illumina Truseq DNA libraries with an insert size of 350 bp. In addition to three separate clone pools, a single clone was selected for sequencing from the clones used to constitute the pools. Single clones were selected on the basis of the highest DNA quality. The single clones and the T_0 isolates were also sequenced using Illumina Truseq nano DNA libraries with 350 bp inserts. Single clones were not sequenced from the control pools. DNA quality checking was carried out according to the protocol outlined in (section 2.3) and libraries were constructed by the CGR, University of Liverpool, according to the protocol laid out in (section 2.5

Sequencing reads were trimmed and quality checked using the method from (section 2.6). T_0 comparator strains were assembles using VelvetOptimiser (Victoria Bioinformatics Consortium) with Kmer sizes from 19 to 99 and Velvet version 1.2.06 (Zerbino & Birney 2008). Annotation was carried out using

PROKKA (Victoria Bioinformatics Consortium). The PacBio assembly of *S. hominis* strain J31 was used as the comparator assembly for this strain.

Good quality filtered reads from experimentally evolved pools and single clones were then aligned to the relevant comparator strains using the BWA, a Burrows-Wheeler aligner (H. Li & Durbin 2009), packages aln and sampe version 0.5.9-r16. SAM files were converted to bcf (binary variant call files) files with samtools which could then be used for SNP calling using the mpileup package. The bcf output file from mpileup was then converted to vcf (variant call format) files and quality filtered. This pipeline was run from the shell script *SNP_analysis.sh*. For pools of clones this quality filtered vcf file along with the mpileup output without base data generated from SNP_analysis.sh were then used to further filter the SNPs to include only those present in 33.33% of reads, which equates to the SNP being present in more than one clone. This further filtering was achieved using the script *significant_snps.sh*.

5.2.5 Modelling proteins containing SNPs

Potential protein domains were identified by aligning the amino acid sequences to Protein Data Bank archive (PDB) and the Conserved Domains Database (CDD) using HHPred (Söding *et al.* 2005) as well as searching the Pfam protein families database (Finn *et al.* 2014). Proteins, excluding membrane proteins and those with predominantly coiled coil structures, were then modelled using the alignment from HHPred. The structures were explored, and figures generated using The Pymol Molecular Graphics System v 1.7.4.4 (Schrödinger, LLC).

5.3 Results and discussion

5.3.1 Intraspecies variation in susceptibility of staphylococci to the lantibiotic nisin

To select strains with a representative baseline susceptibility to nisin the MIC of a panel of *S. aureus*, *S. hominis* and *S. saprophyticus* strains was investigated. A high degree of intraspecies variability was observed for *S. aureus* (Figure 5.1) and *S. saprophyticus* (Figure 5.2) with the nisin MIC ranging from 31.25 to 125 μg/mL for both species. The MIC of *S. aureus* agrees with work by Hiron *et al* who found the nisin MIC of strain HG001 to be >128 μg/mL (Hiron *et al.* 2011). These results however differ significantly from those of Blake *et al* who found that the nisin MIC of SH1000 was 4 μg/mL. Their study does not, however, give details of the method used, and the work presented here used a method similar to that of Hiron *et al*. Differences in MIC method may therefore account for some of the variation between the experiments. Little intraspecies variation was observed among the isolates of *S. hominis* (Figure 5.3) tested, which ranged from 12.5 to 15 μg/mL.

The mean MIC was determined for all three species and an isolate as close as possible to that value was selected for experimental evolution. Type strains of all three species were also included for experimental evolution (Table 5.3).

Table 5.3 Summary of nisin MICs for *S. aureus, S. saprophyticus* and *S. hominis*. End point nisin MICs represent the MIC following experimental evolution in the presence of nisin.

Species	Average MIC (μg/mL)	Isolates	Starting MIC (μg/mL)	End point MIC (μg/mL)
S. aureus	66.04	171	50	10000
		SH1000	100	10000
S. saprophyticus	71.72	CCM883	62.5	10000
		349	40	750
S. hominis	14.58	type	12.5	375
		J31	15	375

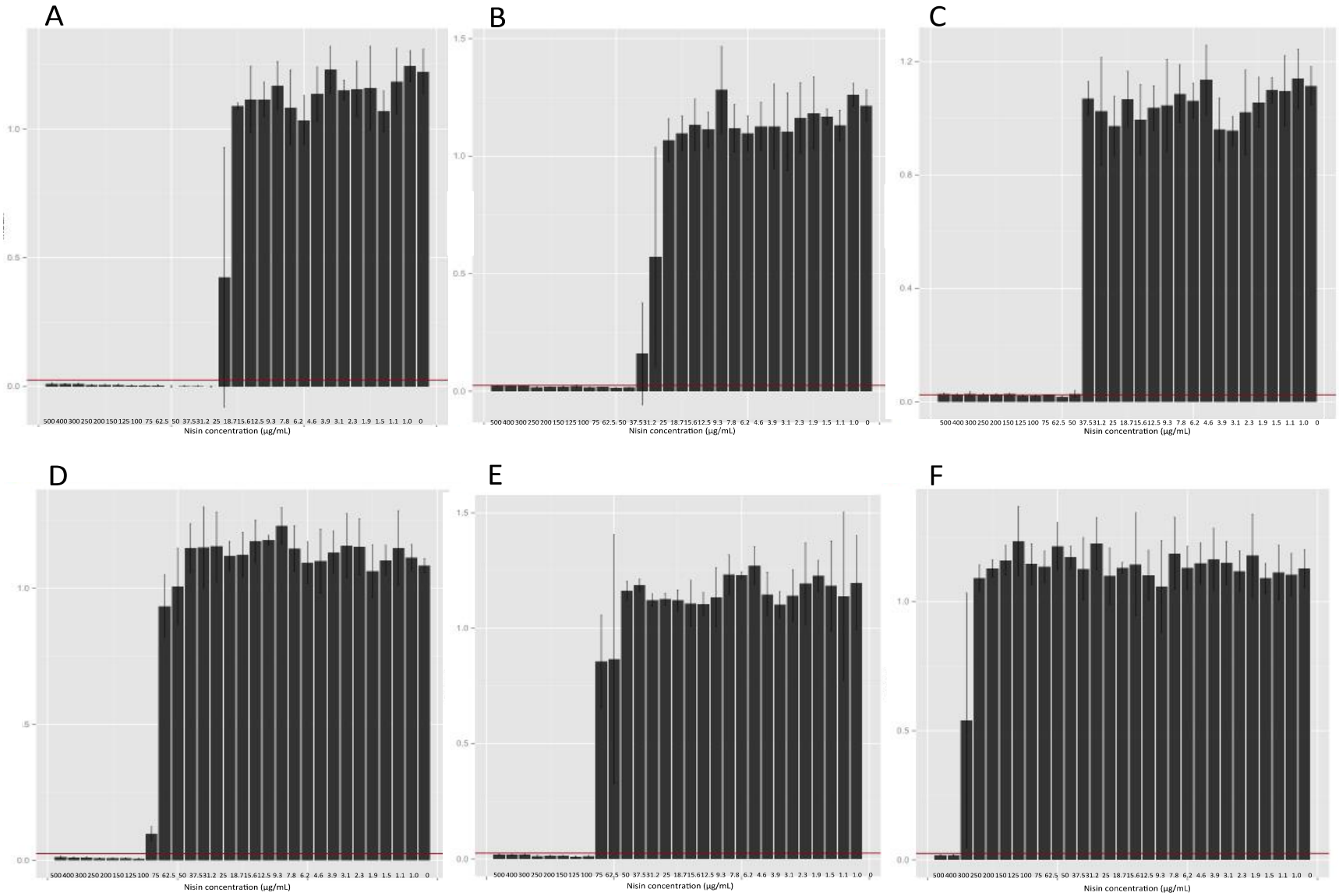


Figure 5.1 Nisin MICs of a panel of *S. aureus* isolates. Strains A) 059 B) 171 C) 014 D) Mu50 E) SH1000 F) MSSA476. Error bars represent the standard deviation of a data set; the red line represents the optical density threshold for bacterial growth.

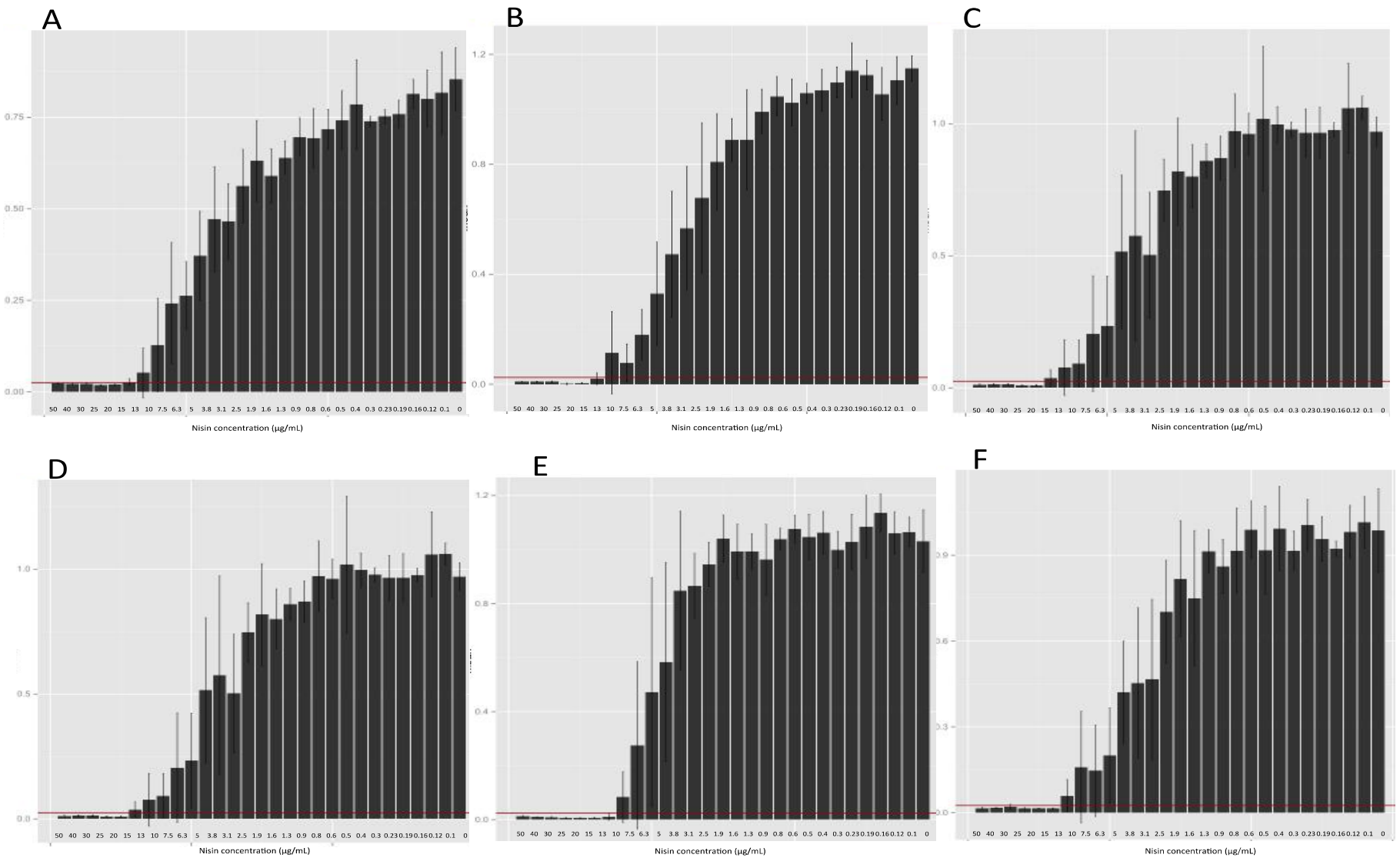


Figure 5.2 Nisin MICs of a panel of *S. hominis* isolates. Strains A) B10 B) I4 C) J6 D) J11 E) *S. hominis* type strain F) J31. Error bars represent the standard deviation of a data set; the red line represents the optical density threshold for bacterial growth.

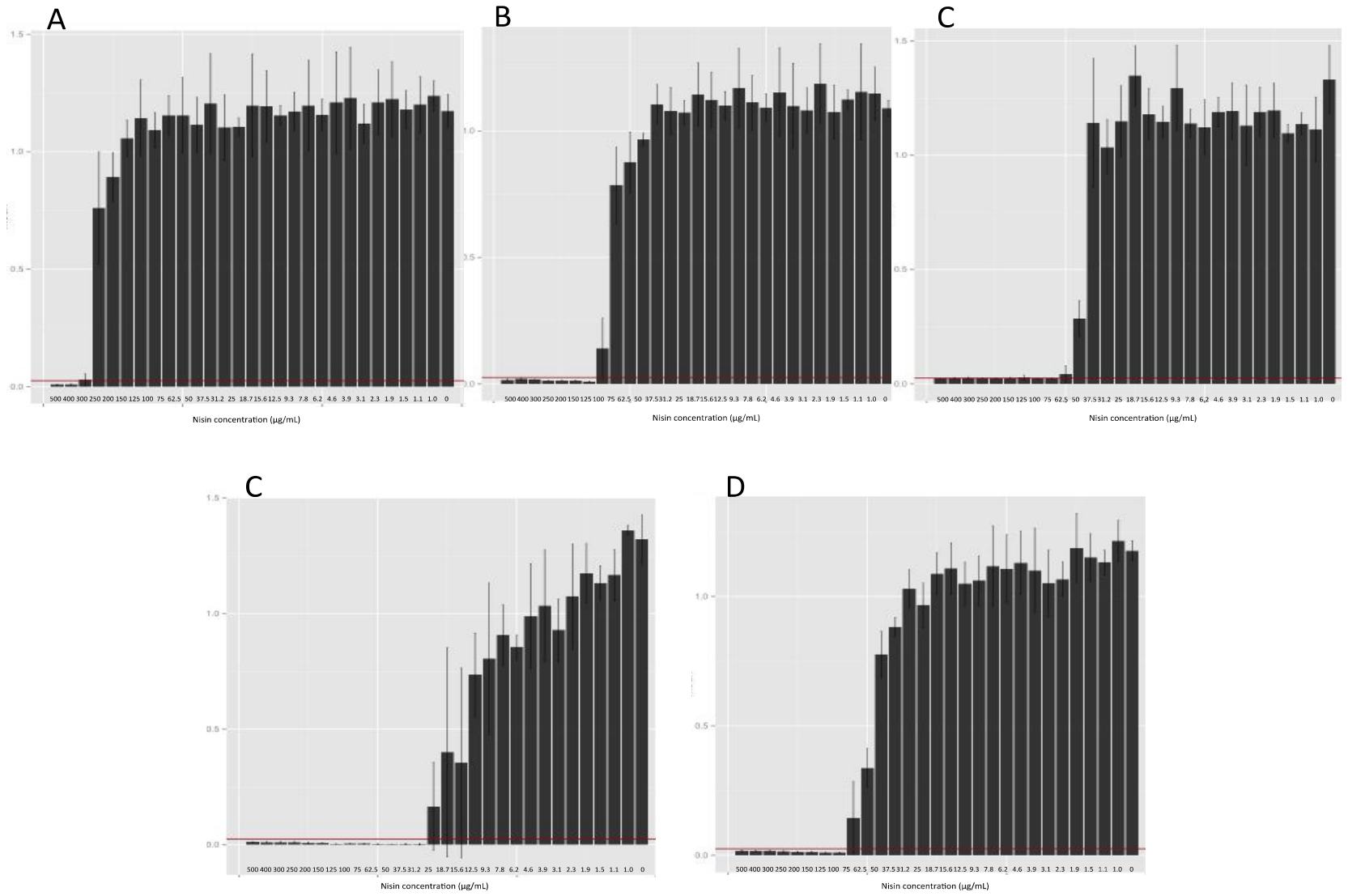


Figure 5.3 Nisin MICs of a panel of *S. saprophyticus* isolates. Strains A) 349 B) 346 C) CCM883 D) 352 E) 409. Error bars represent the standard deviation of a data set; the red line represents the optical density threshold for bacterial growth.

5.3.2 Experimental evolution of *S. aureus*, *S. hominis* and *S. saprophyticus* in the presence of nisin.

All strains selected for experimental evolution were serially passaged in triplicate with increasing concentrations of nisin using a microtiter plate method. Since the nisin was dissolved in sodium citrate, a control experiment was performed in parallel with evolution in the presence of nisin. In this control experiment cells were serially passaged in the equivalent concentration of sodium citrate present in the nisin plates. In this way SNPs selected by passage in the presence of sodium citrate could be distinguished from those selected from passage in the presence of nisin.

Microtiter plates with doubling dilutions of nisin across the plate were used for the serial passages. A fresh plate was used for every passage and all dilutions of nisin were inoculated for every passage. An increase in resistance was taken as comparable growth of a strain to the wild type in a higher concentration of nisin than the previous passage. Strains were passaged once every 24-48 h. Stepwise increases in nisin MIC were observed for all strains tested. No obvious pattern in the rate of resistance acquisition was observed between the species.

Both strains of *S. aureus* exhibited ~100 fold increases in their nisin resistance (Figure 5.5). This is a greater fold increase in *S. aureus* resistance than that observed by Blake *et al*, however there are several differences in the strategies used for experimental evolution between experiments and so results can not easily be compared. Firstly it is not known if their serial passage was carried out in microtiter plates and small volumes, as is the case in this study, or if it was carried out in larger culture volumes such as flasks. MICs do not easily translate between plate-based methods and methods using larger culture volumes. The clinical isolates of *S. saprophyticus* 883 and 349 showed an 80 fold and 4.7 fold increase in nisin resistance respectively. Each of the *S. hominis* strains exhibited around 25-fold increases in their nisin resistance.

The choice of strains to sequence was relatively arbitrary (isolates sequenced are indicated by an asterisk in Figures 5.4 and 5.5): strains J31 and 171 were selected for sequencing post-evolution based upon both being recent skin isolates and therefore deemed to be more representative than the archive *S. hominis* and *S. aureus* type strains tested. To compare mechanisms of increased resistance to nisin across species with and without BraRS, the *S. saprophyticus* strain CCM883 was included for sequencing due to the fold change in nisin resistance being comparable to *S. aureus* and intermediate with *S. hominis* (Figure 5.6).

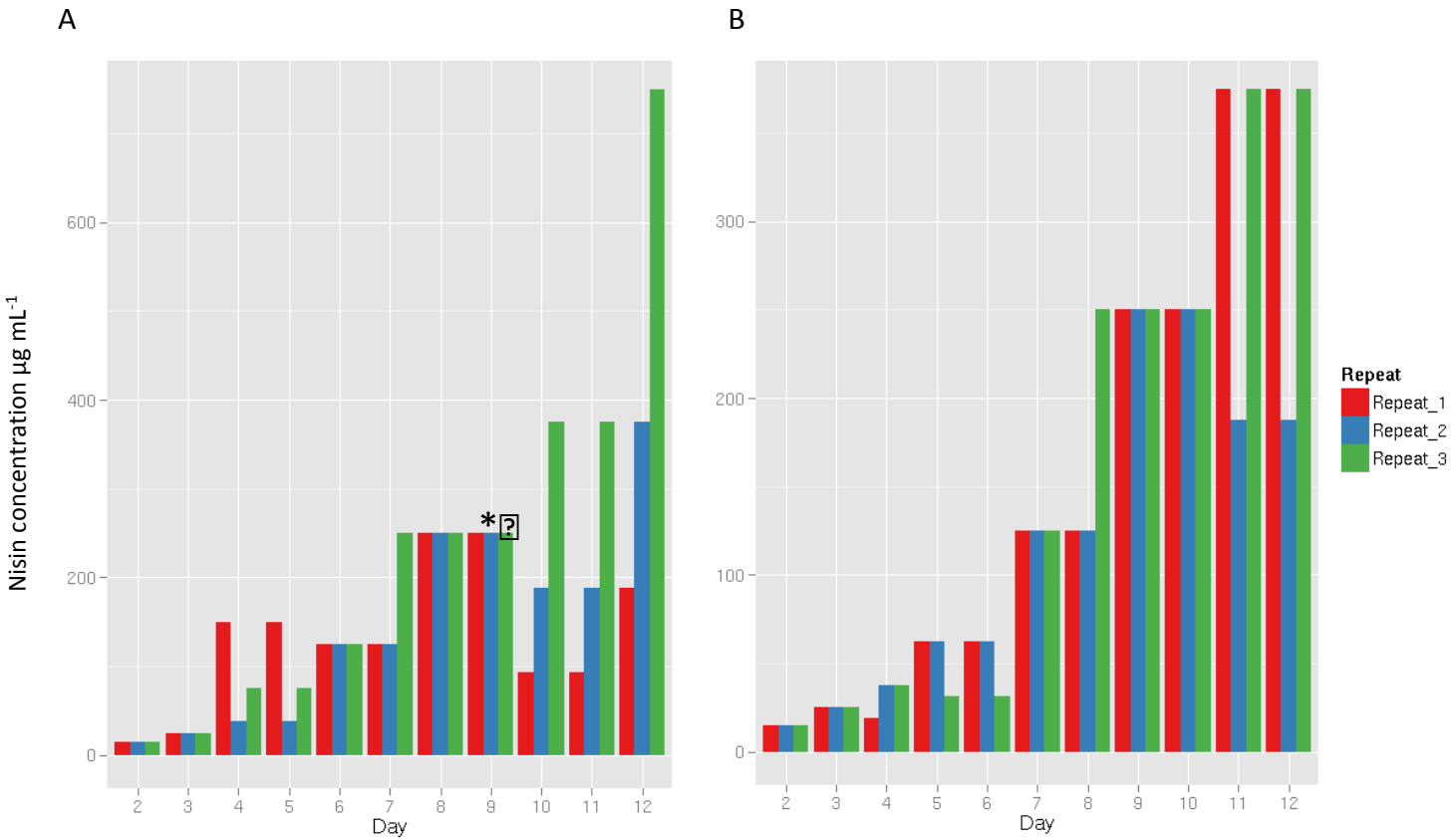


Figure 5.4 Nisin MICs of evolved *S. hominis* strains A) J31 B) HT. Biological repeats are indicated by red, blue and green bars. Isolates selected for sequencing are indicated by asterisks.

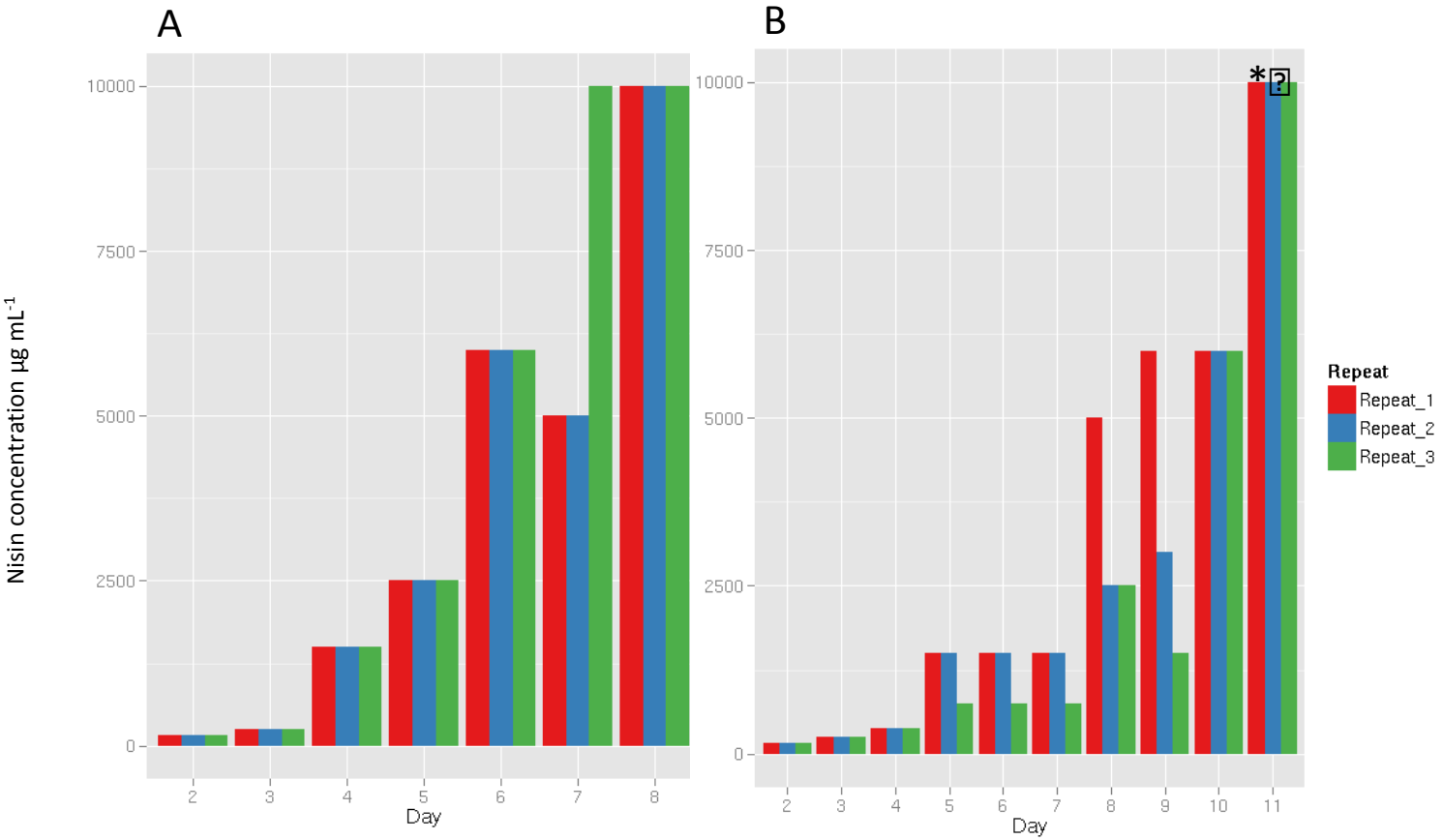


Figure 5.5 Nisin MICs of nisin-evolved *S. aureus* strains A) SH1000 B) 171. Biological repeats are indicated by red, blue and green bars. Isolates selected for sequencing are indicated by asterisks.

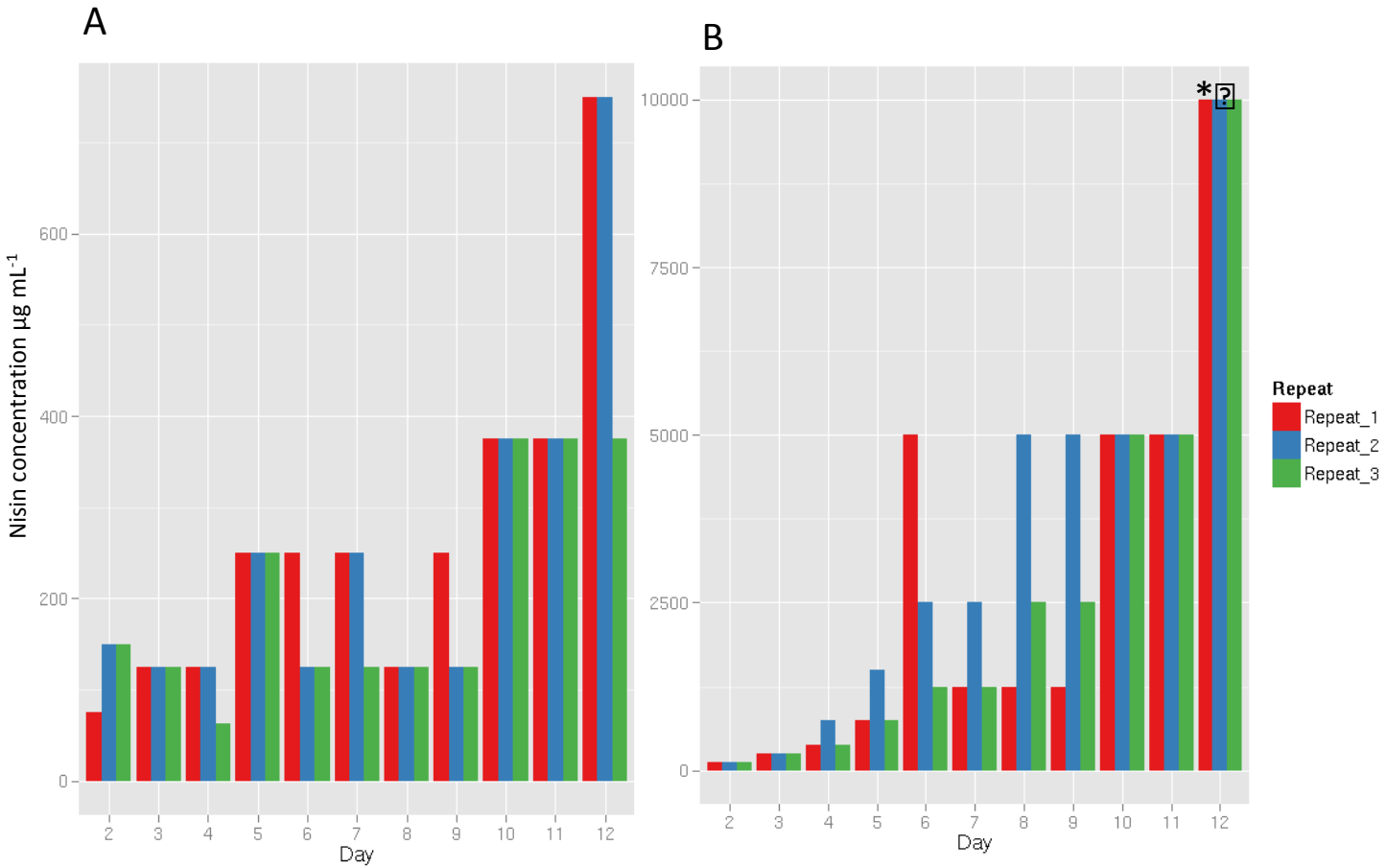


Figure 5.6 Nisin MICs of nisin evolved *S. saprophyticus* strains A) 349 B) 883. Biological repeats are indicated by red, blue and green bars. Clones selected for sequencing are indicated by asterisks.

5.3.3 Nisin selection SNPs in evolved strains

T0 sequenced isolates were assembled and annotated according to the previously described method (chapter 2, Methods). To identify SNPs present, reads from pooled samples and single clones evolved in nisin and the sodium citrate control pools were aligned to these T0 isolates in the case of *S. aureus* 171 and *S. saprophyticus* CCM883, and to the PacBio assembly in the case of *S. hominis* J31. SNPs present in both the control pool and the nisin-evolved pools and single clones were discarded, as these SNPs are unlikely to be due to selection from the presence of nisin. Following this, proteins in which significant SNPs occurred were modelled from the sequences in the reference sample genome and the SNP was imposed upon the structure. SNPs most likely to be involved in resistance to nisin are SNPs present in all three pools of evolved clones.

5.3.3.1 Nisin-selected SNPs of *S. aureus*

The selection for increased nisin resistance in *S. aureus* 171 resulted in only 1 SNP that was present in all three clone pools, and this SNP was located at chromosome position 1820274 in a gene encoding a hypothetical protein containing a transposase domain (Table 5.4a). This SNP was a non-synonymous point mutation within this domain resulting in the substitution of the amino acid leucine for proline. In addition, another SNP was present in more than one clone pool. A serine to threonine (S114T) amino acid change was caused by a base change at the chromosome position 2001512. This chromosome position is within a gene encoding a hypothetical protein of 120 amino acids in length. These 120 amino acids share 95% similarity with 120 amino acids of a larger (582 amino acids) *S. aureus* membrane-associated protein (accession GI486263885) with some homology to a membrane-associated protein, which contains an MHC class II analogue protein (MAP) domain where the amino acid change is located. This protein domain is present in multiple extracellular matrix proteins and facilitates protein-protein interactions (Jönsson et al. 1995). Due to the low percentage of reads covering the position of this base change it is likely that only 1 or 2 of the clones in each pool contained the SNPs, however

due to its presence in multiple pools the base change occurred in multiple independent evolution experiments therefore suggesting that they may be important in nisin resistance.

Table 5.4a Non-synonymous, homozygous SNPs from nisin-evolved isolates of S. aureus strain 171. Pools were comprised of 5 clones from each experimental repeat, for example pool 1 corresponds to experimental repeat 1. Nisin MICs of clones in each pool were confirmed to ensure they were not significantly different.

Prokka annotation	Blast similarity	Pools	Chromosome position	Base change	Amino acid change	% of reads
rpoC	DNA-directed RNA polymerase subunit beta	3	523690	G -> T	Ala448Ser	99.50980392
femB	*femB*	3	1323450	G -> A	Arg215His	63.20754717
phage terminase like protein	Terminase	3	1491745	C -> G	Gly240Ala	100
greA	*greA*	3	1625376	A -> T	Leu76*	37.27272727
hypothetical	Hypothetical protein	1,2,3	1820274	A -> G	Leu39Pro	39.83050847
N/A		3	1820533	T -> A		34.82142857
map-ND2C	Hypothetical protein	2,3	2001512	A -> T	Ser114Thr	34.7826087
braS	Two component sensor histidine kinase	3	2627088	G -> A	Thr175Ile	100
hypothetical protein	Membrane protein	2	615003	C -> T	Gln57*	34.54545455
hypothetical protein	Transposase	2	1337527	T -> C	Val37Ala	33.91304348
hypothetical protein	Hypothetical protein	2	1781422	A -> T	Leu21His	33.48017621
surface precursor protein	Virulence-associated cell-wall-anchored protein SasG	1,2	2481387	A -> G	Phe59Phe	34.09090909
walK	Sensor kinase walK	1	17119	A -> G	His364Arg	100
hypothetical_protein	Hypothetical protein	1	272718	T -> C	Asp23Asp	43.33333333
hypothetical_protein	Hypothetical protein	1	277837	C -> T	Asp23Asp	34.21052632
gltB_1	Glutamate synthase	1	437361	A -> T	Gln797Leu	100
rpoB	DNA-directed RNA polymerase subunit beta	1	520176	C -> T	His506Tyr	100
mraY	Phospho-N-acetylmuramoyl-pentapeptide transferase	1	1103071	G -> A	Val266Ile	100
N/A		1	1821311	G -> T		35
yurK	Transcriptional regulator	1	1998279	C -> T	Gly73Arg	100

Table 5.4b Non-synonymous, homozygous SNPs from nisin evolved clone of S. aureus strain 171. The S. aureus 171 single clone sequences was from pool 1

Prokka annotation	Blast similarity	Chromosome position	Base change	Amino acid change
walK	Sensor protein kinase walK	17119	A -> G	His364Arg
gltB_1	Glutamate synthase	437361	A -> T	Gln797Leu
rpoB	DNA-directed RNA polymerase subunit beta	520176	C -> T	His506Tyr
N/A		846611	A -> G	
mraY	Phospho-N-acetylmuramoyl-pentapeptide transferase	1103071	G -> A	Val266Ile
msrR	Regulatory protein msrR	1309845	G -> T	Glu181*
yurK	Transcriptional regulator	1998279	C -> T	Gly73Arg

Within each of the clone pools, SNPs were identified that were present in 100% of the reads covering their chromosomal location in a single pool indicating that the SNP was present in all five of the clones included in the pool. One of these SNPs was a G -> T base change at the chromosome position 523690 within *rpoC*, which resulted in an alanine to serine (A448S) amino acid change. Alanine and serine are structurally similar, however alanine is non-polar whereas serine is polar. This amino acid change is in domain 2 of the protein which contains the active site, furthermore the alanine to serine residue change appears to be within an NADFDGD motif. This motif is conserved across bacteria and functions to bind the magnesium ion within the active site via the three aspartate residues (Figure 5.7) (Cramer *et al.* 2001) (Zhang *et al.* 1999). Since metal ions catalyse the active site, altering the magnesium ion binding motif may affect the efficiency of the polymerase, and therefore the rate or accuracy of transcription. Other studies have described mutations in the subunits of RNA polymerase as a result of selective pressures. For example, it has been observed that a mutation in *rpoC* causing a glutamate to aspartate amino acid change is involved in resistance of *B. subtilis* to cephalosporin (Lee *et al.* 2013). The location of this mutation however is in the DNA binding cleft of the protein.

A second SNP in a gene encoding a transcription complex protein was identified with a C -> T mutation at chromosomal position 520176 in the *rpoB* gene. This SNP was present in 100% of reads covering the chromosome position from pool 1 and the *S. aureus* 171 single clone which replaces a histidine with a tyrosine (H506Y). Histidine is a basic amino acid and tyrosine is aromatic and polar. This amino acid change is within domain 3 of the protein know as the fork domain which forms the border between the β and β' domains. This domain is highlighted in yellow in the 3D model of the protein (Figure 5.8). Mutations within a 25 amino acid region of this domain of RpoB from *E. coli* are associated with slippage at the RNA-DNA hybrid during transcription which reduces the

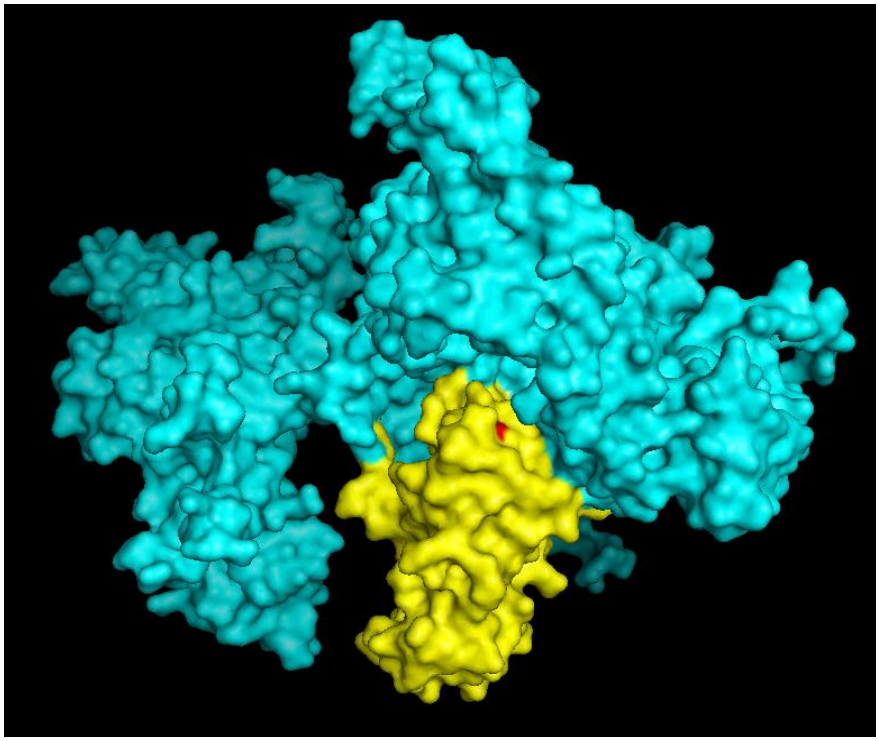


Figure 5.7 3D model of *S. aureus* RpoC with the highlighted substitution Ala448Ser. Domain 2 is highlighted in yellow and the SNP-directed substitution Ala448Ser is highlighted in red.

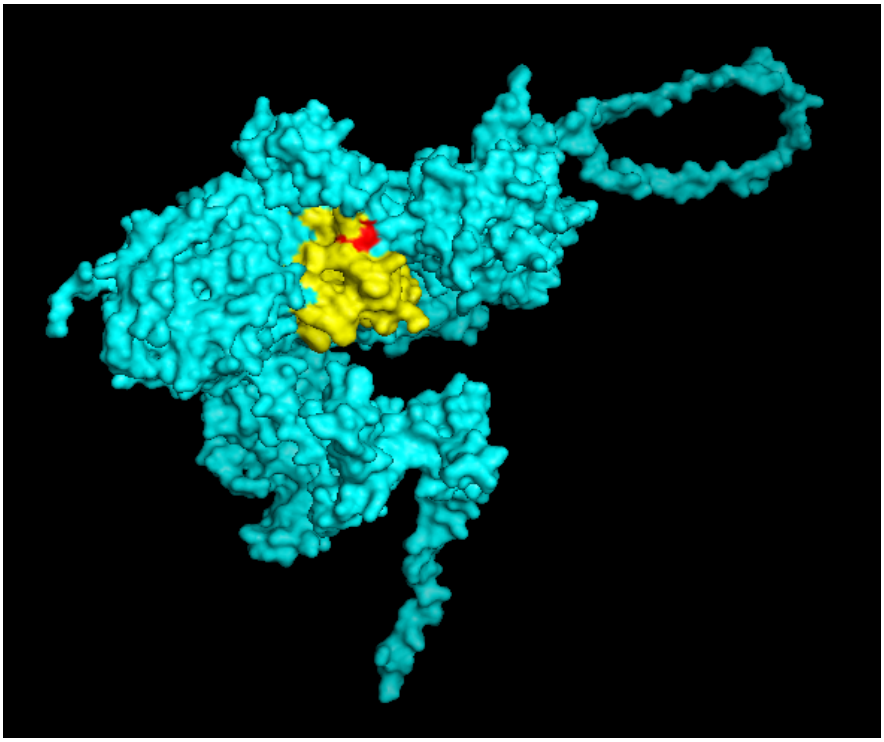


Figure 5.8 3D model of *S. aureus* RpoB with the highlighted substitution His506Tyr. Domain 3 is highlighted in yellow and the SNP-directed substitution His506Tyr is highlighted in red.

fidelity of transcription however the SNP does no fall within this region (Y. N. Zhou *et al.* 2013).

Of key relevance to the current study, a further gene, *braS*, displayed 100% of reads from pool 3 with the presence of a SNP. The identified G -> A base change at chromosome position 2627088 resulted in a threonine to isoleucine (T175I) substitution. Threonine and isoleucine are structurally similar amino acids however the change from threonine to isoleucine replaces a polar residue with a non-polar one. Only a single protein domain is predicted in BraS from *S. aureus 171* by searching against the Pfam database and this is an ATP binding domain encompassing the residues 195 to 294. The SNP therefore does not occur within this domain. In a previous experimental evolution experiment by Blake *et al,* the amino acid change A208E was observed in the same histidine kinase but was positioned within the ATP binding domain (Blake *et al.* 2011). As previously discussed, BraRS is involved in resistance to nisin (section 4.1.4). Mutational assays of the BraRS TCS have shown that inactivation of the genes results in an increased susceptibility to nisin (Hiron *et al.* 2011). Due to the presence of this *braS* mutation in 100% of reads in pool 3 and the presence of the mutation in *rpoC* in 100% of reads covering this locus from pool 3, it is likely that these mutations occurred together in the clones constituting pool 3.

In pool 1, 100% of reads covering a locus had a mutation at chromosome position 17119 which is within the gene *walk,* encoding the sensor protein of the essential TCS WalKR, also known as YycGF. The SNP results in a histidine to arginine (H364R) amino acid change within the PAS domain of the protein. The PAS domain is responsible for signal sensing within a sensor histidine kinase protein and is a diverse and flexible domain (Hefti *et al.* 2004). WalKR is integral to maintaining cell wall metabolism (Delaune *et al.* 2012) and mutations in the TCS are linked with resistance to vancomycin and daptomycin and an associated cell wall-thickening phenotype (Howden *et al.* 2011). The antimicrobial effects of nisin occur as a result of its interaction with cell wall intermediates including lipid I and lipid II, which results in sequestration of these intermediates and also pore formation which uses these intermediates as

docking molecules. A cell wall thickening phenotype that may result from mutation of WalKR could limit the ability of nisin to interact with these target molecules and therefore abrogate its effects.

Pool 1 was the first pool during the experimental evolution to exhibit a large increase in nisin resistance (Figure 5.4a). A large overlap between the WalKR and GraRS regulatory networks has been suggested in *S. aureus* (Falord *et al.* 2011).

A point mutation in the large subunit of glutamate synthase was observed in pool 1 and the single clone sequenced from the encompassing experiment. 100% of reads covering this locus report a base change A -> T which results in the amino acid change glycine to leucine (G797L) within a flavin mononucleotide (FMN) binding domain. This domain has a beta/alpha barrel structure and is one of the catalytic sites of glutamate synthase together with the amidotransferase domain and interacts with FAD or NADPH. The FMN binding site makes a GOGAT domain with the Fe-S cluster. This domain spans residues 806-1179 in *Corynebacterium glutamicum* (Beckers *et al.* 2001). Glutamate synthase involved in the production of D-glutamate which is a component of cell wall peptidoglycans. Of relevance, one study reported that the small subunit of glutamate synthase was upregulated in response to the presence of cell wall active antibiotics, including bacitracin (Utaida *et al.* 2003).

A further gene involved in peptidoglycan synthesis also contains a SNP in both pool 1 and its single clone. *mraY* encodes a transferase which catalyses the formation of lipid I (Bouhss *et al.* 1999), a peptidoglycan precursor with which nisin is known to interact (5.1.1). This SNP results in the amino acid change V266I, which is often considered a functionally conservative replacement, indicating that this SNP might not alter MraY activity.

Finally a C -> T base change in the *yurK* gene encoding a transcriptional regulator at the chromosomal location 1998279 results in a glycine to arginine (G73R) amino acid substitution. This transcriptional regulator is adjacent to an

ABC transporter gene, which it may regulate transcriptionally. ABC transporters are known to transport a range of substrates, including antibiotics such as nisin which is especially relevant here. Certain multidrug ABC transporters have been described with a broad substrate specificity (Poelarends *et al.* 2000).

5.3.3.2 Nisin-selected SNPs of *S. hominis*

Fewer nisin-selected SNPs were present in the *S. hominis* clone pools compared with those present in the *S. aureus* pools (Tables 5.4 and 5.5). No SNPs were present in all three pools. Only 1 SNP was present in more than one pool. An A -> C base change at chromosome position 2211829 that resulted in an asparagine to histidine (N40H) amino acid change in a hypothetical protein with no Pfam protein domains found in the amino acid sequence. Since this SNP was only found in 35.8% of reads covering this locus, it is likely that it was present in just one of the clones. The fact that it was present in more than one pool, however, means that it arose independently in two nisin selection experiments.

Several SNPs were present in 100% of reads covering their chromosomal location. A SNP in the zinc metalloprotease gene *ftsH* was present in 100% of reads covering its location in pool 2 indicating that the SNP was present in every clone. This SNP was also present in the sequenced single clone (Table 5.5a). A C -> A base change at chromosome position 2172470 resulted in an aspartate to glutamate (D171E) amino acid change in the extracellular domain of the protein. Aspartate and glutamate are both structurally similar acidic amino acids and often considered functionally interchangeable. FtsH was found to play a role in osmotic stress and is required for the integration of penicillin binding proteins into the cell membrane. Its expression was found to be upregulated in response to cell wall active antimicrobials in *S. aureus* MRSA252 (Overton *et al.* 2011).

A SNP was present in 100% of reads covering the chromosome position 552630 in pool 3 which resulted in an arginine to histidine (R135H) amino acid change in guanylate kinase, the gene product of *gmk.* This is perhaps somewhat unusual since *gmk* is included in the panel of housekeeping genes used in multi locus

Table 5.5a Non-synonymous, homozygous SNPs from nisin evolved isolates of S. hominis strain J31. Pools were comprised of 5 clones from each experimental repeat, for example pool 1 corresponds to experimental repeat 1. Nisin MICs of clones in each pool were confirmed to ensure they were not significantly different.

Prokka annotation	Blast similarity	Pools	Chromosome position	Base change	Amino acid change	% of reads
rpoB	DNA-directed RNA polymerase subunit beta	1	41317	G -> T	Asp1046Tyr	78.94736842
graS	Sensor histidine kinase	2	147503	C -> T	Ser120Leu	42.12765957
ftsH	Zinc metalloprotease	2	2172470	C -> A	Asp171Glu	100
hypothetical		2,3	2211829	A -> C	Asn40His	34.84848485
N/A		2	2323889	A -> T		60
gmk	Guanylate Kinase	3	552630	G -> A	Arg135His	100
yurK	Transcriptional regulator	3	1212777	G -> A	Gln47*	100

Table 5.5b Non-synonymous, homozygous SNPs from nisin evolved clone of S. hominis strain J31. The S. hominis J31 single clone sequences was from pool 2

Prokka annotation	Blast similarity	Chromosome position	Base change	Amino acid change
ftsH	Zinc metalloprotease	2172470	C -> A	Asp171Glu
N/A		2189828	G -> T	
N/A (plasmid encoded)		2323889	A -> T	

sequence typing due to the lack of selective pressure acting upon the genes and their resultant stable sequences (Chambers & DeLeo 2009).

A C -> T base change in the gene *graS* at chromosome position 147503 substitutes the amino acid serine with leucine (S120L). It is likely that this mutation was present in 2 clones in pool 2 due to the proportion of reads covering this location which carry the alternative base. As previously discussed *graS* encodes the sensor histidine kinase for the TCS GraRS which is involved in resistance to cationic antimicrobial peptides (section 4.1.4). *S. hominis* is part of the group of staphylococcal species which carry both *braRS* and *graRS*, therefore it is intriguing that experimental evolution with nisin resulted in SNPs in different homologous TCS in the species *S. aureus* and *S. hominis* which encode both GraRS and BraRS; a SNP was present in the sensor histidine kinase, *braS* in *S. aureus* pool 3.

A base change of A -> T at chromosome position 2323889 was identified which is 31 base pairs upstream of a replication protein on the *S. hominis* J31 plasmid which carries the gene *qacC.* This gene encodes resistance to cationic compounds by exporting them from the cell in a proton-dependent manner, and has therefore been designated as encoding a multidrug resistance export protein (Paulsen *et al.* 1995). Notably, the plasmid carrying *qacC,* pSepCH in *S. epidermidis* CH occurs alongside the plasmid pCH01 encoding lantibiotic biosynthesis. In *S. warneri* strain ISK-1 a *qacC* plamid, pPI-2, and a lantibiotic biosynthesis plasmid, pPI-1 are both present (3.3.15.1).

There were several similarities between the nisin-selected SNPs that occurred in both *S. aureus* and *S. hominis*. Firstly, an amino acid change in *rpoB* was observed in each species resulting in a tyrosine residue. However, these two tyrosine residues are separated in both the amino acid primary structure and in the folded protein. The H506Y amino acid change in *rpoB* of *S. aureus* 171 is present in the fork region of domain three, whereas in *rpoB* of *S. hominis* J31 the SNP is found at chromosomal location 41317 (D1046Y) which falls in domain six known to be one of the dispensable regions of the protein where random

insertions are found in many bacterial species. Both domain 3 where the SNP is present in *S. aureus* and domain 6 containing the SNP in *S. hominis* are highlighted in Figure 5.9. Whether, given their somewhat proximal structural locality there is any similarity in functional outcome is unclear. Secondly, a base change of G -> A at chromosome position 1212777 in the gene encoding transcriptional regulator YurK resulted in a truncation (N47*) of the protein at the amino acid residue glutamine 47. The amino acid sequence of this transcriptional regulator shares 75% sequence identity with *yurK* which was found to contain a SNP in pool 1 at chromosome position 1998279. In *S. hominis* the transcription factor is also adjacent to the same ABC transporter observed in *S. aureus*. The ABC transporters in the two species share 51.9% identity at the amino acid level. In *S. aureus* 171 the G73R substitution could have an abrogating consequence upon YurK activity to match the transcription termination effect of the SNP in *S. hominis* J31, but this would require further study.

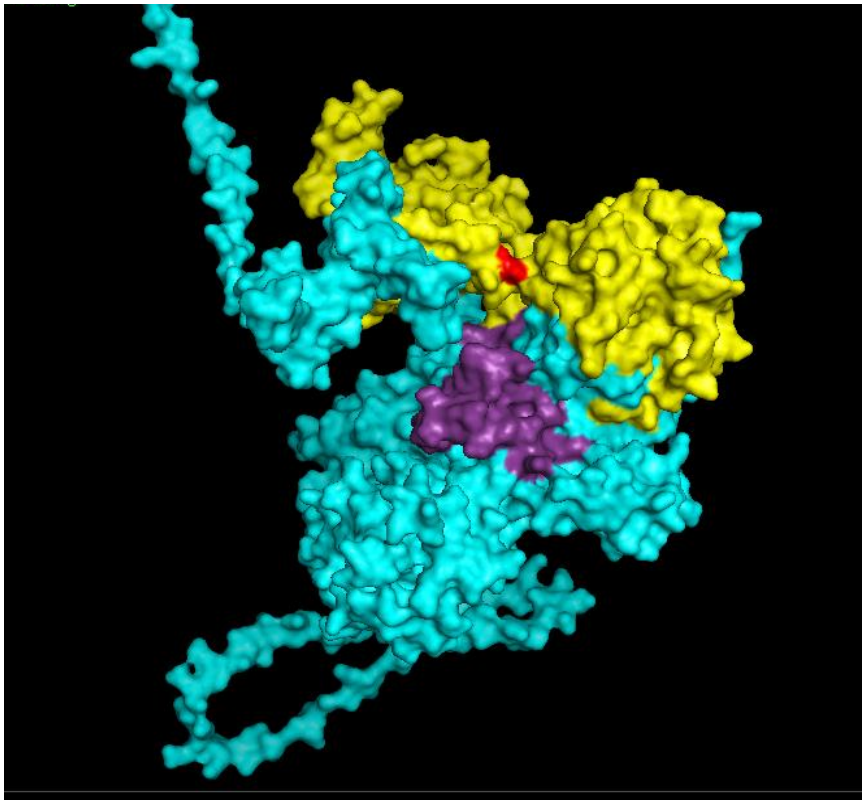


Figure 5.9 3D model of *S. hominis* RpoB with the highlighted substitution Asp1046Tyr. Domain 3 is highlighted in purple and domain 6 is highlighted in yellow. The SNP Asp1046Tyr is highlighted in red.

5.3.3.3 Nisin-selected SNPs of *S. saprophyticus*

As observed with *S. hominis,* the number of SNPs resulting from evolution with nisin selection in *S. saprophyticus* was reduced when compared to *S. aureus*. No identical SNPs were present in all three pools, however SNPs were identified in the gene *graS* in all clones of all pools. Specifically, in pools 2 and 3 a G -> T base change at chromosome position 2069134 resulted in the presence of a serine residue in the place of an alanine residue (A160S), replacing a basic amino acid for a polar one (Table 5.6a). In pool 1 a base change at chromosome position 2068987 substituted a glycine residue for a cysteine residue (G209C), replacing one polar residue with another. This is the only SNP present in the *S. saprophyticus* single clone (Table 5.6b) confirming it is important in the development of nisin resistance. Both of these residue positions are distinct from the amino acid change corresponding to the *graS* mutation in pool 2 of *S. hominis*. The *S. saprophyticus* A160S amino acid change appears to be in the same region of the GraS protein as the S120L amino acid change resulting from the *graS* SNP identified in *S. hominis* pool 2, however the protein model created of GraS from *S. hominis* and *S. saprophyticus* was a low quality model therefore statements must be made with caution. The G209C residue change in *S. saprophyticus* GraS appears to be located within a separate protein domain.

Other SNPs present in more than one pool include a C -> A base change at chromosome position 1537681 which results in an amino acid change from leucine to phenylalanine (L79F) in the transcriptional regulator CodY. This amino acid change from one non-polar amino acid to a more bulky non-polar residue occurs in GAF-like domain of the protein thought to be involved in GTP binding. The presence of a bulky amino acid in this domain may prevent GTP binding and thus attenuate the repressive effect of this regulator. In *S. aureus* CodY in known to be a global regulator which alters expression of virulence genes in response to nutrient availability, and deletion has been shown to enhance virulence in CA-MRSA (Montgomery *et al.* 2012). CodY may be involved in nisin resistance through multiple described mechanisms. For example, it may have a role in the breakdown of the peptide antibiotic through regulation of protease production, it may differentially influence the transcription of *rpoB*

Table 5.6a Non-synonymous, homozygous SNPs from nisin evolved isolates of S. saprophyticus strain 883. Pools were comprised of 5 clones from each experimental repeat, for example pool 1 corresponds to experimental repeat 1. Nisin MICs of clones in each pool were confirmed to ensure they were not significantly different.

Prokka annotation	Blast similarity	Pools	Chromosome position	Base change	Amino acid change	% of reads
codY	Transcriptional regulator	2,3	1537681	C -> A	Leu79Phe	100
pitA	Inorganic phosphate transporter	2,3	2066724	G -> A	Ala195Val	100
graS	Sensor histidine kinase	2,3	2069134	G -> T	Arg160Ser	100
hypothetical	hypothetical	2	140317	G -> T	Val73Leu	38.46153846
integrase core domain	Transposase	2	912319	A -> G	Lys160Arg	33.33333333
marR family protein	Transcriptional regulator	2,3	2136907	G -> A	Thr62Ile	100
hypothetical protein	Transposase	1	911139	T -> A	Met1?	71.42857143
graS	Sensor histidine kinase	1	2068987	C -> A	Gly209Cys	100

Table 5.6b Non-synonymous, homozygous SNPs from nisin evolved clone of S. saprophyticus strain 883. The S. saprophyticus 883 single clone sequences was from pool 1

Prokka annotation	Blast similarity	Chromosome position	Base change	Amino acid change
graS	Sensor histidine kinase	2068987	C -> A	Gly209Cys

which is implicated in nisin resistance in *S. aureus* and *S. hominis* by the results of this chapter, or it may alter gene expression of peptide transporters to export nisin from the cell (Somerville & Proctor 2009). Due to the fact that nisin activity occurs mainly at the cell surface, an increase in the transcription of extracellular proteases may be a mechanism by which resistance to nisin could occur. Furthermore, due to the presence of SNPs in *rpoB* in nisin-evolved strains of *S. aureus* and *S. hominis* it seems likely that there may be differential regulation of *rpoB*.

A protein with similarity to the PitA inorganic phosphate transporter of *E. coli* was found to contain a nisin-selected amino acid change of alanine to valine as a result of a G -> A base change at chromosome position 2066724. Alanine and valine are structurally similar and non-polar. The *S. aureus* homologues of the *pitAB* (SAOUHSC_00669 and SAOUHSC_00670) transporter operon are part of the GraRS regulatory network, and bioinformatic analysis identified a WalR binding site upstream of *S. aureus pitA* (Falord *et al.* 2011). The PitA homologue in *S. saprophyticus* 883 showed 87.5% protein sequence identity with *S. aureus* SAOUHSC_00670 indicating that this protein may be the *S. saprophyticus* homologue and therefore part of the GraRS and/or WalKR regulons.

5.4 Conclusion

In this chapter the capacity for *S. aureus*, *S. saprophyticus* and *S. hominis* to acquire resistance to the lantibiotic nisin was investigated. Specifically, the aim was to correlate nisin-selected resistance using experimental evolution with the resulting genome distribution of SNPs in the context of the *graS* and *braS* genes that are differentially present across the staphylococci. Each of the three staphylococcal species tested exhibited an increased nisin resistance with both strains of *S. hominis* developing a 25-fold increase and both strains of *S. aureus* a 100-fold increase. There was variation in the final nisin MIC between the two isolates of *S. saprophyticus* experimentally evolved with nisin. *S. saprophyticus* strain 349 acquired a 4.7-fold increase in resistance whereas *S. saprophyticus* strain 883 exhibited an 80-fold increase in resistance.

During the experimental evolution three experiments were run for each of the six strains in parallel. Not all experiments exhibited equivalent fold increases in resistance or developed them at the same rate. Furthermore, when the SNPs from the pools of clones sequenced from each experimental repeat were examined, the repertoire of SNPs was different for each of the pools. This indicated that the development of nisin resistance in each parallel experiment occurred through different pathways and involved different genes, or combinations thereof.

In the *S. aureus* nisin selection pool 3 clones contained a SNP in the *braS* gene. This gene was previously implicated in resistance to nisin (Kolar *et al.* 2011) (Blake *et al.* 2011). Despite this, pool 1 which did not contain a SNP in either *braS* or its regulator *braR* evolved resistance to nisin more quickly than pool 3. The SNPs found in pool 1 were found in *walK*, glutamate synthase and *mraY*. These genes are involved in peptidoglycan synthesis and turnover, and synthesis of lipid precursors known to be targets for the mechanism of action of nisin. This proposes an alternative strategy for nisin resistance whereby lipid precursors are synthesised at an increased rate to counteract the mechanism of nisin to sequester them.

In *S. aureus* and *S. saprophyticus* all three sequenced pools reached the same end point nisin MIC. This suggests that, although crucial SNPs in the evolution of increased nisin resistance may have occurred at different passage stages resulting in increased resistance occurring at different rates in each pool, all pathways in the different pools are ultimately effective in conferring nisin resistance in this experiment. This is not the case for *S. hominis* strain J31, however, since pool 3 is the only pool to show a 25-fold increase in nisin resistance. Therefore this pathway is most effective in conferring increased resistance to nisin in this species. The presence of a SNP in the *yurK* transcriptional regulator gene in 100% of reads covering the region in this *S. hominis* pool, and in *S. aureus* pool 1, implicates *yurK* and potentially the adjacent ABC transporter as being important in nisin resistance in both species. Intriguingly, a MarR family transcriptional regulator was implicated in *S. saprophyticus* nisin resistance in pools 2 and 3 which could be functionally equivalent to YurK.

Results from all three species highlight the complex interplay between TCSs as SNPs were present in genes from the regulons of GraRS, BraRS and WalKR as well as within the genes encoding the TCSs themselves. The regulons of GraRS and BraRS are known to overlap, as are those of GraRS and WalKR. Furthermore, Hiron *et al* showed that deletion mutants of GraRS and BraRS resulted in increased susceptibility to nisin in *S. aureus* (Hiron *et al.* 2011). The fact that SNPs were present in *walK* and WalKR regulon genes in all three species supports the possibility that this TCS and its regulatory network may also be involved in nisin resistance in addition to GraRS and BraRS. Due to the potentially interconnected nature of the three TCSs, although SNPs were found in different genes in each of the species, it may be the case that many of the SNPs occurred within the same regulatory network of genes to produce somewhat equivalent resistance outcomes.

Chapter 6

Conclusions and future work

The aim of this thesis was the comparative study of the major human staphylococcal species to gain insights into their complement of available genes that make them successful at thriving on skin. To complete this aim, firstly sequencing of the genomes of those species not publically available was undertaken. Secondly, in order to enlarge the current knowledge of *S. hominis*, the second most common species colonising the human skin, and therefore allow comparison with the major skin coloniser *S. epidermidis,* as well as an exploration of the gene content of *S. hominis* as a species, a set of recent skin isolates of the latter species was sequenced. Thirdly, differential presence of genomic components in the major staphylococcal species was then investigated to elucidate new insights into human skin colonisation, with focus on two related TCS.

6.1 Genome sequencing and annotation

Over the course of this project, sequencing technologies have advanced such that the volume of genomic data has increased greatly. This has led to both benefits and drawbacks in scientific research.

The volume of data has led to the development of the RefSeq database, which aims to reduce the redundancy of available information, and integrate data types such as transcriptomic, proteomic and annotation (Pruitt *et al.* 2012). Furthermore, sequencing of large strain sets has enabled pan-genome analyses which take into account the diversity and genetic complexity of a single species. For the staphylococci, large scale sequencing efforts and pan-genome analyses have been completed for *S. aureus* (Kelly 2013), (Méric *et al.* 2015) *S. epidermidis* (Conlan *et al.* 2012), (Kelly 2013) (Méric *et al.* 2015) and *S. haemolyticus* (Cavanagh *et al.* 2014). Until now no such analysis had been carried out for *S. hominis.* The *S. hominis* pan-genome is discussed below (section 6.2).

The long read sequencing technology PacBio has enabled the sequencing and assembly of the *S. hominis* isolate J31 into five contigs. One of these contigs is proposed to encode the whole bacterial chromosome, and the other four are

proposed to encode plasmids. Discrepancies between the annotation of the Illumina-sequenced *S. hominis* J31 and the PacBio assembly indicate the need for localised Sanger sequencing of these regions of the genome to resolve coding sequence prediction (section 3.3.11). One possible reason for these discrepancies is the reliance of this study upon computational methods of annotation.

The process of computational functional predictions, and the reduced incidence of experimental verification of these functional predictions, has led to mis-annotation of protein sequences and propagation of these errors in protein sequence databases. This project relies largely on such computational function prediction and protein sequence databases. The presence of a TCS similar to that of BraRS in *S. aureus*, for example, was predicted by computational annotation in a group of species (section 4.3.3.1), and its identity explored and confirmed through bioinformatics analysis. Experimental confirmation of its function in species other than *S. aureus*, however was not performed except with the use of experimental evolution.

6.2 The *S. hominis* pan-genome

The main weakness of this pan-genome study is the restricted *S. hominis* strain set included in the analysis; it also highlights the advances in genomic study over recent years The largest published pan-genome sequencing efforts of *S. aureus, S. epidermidis* and *S. haemolyticus* involve 168, 139 and 134 genomes respectively. The pangenome analysis of *S. hominis* includes only 11 strains which falls far short of these numbers. Additionally, the Liverpool strains were sampled from a single geographic area, albeit the study volunteers migrated from different parts of the UK, and they worked in the same building and environment. This could be argued to have limited the diversity of the strains and therefore skewed results associated with the pan-genome, such as increasing the apparent number of strain-specific genes or reducing the apparent size of the pan-genome. The pan-genome of *S. hominis* was approaching closure at 11 genomes, meaning that few new genes were added to the pan-genome upon addition of each strain. With increased strain diversity it

is highly likely that the size of the *S. hominis* pan-genome would increase from 3115 genes.

The analysis of the *S. hominis* pan genome revealed the conserved presence of several genes proposed as virulence factors in *S. hominis*: *shsA*; *braRS* and *stp1/stk1*. The putative *S. hominis* cell wall-associated protein ShsA shares 16% protein sequence identity with *S. aureus* AdsA, previously known as SasH. ShsA is also predicted to share 5'-nucleosidase activity with AdsA, this enzyme mediates adenosine synthesis activity in *S. aureus*, which in turn reduces its killing by neutrophils. This function designates AdsA as a virulence factor of *S. aureus*. The two component system BraRS is associated with virulence phenotypes in *S. aureus,* such as biofilm formation and resistance to macrophage killing (Kolar 2012), despite its presence in the non-pathogenic species *S. hominis* among others. Finally, the serine/threonine kinase Stp1/Stk1 is also associated with virulence in a mouse model due to reduced survival in the kidney of an *stk1* mutant (Débarbouillé *et al.* 2009).

Virulence, by definition, is the degree of pathogenicity of an organism, and in turn the definition of pathogenicity is the ability to cause disease. If genes labelled as virulence factors, or as being able to cause virulence phenotypes, are also present in non-pathogenic backgrounds such as those discussed above it could be argued that they play a role as a survival factor. Genes with roles in evading the host immune system, or antibiotic resistance, or adhesion and colonisation may have homologues in non-pathogenic staphylococci and may have primary functions as survival factors. Although these survival factors may help *S. aureus* to cause disease in the sense that they allow the bacteria to survive and propagate sufficiently to seed the tissues where disease is caused, they may not be directly involved in the disease symptoms, themselves. Proteins such as SasG and its *S. epidermidis* homologue Aap, and Ebh and its *S. epidermidis* homologue Embp are often discussed in the context of virulence in *S. aureus*, however they all play roles during different stages of colonisation by interacting with the host (Roche, Meehan, *et al.* 2003; Christner *et al.* 2010). It is perhaps unsurprising, therefore, that such surface proteins which interact with

host molecules should be recruited into the virulence pathway in *S. aureus* (Roche, Massey, *et al.* 2003) (Cheng *et al.* 2014). Understanding the homologues of these virulence-associated proteins as survival factors in non-pathogenic species may enhance the understanding of their evolution and recruitment into pathogenesis of *S. aureus.*

6.3 Clade groups of the *Staphylococcus* genus and phylogenetic relationships can be explored by orthologous genes

A previous comparative genomic analysis of the genus *Staphylococcus* was carried out by Suzuki *et al.* In their study the phylogenetic relationships of the staphylococci were inferred from the presence of a group of 491 shared orthologues among 29 strains of 12 species of *Staphylococcus.* The phylogenetic relationships revealed *S. pseudintermedius* as the most basal lineage, with *S. carnosus* as the next most basal lineage. After these two species, the 10 other species included in their analysis were often isolated from human skin, therefore the study inferred from this that the evolution of human adaptation occurred after the branching of these 10 other species from *S. carnosus* (Suzuki *et al.* 2012).

The same study also investigated whether the relationships of the 29 staphylococcal strains revealed by the 491 shared orthologues were maintained when total gene content was investigated. In contrast to this study, and other studies of staphylococcal phylogeny (Takahashi *et al.* 1999; Lamers *et al.* 2012) the method employed by Suzuki *et al* revealed that the *S. aureus* strains clustered in one clade and the rest of the *Staphylococcus* species into another, and that *S. simiae* clustered with *S. carnosus*. Clustering of *S. hominis* and *S. haemolyticus* was not observed and *S. saprophyticus* clustered among the *S. epidermidis* group of species (Suzuki *et al.* 2012).

The study presented here carried out a threefold analysis of: the total gene content of *Staphylococcus* species using orthologous clusters of genes; the groups of staphylococci that cluster together in the analysis of these orthologous clusters; and in turn the genes that drive the formation of these

clusters. In conjunction with the RandomForest algorithm this analysis revealed the possible genetic basis of the lifestyles of the groups of staphylococcal species. Groups of staphylococcal species have previously been proposed by multiple studies. Kloos *et al* originally proposed that staphylococcal species were divided into multiple species groups on the basis of DNA-DNA hybridisation methods (Kloos & Schleifer, 1991). These species groups, represented by *S. epidermidis, S. saprophyticus, S. simulans, S. intermedius, S. hyicus, S. sciuri, S. aureus and S. caseolyticus* were also identified by Takahashi *et al* on the basis of 16S rDNA sequence analysis, with the addition of an *S. auricularis* species group (Takahashi *et al.* 1999). *S. caseolyticus* has since been designated as a micrococcus and renamed *Micrococcus caseolyticus,* therefore only eight of the nine species groups represent *Staphylococcus* species. Lamers *et al* proposed a refinement of the *S. epidermidis* and *S. aureus* groups into one group, and the *S. hyicus* and *S. intermedius* groups also into one group (Lamers *et al.* 2012). In the analysis here of orthologous groups of genes across the publically available genomes of staphylococci these species groups were also broadly conserved. The clustering of species of the *S. sciuri* group, the *S. simulans* group, the *S. intermedius* group, the *S. saprophyticus* group, the *S. epidermidis* group, and the *S. aureus* included in the OrthoMCL analysis were conserved in this thesis (Figure 4.1, section 4.3.2). Of note, the *S. epidermidis* species group was the least well resolved in the 16S rDNA analysis and the comprising species were described as being "somewhat" closely related to one another (Takahashi *et al.* 1999). This species group is broken into smaller clusters which are reflected in the OrthoMCL analysis: *S. lugdunensis* clusters alone; *S. hominis* and *S. haemolyticus* cluster together; *S. epidermidis* and *S. capitis* cluster together. Furthermore, as Lamers *et al* propose the *S. epidermidis-aureus* species group, (Lamers *et al.* 2012) here group A was separated into the *S. epidermidis* group and the *S. aureus* group. The genetic basis of the two groups showed a high degree of overlap in the randomForest analysis.

Although the phenotypic basis of these groups was explored by Kloos *et al* (Kersters & Vancanneyt 2005) the genetic basis driving them has not previously been explored. As the grouping and clustering patterns observed by Kloos &

Schleifer and Takahashi *et al* are maintained on the basis of orthologous gene content, a subsequent analysis which more thoroughly covers the genus *Staphylococcus,* coupled with use of algorithms, such as a randomForest analysis should be used to investigate the genetic drivers behind each of the groups. This interrogative strategy would help to explore further the lifestyles associated with the eight species groups. Additionally, the analysis by Takahashi *et al* included only 38 species of staphylococci, and the OrthoMCL analysis in this study included only 20 species (Takahashi *et al.* 1999). Therefore representation of the remaining species from the 51 *Staphylococcus* species that currently constitute the genus may result in additional species groups and clusters.

6.4 The staphylococcal response to antimicrobial peptides

BraRS has a role in resistance to nisin and bacitracin in *S. aureus*, however it is not known if this role also extends to the other species that encode BraRS. SNPs were present in *braS* of *S. aureus* evolved in the presence of nisin, however this was not observed with the *braRS*-positive species *S. hominis.* Hiron *et al* showed that inactivation of both *braRS* and its ortholog *graRS* produced a similar increase in sensitivity to nisin, however the same gene mutants showed differential sensitivity to bacitracin, with the *braRS* mutant having a greater sensitivity to the latter than the *graRS* mutant (Hiron *et al.* 2011). Experimental evolution of both *S. aureus* and *S. hominis* in the presence of bacitracin and investigation of the resulting SNPs may reveal differential loci under selection in the presence of both antibiotics. Furthermore, inactivation of *S. hominis braRS* and *graRS,* and comparative analysis of the MICs of both mutants with bacitracin and nisin would reveal if the two TCSs respond comparatively in *S. aureus* and *S. hominis.* This strategy could be repeated for the transcription factor YurK and its adjacent ABC transporter in both *S. hominis* and *S. aureus* to assess their involvement in nisin and bacitracin resistance.

Additional complexity of inter-relationships between regulons of genes is revealed by the genes under selection pressure during experimental evolution in the presence of nisin. The regulons of the TCSs GraRS and BraRS, as well as

GraRS and WalKR are known to overlap (Falord *et al.* 2011), and SNPs are present in these three TCSs and their regulons in evolved clones and clone pools. In addition to their overlapping regulons, it is possible that the TCSs BraRS and GraRS themselves are regulated differentially. The TCS Stk1/Stp1 is involved in the regulation of GraRS through phosphorylation of three conserved residues in the DNA binding domain of GraR. The regulons of Stk1/Stp1 and GraRS also overlap, for example transcriptomic analysis showed differential expression of the GraRS regulated genes VraFG and the *agr* operon in wild type and a Stk/Stp1 knockout mutant (Donat *et al.* 2009). Phosphorylation by Stk1 is vital to the regulation of D-alanylation of cell wall teichoic acids and substitution of the phosphorylation sites results in the same phenotype as a *graR* inactivation mutant. Additionally, despite conservation of two of the threonine residues, BceR of the BceRS TCS in *Bacillus subtilis* is not phosphorylated by Stk1 (Fridman *et al.* 2013). BceRS is homologous to the BraRS TCS, therefore it may be possible that BraRS and GraRS are differentially regulated and therefore activate an overlapping regulon of genes through different mechanisms to ensure a breadth of responses to the diversity of antimicrobial peptides and antibiotics encountered by the staphylococci. This kinase/phosphatase is conserved in both the BraRS positive species from this analysis *S. hominis* J31and *S. aureus* 171 as well as the BraRS-negative species *S. saprophyticus* 883 with an average of 63.7% protein sequence identity. Knockout mutants of Stk1/Stp1 in both *S. hominis* and *S. saprophyticus* would show if the regulation of GraRS by this signal transduction system is conserved across the staphylococci. Furthermore, the study by Fridman *et al* did not investigate phosphorylation of BraR by Stk1 in *S. aureus* (Fridman *et al.* 2013), investigation of this is necessary to explore if BraRS and GraRS are both phosphorylated by Stk1 and therefore under its regulation, thus adding an additional aspect of cross talk between TCSs.

In conclusion, this study set out to identify the genetic toolbox at the disposal of major skin staphylococci, including a detailed investigation of the genetic complement of *S. hominis*. Although this was achieved through in-depth interrogation of the *S. hominis* genome and investigation of the shared genes of

skin colonising species of staphylococci, no inferences about how and where these genes are used can be made. In addition to the genomic data from skin-colonising bacteria, greater detail is required about the specific skin locations and micro-niches they inhabit is required. Understanding of the fine architecture and topography of the skin surface could perhaps be obtained from microscopy of skin explants. Skin sampling techniques able to distinguish between the upper levels of the skin surface, and the creases and reliefs would give deeper insight into the structure of bacterial communities. Such analysis might enable propositions as to the likely environmental drivers of speciation.

Remodelling of bacterial communities has been proposed as a method for restoring health where there is disease and to limit effects upon personal hygiene (Manichanh *et al.* 2010). In order to remodel a bacterial community it is essential to know:

1. What species of bacteria constitute this community
2. How these species interact with one another
3. How these species interact with the host

Until the ecology of the skin, and how it relates to the fine topography of the niche, is better understood through more sophisticated sampling techniques, remodelling of the skin microbiome will remain a challenge.

Supplementary information

Table S1 Description of the coding sequences of the *S. hominis* complete phage CDS highlighted in bold represent BLAST hits to PHASTs non-redundant protein database, all other CDS represent BLAST hits to PHAST's phage/prophage protein database.

CDS position	BLAST hit	*S. hominis* protein ID
B10 intact phage 1		
83326..83338	***attL* TTTATATCAAATC**	
complement(83848..85122)	PHAGE_Bacill_G_NC_023719: gp344	B10-PB_00084
complement(85067..86452)	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: integrase	B10-PB_00085
86531..86543	***attL* TATGATATAATTT**	
86575..87183	PHAGE_Staphy_phiSauS_IPLA35_NC_011612: hypothetical protein SauSIPLA35_gp02	B10-PB_00086
complement(87180..87353)	**hypothetical protein**	**B10-PB_00087**
complement(87417..87764)	**hypothetical protein**	**B10-PB_00088**
complement(87818..88429)	PHAGE_Staphy_phi2958PVL_NC_011344: probable transcriptional repressor	B10-PB_00089
88602..88829	PHAGE_Staphy_phi2958PVL_NC_011344: hypothetical protein phi2958PVL_gp07	B10-PB_00090
88857..89654	PHAGE_Staphy_phiRS7_NC_022914: pathogenicity island protein	B10-PB_00091
89658..90503	**hypothetical protein**	**B10-PB_00092**
90742..90987	PHAGE_Staphy_YMC/09/04/R1988_NC_022758: hypothetical protein	B10-PB_00093
complement(90956..91321)	PHAGE_Staphy_phi7401PVL_NC_020199: hypothetical protein	B10-PB_00094
91374..91505	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	B10-PB_00095
91480..91665	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	B10-PB_00096
91744..92031	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	B10-PB_00097
92003..92230	PHAGE_Staphy_PH15_NC_008723: hypothetical protein ph45	B10-PB_00098
92220..92885	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: erf superfamily protein	B10-PB_00099
92899..93312	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: single stranded DNA binding protein	B10-PB_00100
93327..93998	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	B10-PB_00101
93995..94732	PHAGE_Staphy_phiRS7_NC_022914: DnaD and phage-associated domain protein	B10-PB_00102
94729..95088	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	B10-PB_00103
95081..96316	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: DNA helicase	B10-PB_00104
96313..96537	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	B10-PB_00105
96512..96754	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	B10-PB_00106
96765..97193	PHAGE_Staphy_P954_NC_013195: DNA methyltransferase	B10-PB_00107
97190..97606	PHAGE_Staphy_CNPH82_NC_008722: hypothetical protein cn50	B10-PB_00108

97593..97967	PHAGE_Staphy_vB_SepS_SEP9_NC_023582: hypothetical protein	B10-PB_00109
97967..98326	**hypothetical protein**	**B10-PB_00110**
98326..98499	**hypothetical protein**	**B10-PB_00111**
98504..98707	**hypothetical protein**	**B10-PB_00112**
98704..99156	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	B10-PB_00113
99159..99875	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: endonuclease	B10-PB_00114
99872..100057	PHAGE_Staphy_CNPH82_NC_008722: hypothetical protein cn56	B10-PB_00115
100054..100473	PHAGE_Staphy_37_NC_007055: ORF027	B10-PB_00116
100474..100818	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: nuclease	B10-PB_00117
100821..101129	**hypothetical protein**	**B10-PB_00118**
101130..101480	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: DNA binding protein	B10-PB_00119
101529..101807	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	B10-PB_00120
101800..101982	PHAGE_Staphy_StauST398_1_NC_021326: integrase regulator	B10-PB_00121
101982..102404	**hypothetical protein**	**B10-PB_00122**
102409..102633	**hypothetical protein**	**B10-PB_00123**
102637..102825	PHAGE_Staphy_2638A_NC_007051: ORF083	B10-PB_00124
102879..103166	**hypothetical protein**	**B10-PB_00125**
103193..103414	PHAGE_Bacill_phi105_NC_004167: immunity repressor	B10-PB_00126
103428..103886	PHAGE_Staphy_2638A_NC_007051: ORF020	B10-PB_00127
103987..104712	PHAGE_Lister_B025_NC_009812: gp65	B10-PB_00128
104856..105239	PHAGE_Staphy_phiBU01_NC_026016: N/A	B10-PB_00129
105226..106893	PHAGE_Lister_LP_101_NC_024387: terminase large subunit	B10-PB_00130
106913..108091	PHAGE_Lister_LP_101_NC_024387: portal protein	B10-PB_00131
108054..108767	PHAGE_Staphy_2638A_NC_007051: ORF013	B10-PB_00132
108779..109987	PHAGE_Lister_LP_101_NC_024387: major capsid protein	B10-PB_00133
110005..110292	PHAGE_Lister_LP_101_NC_024387: hypothetical protein LP101_007	B10-PB_00134
110276..110644	PHAGE_Staphy_StB20_NC_019915: phage head-tail joining protein	B10-PB_00135
110634..111023	PHAGE_Staphy_phiSa119_NC_025460: N/A	B10-PB_00136
111023..111439	PHAGE_Staphy_phiBU01_NC_026016: N/A	B10-PB_00137
111452..112075	PHAGE_Staphy_StB20_NC_019915: major tail protein	B10-PB_00138
112096..112275	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: rho termination protein	B10-PB_00139
112343..112714	PHAGE_Staphy_StB20_NC_019915: hypothetical protein	B10-PB_00140
112934..118216	PHAGE_Staphy_phi13_NC_004617: tail length tape measure protein	B10-PB_00141
118220..119065	PHAGE_Staphy_StauST398_2_NC_021323: phage tail tape measure protein-like protein	B10-PB_00142
119078..120844	PHAGE_Staphy_phiSauS_IPLA88_NC_011614: putative tail fiber protein	B10-PB_00143
120896..122866	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: virion associated hydrolase	B10-PB_00144
122871..123143	**hypothetical protein**	**B10-PB_00145**
123200..123691	PHAGE_Staphy_StB27_NC_019914: holin	B10-PB_00146
123741..124040	PHAGE_Entero_BC_611_NC_018086: Holin	B10-PB_00147

124040..125749	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: amidase	B10-PB_00148
128713..128725	***attR*** **TATGATATAATTT**	
131512..131524	***attR*** **TTTATATCAAATC**	
C80 intact phage 3		
512225..512236	*attL* TTTGTTTAATTT	0
516695..517276	PHAGE_Uganda_virus_NC_014791: polyprotein	C80-PB_00520
517276..517797	**hypothetical protein**	**C80-PB_00521**
518020..518094	**tRNA**	
complement(518225..518608)	PHAGE_Staphy_phiBU01_NC_026016: N/A	C80-PB_00523
complement(518568..519128)	PHAGE_Staphy_phiBU01_NC_026016: N/A	C80-PB_00524
complement(519304..519975)	PHAGE_Staphy_CNPH82_NC_008722: hypothetical protein cn33	C80-PB_00525
complement(520025..520576)	PHAGE_Lactob_prophage_Lj965_NC_005355: putative superinfection immunity protein	C80-PB_00526
complement(520827..521819)	PHAGE_Bacill_SPBc2_NC_001884: putative DNase/RNase endonuclease	C80-PB_00527
536746..536757	*attR* TTTGTTTAATTT	
I4 Intact phage 4		
2201825..2201849	*attL* TTATTCCAATTGCTTTATTGACGTT	
complement(2203767..2204735)	PHAGE_Strept_pyogenes_315_1_NC_004584: hypothetical protein SpyM3_0691	I4-PB_02212
2205205..2205897	PHAGE_Choris_'L'_NC_021248: N1R/p28-like protein	I4-PB_02213
complement(2206302..2207495)	PHAGE_Amsact_moorei_entomopoxvirus__L__NC_002520: hypothetical protein AMV156	I4-PB_02214
complement(2207480..2210383)	PHAGE_Mythim_'L'_NC_021246: unknown similar to AMEV156	I4-PB_02215
complement(2210391..2211293)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein	I4-PB_02216
complement(2211361..2211531)	**Lantibiotic Pep5 precursor**	**I4-PB_02217**
complement(2211589..2211777)	**hypothetical protein**	**I4-PB_02218**
complement(2211852..2212595)	PHAGE_Tricho_ni_ascovirus_2c_NC_008518: hypothetical protein TNAV2c_gp071	I4-PB_02219
2213687..2214382	PHAGE_Anomal_entomopoxvirus_NC_023426: putative ATP-binding cassette transporter	I4-PB_02220
2214384..2214593	**ABC-2 family transporter protein**	**I4-PB_02221**
2214554..2215084	**ABC-type transport system involved in multi-copper enzyme maturation, permease component**	**I4-PB_02222**
complement(2215657..2215785)	Putative transposon Tn552 DNA-invertase bin3	I4-PB_02223
complement(2215796..2216077)	PHAGE_Entero_phi80_NC_021190: N/A	I4-PB_02224
2216718..2219357	PHAGE_Megavi_lba_NC_020232: hypothetical protein	I4-PB_02225
2219381..2220325	**Glycine betaine-binding protein precursor**	**I4-PB_02226**
2220674..2221243	**putative dihydroxyacetone kinase regulator**	**I4-PB_02227**
complement(2222329..2222745)	**Glyoxalase-like domain protein**	**I4-PB_02228**
complement(2222852..2223352)	**MepB protein**	**I4-PB_02229**

complement(2223479..2223622)	**hypothetical protein**	**I4-PB_02230**
complement(2223723..2224259)	PHAGE_Bacill_Waukesha92_NC_025424: N/A	I4-PB_02231
2224616..2224807	**hypothetical protein**	**I4-PB_02232**
complement(2225167..2225889)	**ABC-2 family transporter protein**	**I4-PB_02233**
complement(2225905..2226540)	PHAGE_Bacill_G_NC_023719: gp245	I4-PB_02234
2226688..2226954	**hypothetical protein**	**I4-PB_02235**
complement(2227212..2227490)	**hypothetical protein**	**I4-PB_02236**
complement(2227497..2228234)	**hypothetical protein**	**I4-PB_02237**
complement(2228231..2228983)	**hypothetical protein**	**I4-PB_02238**
complement(2228984..2229892)	PHAGE_Anomal_entomopoxvirus_NC_023426: putative ATP-binding cassette transporter	I4-PB_02239
2230368..2230691	PHAGE_Acinet_Acj61_NC_014661: putative quaternary ammonium compound-resistance protein qacE	I4-PB_02240
complement(2231287..2231472)	**hypothetical protein**	**I4-PB_02241**
complement(2231469..2232314)	**Replication protein**	**I4-PB_02242**
2234074..2234910	**Divergent AAA domain protein**	**I4-PB_02243**
2234918..2234942	*attR* TTATTCCAATTGCTTTATTGACGTT	
2235017..2235493	**Firmicute plasmid replication protein (RepL)**	**I4-PB_02244**
2236765..2237292	PHAGE_Entero_phiX174_sensu_lato_NC_001422: major spike protein	I4-PB_02245
2237301..2238287	PHAGE_Entero_phiX174_sensu_lato_NC_001422: minor spike protein	I4-PB_02246
2238351..2239892	PHAGE_Entero_phiX174_sensu_lato_NC_001422: DNA replication initiation protein gpA	I4-PB_02247
2240006..2240149	PHAGE_Entero_phiX174_sensu_lato_NC_001422: C	I4-PB_02248
2240146..2240604	PHAGE_Entero_phiX174_sensu_lato_NC_001422: external scaffolding protein	I4-PB_02249
2240604..2240720	PHAGE_Entero_phiX174_sensu_lato_NC_001422: DNA packaging protein	I4-PB_02250
2240757..2242040	PHAGE_Entero_phiX174_sensu_lato_NC_001422: capsid protein	I4-PB_02251
ZBW5 intact phage 1		
502486..503067	PHAGE_Cassav_brown_streak_virus_NC_012698: polyprotein	ZBW5-PB_00512
503067..503579	**hypothetical protein**	**ZBW5-PB_00513**
503803..503877	**tRNA**	
503862..503882	***attL*** **ATCCCTCCCAGGACGCTAATA**	
complement(504008..505033)	PHAGE_Staphy_JS01_NC_021773: integrase	ZBW5-PB_00515
complement(505091..505732)	PHAGE_Deep_s_D6E_NC_019544: structure protein	ZBW5-PB_00516
complement(505832..506293)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_00517
complement(506306..506632)	PHAGE_Lactob_Sha1_NC_019489: bifunctional S24 family peptidase/transcriptional regulator	ZBW5-PB_00518
506853..507050	PHAGE_Lactoc_bIL286_NC_002667: repressor	ZBW5-PB_00519

507095..507697	**hypothetical protein**	**ZBW5-PB_00520**
507660..507974	PHAGE_Staphy_69_NC_007048: ORF045	ZBW5-PB_00521
507979..508167	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_00522
508230..508997	**hypothetical protein**	**ZBW5-PB_00523**
508999..509280	**hypothetical protein**	**ZBW5-PB_00524**
509240..509464	PHAGE_Staphy_StB20_NC_019915: hypothetical protein	ZBW5-PB_00525
509457..510119	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: erf superfamily protein	ZBW5-PB_00526
510130..510552	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: single stranded DNA binding protein	ZBW5-PB_00527
510736..511404	PHAGE_Staphy_X2_NC_007065: ORF017	ZBW5-PB_00528
511404..512102	PHAGE_Staphy_CNPH82_NC_008722: putative DNA replication protein	ZBW5-PB_00529
512108..512464	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: hypothetical protein	ZBW5-PB_00530
512451..513698	PHAGE_Staphy_CNPH82_NC_008722: putative DNA helicase	ZBW5-PB_00531
513698..513919	PHAGE_Staphy_StB20_NC_019915: hypothetical protein	ZBW5-PB_00532
513897..514142	PHAGE_Staphy_StB20_NC_019915: hypothetical protein	ZBW5-PB_00533
514153..514581	PHAGE_Staphy_P954_NC_013195: DNA methyltransferase	ZBW5-PB_00534
514578..514997	PHAGE_Staphy_CNPH82_NC_008722: hypothetical protein cn50	ZBW5-PB_00535
514984..515358	PHAGE_Staphy_6ec_NC_024355: hypothetical protein PHAGE6E_115	ZBW5-PB_00536
515358..515531	**hypothetical protein**	**ZBW5-PB_00537**
515536..515739	**hypothetical protein**	**ZBW5-PB_00538**
515736..516200	PHAGE_Staphy_EW_NC_007056: ORF046	ZBW5-PB_00539
516364..516384	***attR*** **ATCCCTCCCAGGACGCTAATA**	
complement(516461..517216)	PHAGE_Staphy_2638A_NC_007051: ORF015	ZBW5-PB_00540
ZBW5 intact phage 5		
2253755..2253766	***attL*** **ATTATCTTTATC**	
2258306..2260987	PHAGE_Megavi_lba_NC_020232: hypothetical protein	ZBW5-PB_02267
2261011..2261955	**Glycine betaine-binding protein precursor**	**ZBW5-PB_02268**
2262305..2262598	**putative dihydroxyacetone kinase regulator**	**ZBW5-PB_02269**
2262652..2262864	**hypothetical protein**	**ZBW5-PB_02270**
2263280..2264452	PROPHAGE_Escher_CFT073: transposase	ZBW5-PB_02271
2264883..2266091	**Putative surface protein precursor**	**ZBW5-PB_02272**
2266045..2266725	PHAGE_Staphy_CNPH82_NC_008722: conserved phage protein	ZBW5-PB_02273
2266722..2266907	PHAGE_Staphy_CNPH82_NC_008722: hypothetical protein cn56	ZBW5-PB_02274
2266904..2267251	PHAGE_Staphy_CNPH82_NC_008722: putative nuclease protein	ZBW5-PB_02275
2267254..2267562	PHAGE_Staphy_S25_3_NC_022920: hypothetical protein	ZBW5-PB_02276
2267549..2267740	**hypothetical protein**	**ZBW5-PB_02277**
2267741..2268259	PHAGE_Staphy_StauST398_3_NC_021332: hypothetical protein	ZBW5-PB_02278
2268318..2268737	**hypothetical protein**	**ZBW5-PB_02279**
2268753..2269034	PHAGE_Staphy_PH15_NC_008723: hypothetical protein ph64	ZBW5-PB_02280

2269024..2269200	**Transcriptional activator RinB**	**ZBW5-PB_02281**
2269200..2269622	**hypothetical protein**	**ZBW5-PB_02282**
2269627..2269851	**hypothetical protein**	**ZBW5-PB_02283**
2269855..2270043	PHAGE_Staphy_2638A_NC_007051: ORF083	ZBW5-PB_02284
2270057..2270455	PHAGE_Staphy_phiRS7_NC_022914: hypothetical protein	ZBW5-PB_02285
2270594..2270920	PHAGE_Staphy_2638A_NC_007051: ORF026	ZBW5-PB_02286
2271074..2271433	PHAGE_Staphy_2638A_NC_007051: ORF031	ZBW5-PB_02287
2271430..2273091	PHAGE_Staphy_2638A_NC_007051: ORF005	ZBW5-PB_02288
2273111..2274256	PHAGE_Staphy_P954_NC_013195: HK97 family phage portal protein	ZBW5-PB_02289
2274249..2275112	PHAGE_Staphy_2638A_NC_007051: ORF013	ZBW5-PB_02290
2275174..2276352	PHAGE_Lister_LP_101_NC_024387: major capsid protein	ZBW5-PB_02291
2276370..2276666	PHAGE_Staphy_2638A_NC_007051: ORF036	ZBW5-PB_02292
2276656..2277027	PHAGE_Staphy_phiBU01_NC_026016: N/A	ZBW5-PB_02293
2277017..2277439	PHAGE_Staphy_2638A_NC_007051: ORF023	ZBW5-PB_02294
2277436..2277843	PHAGE_Staphy_2638A_NC_007051: ORF024	ZBW5-PB_02295
2277844..2278458	PHAGE_Staphy_P954_NC_013195: phage tail superfamily	ZBW5-PB_02296
2278472..2278639	PHAGE_Staphy_2638A_NC_007051: ORF096	ZBW5-PB_02297
2278703..2279065	PHAGE_Staphy_2638A_NC_007051: ORF028	ZBW5-PB_02298
2279329..2285580	PHAGE_Staphy_phiRS7_NC_022914: tail length tape measure protein	ZBW5-PB_02299
2285581..2287071	PHAGE_Staphy_phiBU01_NC_026016: N/A	ZBW5-PB_02300
2287087..2292195	PHAGE_Staphy_phi5967PVL_NC_019921: phage minor structural protein	ZBW5-PB_02301
2292173..2292328	PHAGE_Staphy_phiSa119_NC_025460: N/A	ZBW5-PB_02302
2292372..2292671	PHAGE_Staphy_StauST398_4_NC_023499: hypothetical protein	ZBW5-PB_02303
2292682..2292978	PHAGE_Staphy_vB_SepiS_phiIPLA7_NC_018284: hypothetical protein	ZBW5-PB_02304
2293403..2294260	**hypothetical protein**	**ZBW5-PB_02305**
complement(2294401..2294805)	PHAGE_Staphy_phiETA_NC_003288: hypothetical protein phiETA_02	ZBW5-PB_02306
complement(2294808..2295209)	PHAGE_Staphy_42E_NC_007052: ORF031	ZBW5-PB_02307
complement(2295407..2297143)	PHAGE_Staphy_StB27_NC_019914: lysine	ZBW5-PB_02308
complement(2297156..2297443)	PHAGE_Staphy_StB27_NC_019914: holin	ZBW5-PB_02309
complement(2297616..2299055)	**Bifunctional AAC/APH**	**ZBW5-PB_02310**
complement(2299056..2299460)	PHAGE_Erwini_phiEaH2_NC_019929: N-acetyltransferase GCN5	ZBW5-PB_02311
complement(2299597..2299740)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_02312
complement(2299733..2300077)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_02313
complement(2300091..2301137)	PHAGE_Staphy_StB27_NC_019914: tail protein	ZBW5-PB_02314
complement(2301190..2303127)	PHAGE_Staphy_StB27_NC_019914: peptidoglycan hydrolase	ZBW5-PB_02315
complement(2303177..2303668)	PHAGE_Staphy_StB27_NC_019914: holin	ZBW5-PB_02316

complement(2303690..2304886)	PHAGE_Staphy_StB27_NC_019914: tail protein	ZBW5-PB_02317
complement(2304899..2306773)	PHAGE_Staphy_StB27_NC_019914: Zn carboxypeptidase	ZBW5-PB_02318
complement(2306784..2308034)	PHAGE_Staphy_StB27_NC_019914: endopeptidase	ZBW5-PB_02319
complement(2308044..2309000)	PHAGE_Staphy_StB27_NC_019914: tail protein	ZBW5-PB_02320
complement(2309013..2312405)	PHAGE_Staphy_StB27_NC_019914: tail tape measure protein	ZBW5-PB_02321
complement(2312405..2312770)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_02322
complement(2312800..2313171)	PHAGE_Staphy_StB27_NC_019914: tail assembly protein	ZBW5-PB_02323
complement(2313240..2313833)	PHAGE_Staphy_StB27_NC_019914: major tail protein	ZBW5-PB_02324
complement(2313848..2314243)	PHAGE_Staphy_StB27_NC_019914: head structural protein	ZBW5-PB_02325
complement(2314254..2314604)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_02326
complement(2314594..2314908)	PHAGE_Staphy_StB27_NC_019914: minor head protein	ZBW5-PB_02327
complement(2314905..2315240)	PHAGE_Staphy_StB27_NC_019914: DNA packaging minor head protein	ZBW5-PB_02328
complement(2315268..2316179)	PHAGE_Staphy_StB27_NC_019914: major head protein	ZBW5-PB_02329
complement(2316195..2316785)	PHAGE_Staphy_StB27_NC_019914: scaffold protein	ZBW5-PB_02330
complement(2316894..2317043)	**hypothetical protein**	**ZBW5-PB_02331**
complement(2317036..2318073)	PHAGE_Staphy_StB27_NC_019914: minor head protein	ZBW5-PB_02332
complement(2318057..2319508)	PHAGE_Staphy_StB27_NC_019914: portal protein	ZBW5-PB_02333
complement(2319509..2320795)	PHAGE_Staphy_StB27_NC_019914: TerL	ZBW5-PB_02334
complement(2320788..2321240)	PHAGE_Staphy_StB27_NC_019914: TerS	ZBW5-PB_02335
complement(2321438..2321854)	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: RinA	ZBW5-PB_02336
complement(2321928..2322374)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_02337
2322138..2322149	***attR*** **ATTATCTTTATC**	
complement(2322371..2322538)	PHAGE_Staphy_StB27_NC_019914: RinB	ZBW5-PB_02338
complement(2322531..2322809)	PHAGE_Staphy_StB27_NC_019914: hypothetical protein	ZBW5-PB_02339
complement(2322823..2322987)	**hypothetical protein**	**ZBW5-PB_02340**
complement(2323007..2323321)	**hypothetical protein**	**ZBW5-PB_02341**
complement(2323410..2323961)	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: dUTPase	ZBW5-PB_02342
complement(2323962..2324138)	**hypothetical protein**	**ZBW5-PB_02343**
complement(2324131..2324418)	**hypothetical protein**	**ZBW5-PB_02344**

complement(2324422..2324766)	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: nuclease	ZBW5-PB_02345
complement(2324767..2325186)	PHAGE_Staphy_37_NC_007055: ORF027	ZBW5-PB_02346
complement(2325183..2325368)	PHAGE_Staphy_vB_SepiS_phiIPLA5_NC_018281: hypothetical protein	ZBW5-PB_02347
complement(2325365..2326045)	PHAGE_Staphy_CNPH82_NC_008722: conserved phage protein	ZBW5-PB_02348
complement(2326113..2326517)	**Replication protein**	**ZBW5-PB_02349**
complement(2326762..2328024)	PHAGE_Pieris_rapae_granulovirus_NC_013797: desmoplakin	ZBW5-PB_02350

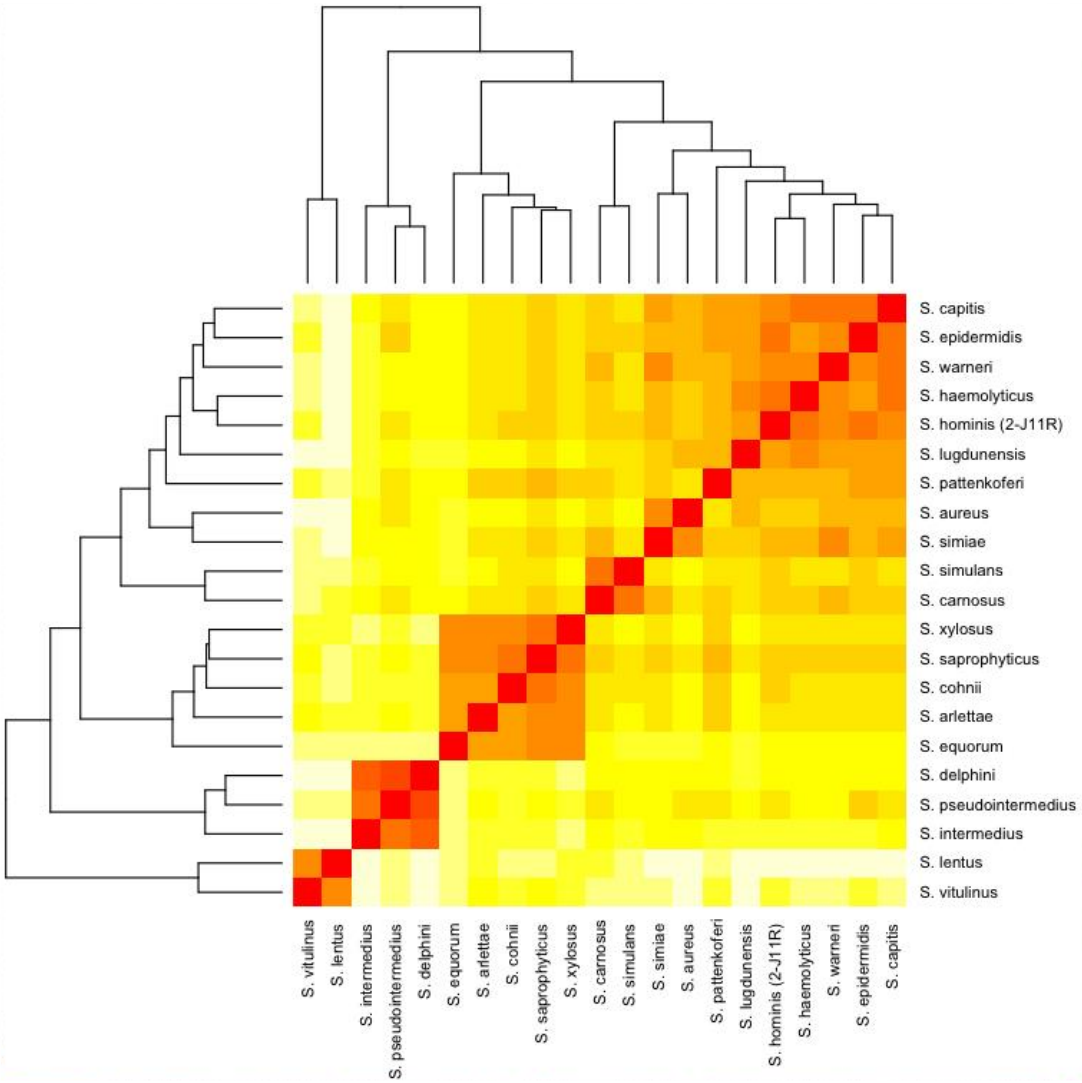


Figure S.1 Heat map representation of the presence of clustered orthologous proteins across the staphylococci. Swapping the *S. hominis* strains J6 for J11 revealed no difference in the groups of staphylococci discussed in Chapter 4. A high number of shared clusters of orthologous proteins is indicated by red colouration, few shared clusters is indicated by white.

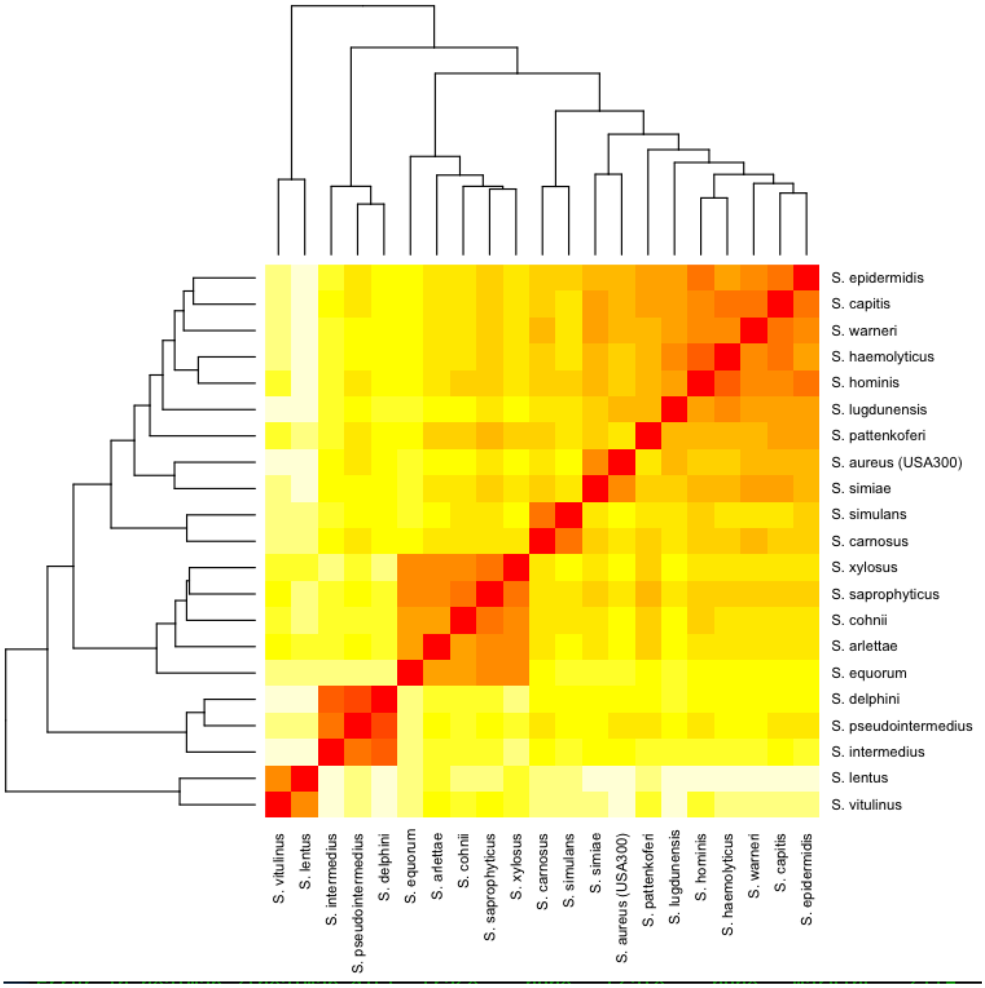


Figure S.2 Heat map representation of the presence of clustered orthologous proteins across the staphylococci. Swapping the *S. aureus* strain newman for USA300 revealed no difference in the groups of staphylococci discussed in Chapter 4. A high number of shared clusters of orthologous proteins is indicated by red colouration, few shared clusters is indicated by white.

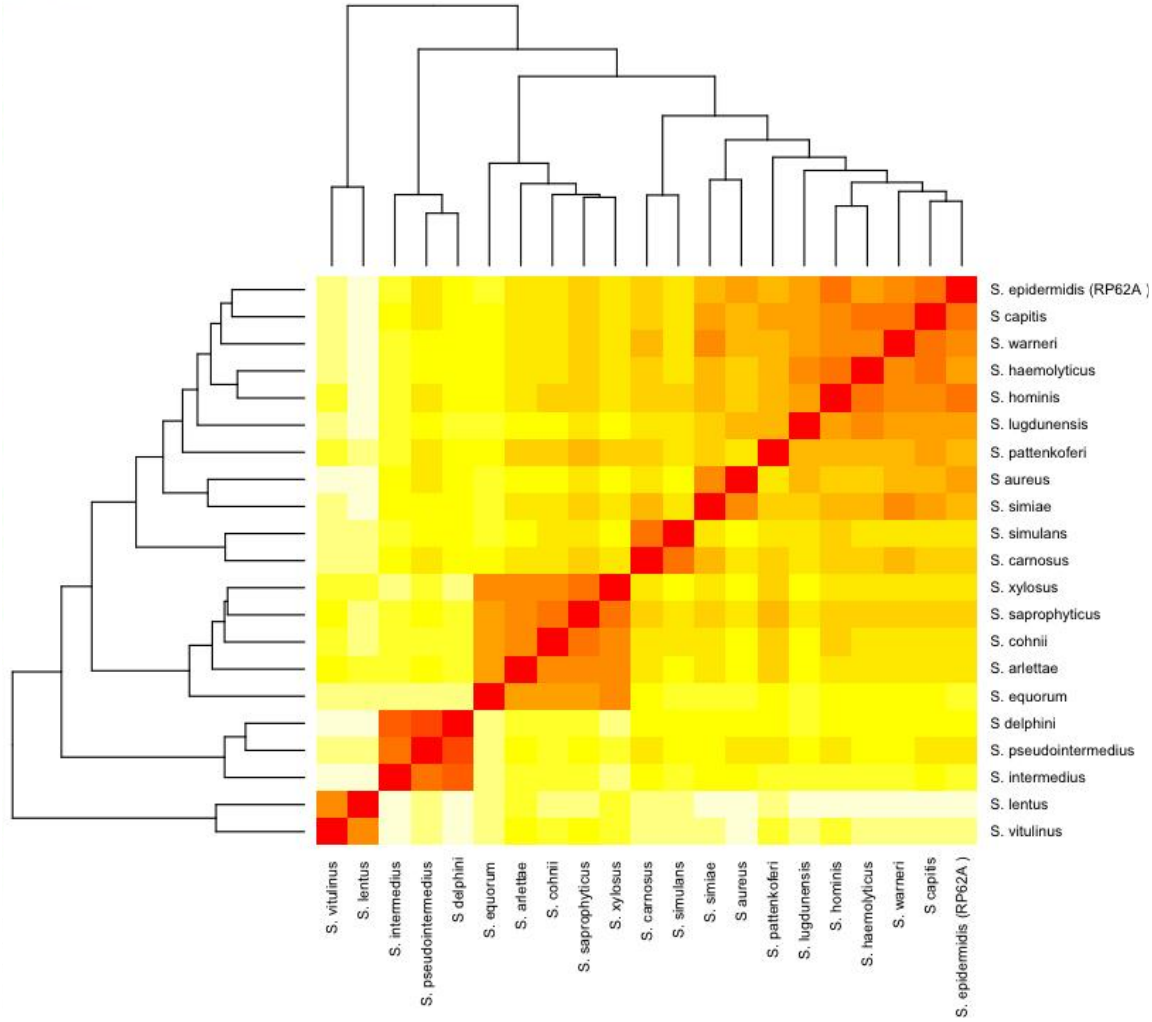


Figure S.3 Heat map representation of the presence of clustered orthologous proteins across the staphylococci. Swapping the *S. epidermidis* strain ATCC12228 for RP62A revealed no difference in the groups of staphylococci discussed in Chapter 4. A high number of shared clusters of orthologous proteins is indicated by red colouration, few shared clusters is indicated by white.

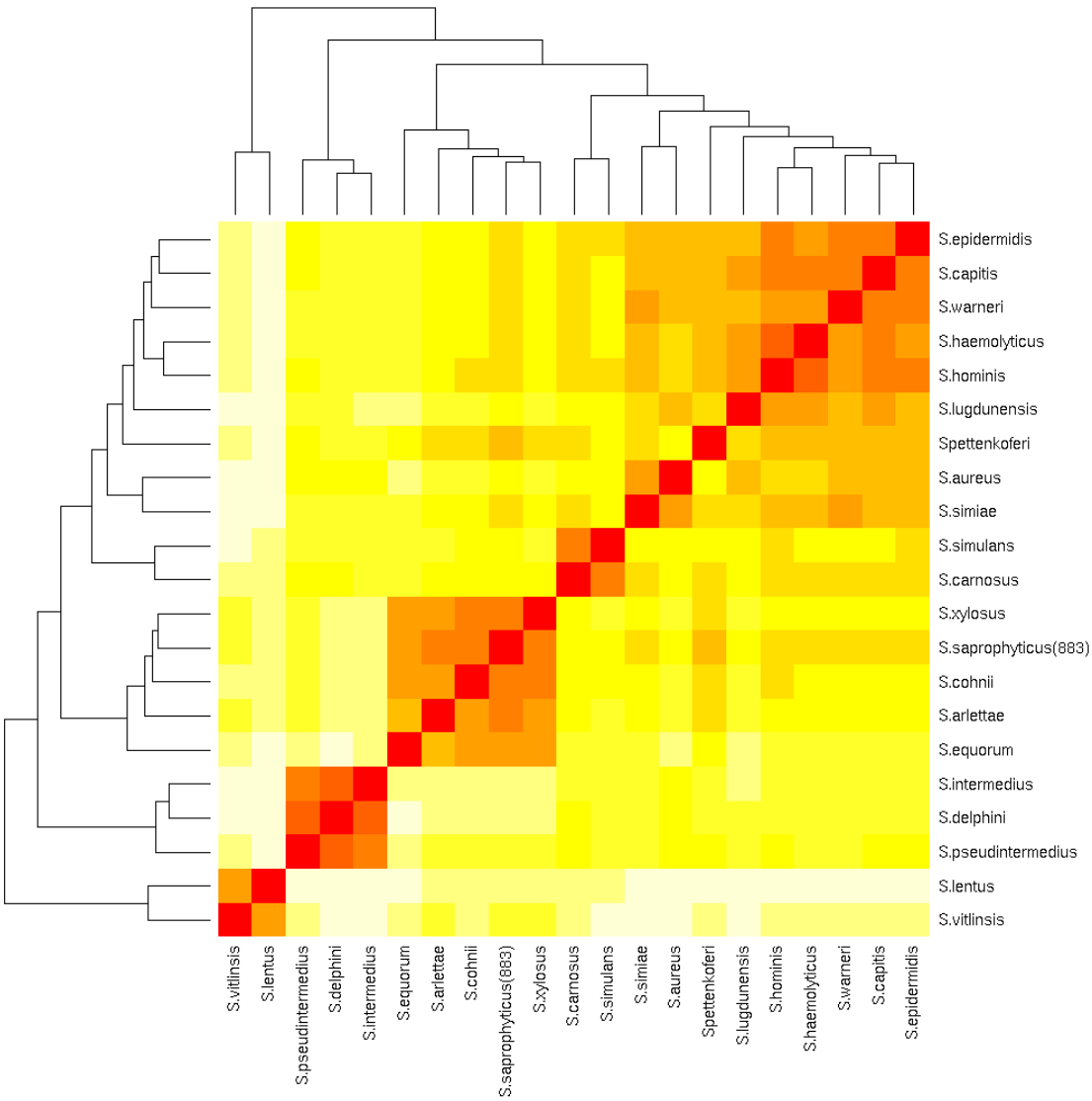


Figure S.4 Heat map representation of the presence of clustered orthologous proteins across the staphylococci. Swapping the *S. saprophyticus* strain strain ATCC 15305 for CCM883 revealed no difference in the groups of staphylococci discussed in Chapter 4. A high number of shared clusters of orthologous proteins is indicated by red colouration, few shared clusters is indicated by white.

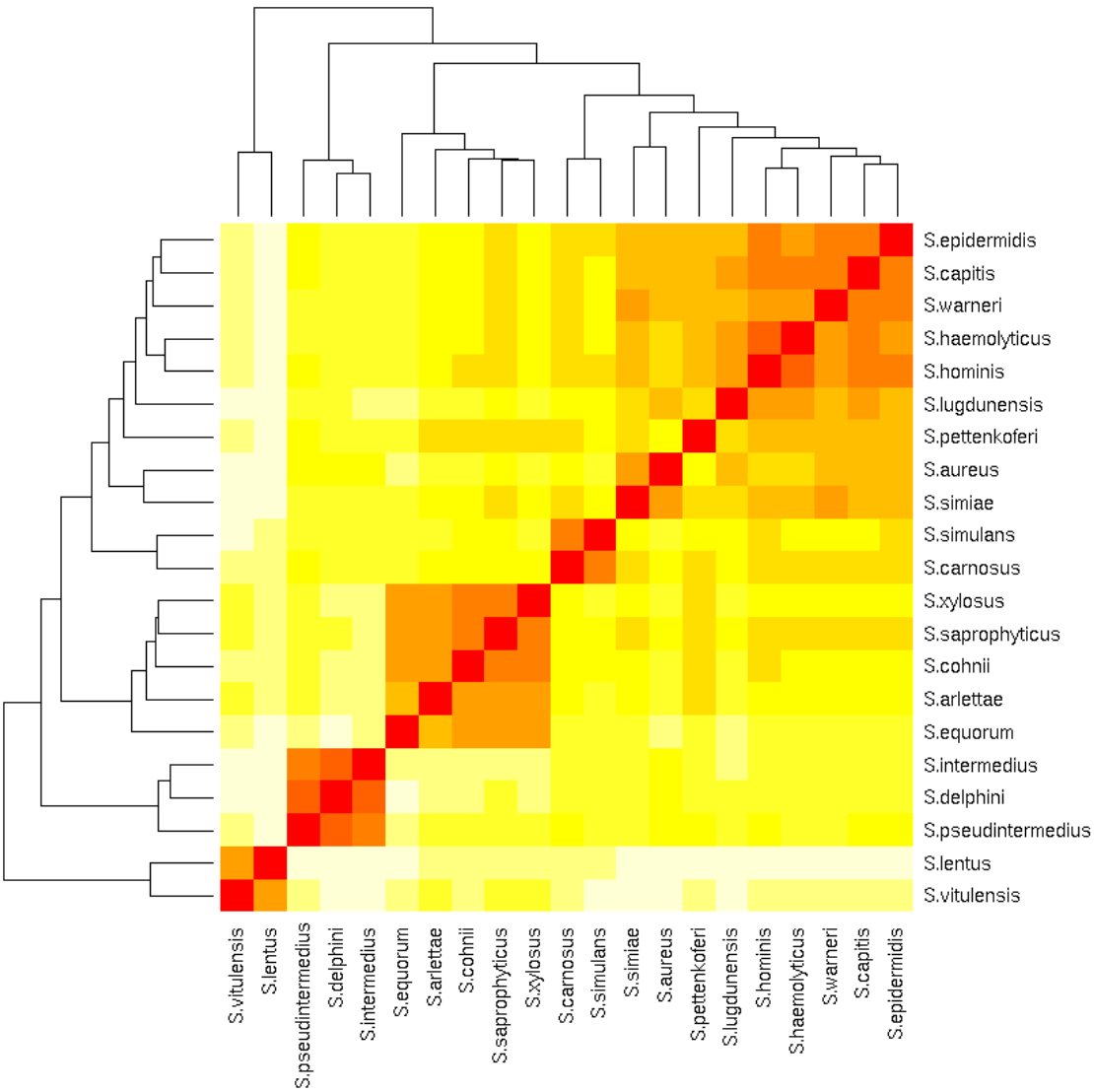


Figure S.5 Heat map representation of the presence of clustered orthologous proteins across the staphylococci. Swapping the *S. pseudintermedius* strain strain NCTC_11048 for ED99 revealed no difference in the groups of staphylococci discussed in Chapter 4. A high number of shared clusters of orthologous proteins is indicated by red colouration, few shared clusters is indicated by white.

Table S.2 A description, usage instructions and input files required for all bespoke scripts used in this thesis. Scripts are provided on the memory stick accompanying this thesis

Thesis section	Author		Description	Usage
3.2.1, 3.2.7, 4.2.2	Thomas Craig	*Table_alt_cluster_singleton.py*	Converts the clusters from an OrthoMCL output file into a table of the presence (1) and absence (0) of these clusters in each genome used in the analysis	*Python Table_alt_cluster_singleton.py concatenated_proteins.faa orthomcl_output.out* **Input files:** • Concatenated fasta file of all proteins used in the OrthoMCL analysis • OrthoMCL output files **Output files** • Presence (1)/Absence (0) table
3.2.1	Ben Wareham	*core_pan_plot*	In conjunction with wrapper, once loaded into R the script calculates the core and pan-genome sizes for x number of strain combinations increasing from 1 to x	*Wrapper(data,x) > table.csv* **Input files** • Tab-delimited output from the script *Table_alt_cluster_singleton.py* **Output files** • Comma sepatated file with 3 columns for the combinations of strains up to x. Col1=Strains IDs Col2=Size of core genome Col3=Size of pan genome
3.2.1	Ben Wareham	*New_genes_added*	In conjunction with wrapper, once loaded into R the script calculates the new genes added for x number of strain combinations increasing from 1 to x	*Wrapper(data,x) > table.csv* **Input files** • Tab-delimited output from the script *Table_alt_cluster_singleton.py* **Output files**

				• Comma sepatated file with 2 columns for the combinations of strains up to x. Col1=Strains IDs Col2=Number of new genes added
3.2.2	Jennifer Kelly	*extract_COGID*	Adds COG annotation to OrthoMCL clusters if all annotations in a cluster are the same	*Perl extract_COGID.pl orthomcl_output.out CODID_list.txt > COG_annotated_clusters.txt* **Input files** • OrthoMCL output file • COGID list: tab delimited file of all proteins used in the OrthoMCL analysis and their COGID annotated by WebMGA (http://weizhong-lab.ucsd.edu/metagenomic-analysis/) **Output files** •
3.2.4	Rosanna Coates-Brown	*LPXTG_capture.pl*	Searches the input genome for the extended LPXTG motig and writes the corresponding protein to a multi fasta file	*Perl LPXTG_capture.pl query_genome.fasta > output.fasta* **Input files** • Target genome in fasta format **Output files** • Multi fasta file of LPXTG protein sequences
5.2.4	Rosanna Coates-Brown	*SNP_analysis.sh*	Aligns reads from experimentally evolved strains to wild type reference strains using BWA, converts resulting SAM files to BAM files, uses these BAM files for SNP calling with bcftools and quality filters these SNPs	*bash SNP_analysis.sh <reference_genome.fasta> <read_1.fastq> <read_2.fastq> <output_prefix>* **Input files** • The reference genome the reads will be aligned to • The forward and reverse sequencing reads in separate fastq

				files • A prefix for the output files **Output files** • Prefix_R[1/2].sai file containing coordiates of aligned reads • Sam file of reads mapped to the reference genome-prefix_mappedfile.sam • Bam file of above • Sorted bam file of above • Sam file of the headers from sorted bam file above • Bam file of above with duplicate mapped reads removed • Bcf file of above • Vcf file of SNPs called using bcf tools • Quality filtered vcf file produced using vcfutils.pl
5.2.4	Rosanna Coates-Brown	*Significant_SNPs.sh*	Calls the scripts heterozygote_allele_percentages_basic.pl and snp_final_parse.pl Filters the output from *SNP_analysis.sh* to find the percentage of alternative base calls and include only alternative base calls in >33.33% of reads	*Bash Significant_SNPs.sh <mpileup_nobaseinfo.bcf> <qualfiltered.vcf> <prefix_for_output>* **Input files** • Output file from *SNP_analysis.sh* **Output files** • Text file of filtered snps

References

Aakra, A. et al., 2007. Survey of genomic diversity among Enterococcus faecalis strains by microarray-based comparative genomic hybridization. *Applied and Environmental Microbiology*, 73(7), pp.2207–2217.

Abtin, A. et al., 2009. Degradation by Stratum Corneum Proteases PreventsEndogenous RNase Inhibitor from BlockingAntimicrobial Activities of RNase 5 and RNase 7. 129(9), pp.2193–2201.

Alkhatib, Z. et al., 2012. Lantibiotics: how do producers become self-protected? *Journal of Biotechnology*, 159(3), pp.145–154.

Alkhatib, Z. et al., 2014. The C-terminus of nisin is important for the ABC transporter NisFEG to confer immunity in Lactococcus lactis. *MicrobiologyOpen*, 3(5), pp.752–763.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Anon, 2008. Description. pp.1–4.

Arciola, C.R. & Baldassarri, L., 2001. Presence of icaA and icaDgenes and slime production in a collection of staphylococcal strains from catheter-associated infections. *Journal of clinical ….*

Arêde, P., Ministro, J. & Oliveira, D.C., 2013. Redefining the role of the β-lactamase locus in methicillin-resistant Staphylococcus aureus: β-lactamase regulators disrupt the MecI-mediated strong repression on mecA and optimize the phenotypic expression of resistance in strains with constitutive mecA expression. *Antimicrobial Agents and Chemotherapy*, 57(7), pp.3037–3045.

Asaduzzaman, S.M. et al., 2006. Lysine-oriented charges trigger the membrane binding and activity of nukacin ISK-1. *Applied and Environmental Microbiology*, 72(9), pp.6012–6017.

Aso, Y. et al., 2005. Description of complete DNA sequence of two plasmids from the nukacin ISK-1 producer, Staphylococcus warneri ISK-1. *Plasmid*, 53(2), pp.164–178.

Augustin, J. et al., 1992. Genetic analysis of epidermin biosynthetic genes and epidermin-negative mutants of Staphylococcus epidermidis. *European journal of biochemistry / FEBS*, 204(3), pp.1149–1154.

Avdeeva, S.V. et al., 2006. Human Angiogenin Lacks Specific Antimicrobial Activity. *Current Microbiology*, 53(6), pp.477–478.

Baar, C. et al., 2003. Complete genome sequence and analysis of Wolinella succinogenes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), pp.11690–11695.

Bacon, D.J. et al., 2000. Involvement of a plasmid in virulence of Campylobacter jejuni 81-176. *Infection and Immunity*, 68(8), pp.4384–4390.

Bankevich, A. et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5), pp.455–477.

Bannerman, T.L., Wadiak, D.L. & Kloos, W.E., 1991. Susceptibility of Staphylococcus species and subspecies to teicoplanin. *Antimicrobial Agents and Chemotherapy*, 35(9), pp.1919–1922.

Barba, M., Czosnek, H. & Hadidi, A., 2014. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6(1), pp.106–136.

Barbier, F. et al., 2011. High prevalence of the arginine catabolic mobile element in carriage isolates of methicillin-resistant Staphylococcus epidermidis. *The Journal of antimicrobial chemotherapy*, 66(1), pp.29–36.

Barros, E.M. et al., 2012. Staphylococcus haemolyticus as an important hospital pathogen and carrier of methicillin resistance genes. *Journal of Clinical Microbiology*, 50(1), pp.166–168.

Bauer, R. & Dicks, L.M.T., 2005. Mode of action of lipid II-targeting lantibiotics. *International Journal of Food Microbiology*, 101(2), pp.201–216.

Baviera, G. et al., 2014. Microbiota in healthy skin and in atopic eczema. *BioMed research international*, 2014, p.436921.

Beard, S.J. et al., 2000. Evidence for the transport of zinc(II) ions via the pit inorganic phosphate transport system in Escherichia coli. *FEMS Microbiology Letters*, 184(2), pp.231–235.

Beckers, G., Nolden, L. & Burkovski, A., 2001. Glutamate synthase of Corynebacterium glutamicum is not essential for glutamate synthesis and is regulated by the nitrogen status. *Microbiology (Reading, England)*, 147(Pt 11), pp.2961–2970.

Benjamini, Y. & Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10), pp.e72–e72.

Bera, A. et al., 2006. Influence of Wall Teichoic Acid on Lysozyme Resistance in Staphylococcus aureus. *Journal of Bacteriology*, 189(1), pp.280–283.

Bera, A. et al., 2004. Why are pathogenic staphylococci so lysozyme resistant? The peptidoglycan O-acetyltransferase OatA is the major determinant for lysozyme resistance of Staphylococcus aureus. *Molecular Microbiology*, 55(3), pp.778–787.

Berlin, K. et al., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6), pp.623–630.

Bierbaum, G. et al., 1996. The biosynthesis of the lantibiotics epidermin, gallidermin, Pep5 and epilancin K7. *Antonie van Leeuwenhoek*, 69(2), pp.119–127.

Bissett, D.L. & Anderson, R.L., Lactose and d-Galactose Metabolism in Group N Streptococci: Presence of Enzymes for Both the d-Galactose 1-Phosphate and d-Tagatose 6-Phosphate Pathways1.

Blake, K.L., Randall, C.P. & O'Neill, A.J., 2011. In vitro studies indicate a high resistance potential for the lantibiotic nisin in Staphylococcus aureus and define a genetic basis for nisin resistance. *Antimicrobial Agents and Chemotherapy*, 55(5), pp.2362–2368.

Blanpain, C. & Fuchs, E., 2009. Epidermal homeostasis: a balancing act of stem cells in the skin. *Nature Reviews Molecular Cell Biology*, 10(3), pp.207–217.

Boetzer, M. & Pirovano, W., 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics*, 15, p.211.

Boetzer, M. et al., 2011. Scaffolding pre-assembled contigs using SSPACE. *Journal of Gerontology*, 27(4), pp.578–579.

Boles, B.R. & Horswill, A.R., 2008. Agr-mediated dispersal of Staphylococcus aureus biofilms. *PLoS Pathogens*, 4(4), pp.e1000052–e1000052.

Bollenbach, T. et al., 2009. Nonoptimal Microbial Response to Antibiotics Underlies Suppressive Drug Interactions. *Cell*, 139(4), pp.707–718.

Bore, E. et al., 2007. Acid-shock responses in Staphylococcus aureus investigated by global gene expression analysis. *Microbiology (Reading, England)*, 153(Pt 7), pp.2289–2303.

Borges-Walmsley, M.I. & Walmsley, A.R., 2001. The structure and function of drug pumps. *Trends in Microbiology*, 9(2), pp.71–79.

Bouchami, O. et al., 2011. Molecular Epidemiology of Methicillin-Resistant Staphylococcus hominis (MRSHo): Low Clonality and Reservoirs of SCCmec Structural Elements V. Chaturvedi, ed. *PLoS ONE*, 6(7), p.e21940.

Bouhss, A. et al., 1999. Topological analysis of the MraY protein catalysing the first membrane step of peptidoglycan synthesis. *Molecular Microbiology*, 34(3), pp.576–585.

Bowden, M.G. et al., 2005. Identification and preliminary characterization of cell-wall-anchored proteins of Staphylococcus epidermidis. *Microbiology (Reading, England)*, 151(Pt 5), pp.1453–1464.

Bowden, M.G. et al., 2002. Is the GehD lipase from Staphylococcus epidermidis a collagen binding adhesin? *Journal of Biological Chemistry*, 277(45), pp.43017–43023.

Breiman, L., 2001. Random forests. *Machine learning*.

Breukink, E. et al., 1999. Use of the cell wall precursor lipid II by a pore-forming peptide antibiotic. *Science*, 286(5448), pp.2361–2364.

Brötz, H. et al., 1998. Role of lipid-bound peptidoglycan precursors in the formation of pores by nisin, epidermin and other lantibiotics. *Molecular Microbiology*, 30(2), pp.317–327.

Bubeck Wardenburg, J. et al., 2007. Poring over pores: alpha-hemolysin and Panton-Valentine leukocidin in Staphylococcus aureus pneumonia. *Nature Medicine*, 13(12), pp.1405–1406.

Bubeck Wardenburg, J., Williams, W.A. & Missiakas, D., 2006. Host defenses against Staphylococcus aureus infection require recognition of bacterial lipoproteins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(37), pp.13831–13836.

Buell, C.R. et al., 2003. The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18), pp.10181–10186.

Burgess-Herbert, S.L. & Euling, S.Y., 2013. Use of comparative genomics approaches to characterize interspecies differences in response to environmental chemicals: challenges, opportunities, and research needs. *Toxicology and applied pharmacology*, 271(3), pp.372–385.

Burian, M. et al., 2010. Temporal expression of adhesion factors and activity of global regulators during establishment of Staphylococcus aureus nasal colonization. *The Journal of Infectious Diseases*, 201(9), pp.1414–1421.

Burts, M.L., DeDent, A.C. & Missiakas, D.M., 2008. EsaC substrate for the ESAT-6 secretion pathway and its role in persistent infections of Staphylococcus aureus. *Molecular Microbiology*, 69(3), pp.736–746.

Cain, B.D. et al., 1993. Amplification of the bacA gene confers bacitracin resistance to Escherichia coli. *Journal of Bacteriology*, 175(12), pp.3784–3789.

Calcutt, M.J. et al., 2013. Draft Genome Sequence of Staphylococcus simulans UMC-CNS-990, Isolated from a Case of Chronic Bovine Mastitis. *Genome announcements*, 1(6).

Canchaya, C. et al., 2003. Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, 6(4), pp.417–424.

Candi, E., Schmidt, R. & Melino, G., 2005. The cornified envelope: a model of cell death in the skin. *Nature Reviews Molecular Cell Biology*, 6(4), pp.328–340.

Capone, K.A. et al., 2011. Diversity of the human skin microbiome early in life. *Journal of Investigative Dermatology*, 131(10), pp.2026–2032.

Carlsson, F. et al., 2006. Signal sequence directs localized secretion of bacterial surface proteins. *Nature*, 442(7105), pp.943–946.

Cartron, M.L. et al., 2014. Bactericidal activity of the human skin fatty acid cis-6-hexadecanoic acid on Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 58(7), pp.3599–3609.

Carver, T. et al., 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Journal of Gerontology*, 28(4), pp.464–469.

Carver, T.J. et al., 2005. ACT: the Artemis Comparison Tool. *Bioinformatics*, 21(16), pp.3422–3423.

Cavaco, L.M. et al., 2010. Cloning and occurrence of czrC, a gene conferring cadmium and zinc resistance in methicillin-resistant Staphylococcus aureus CC398 isolates. *Antimicrobial Agents and Chemotherapy*, 54(9), pp.3605–3608.

Cavanagh, J.P. et al., 2014. Whole-genome sequencing reveals clonal expansion of multiresistant Staphylococcus haemolyticus in European hospitals. *Journal of Antimicrobial Chemotherapy*, 69(11), pp.2920–2927.

Ceotto, H. et al., 2010. Nukacin 3299, a lantibiotic produced by Staphylococcus simulans 3299 identical to nukacin ISK-1. *Veterinary microbiology*, 146(1-2), pp.124–131.

Chalker, A.F. et al., 2000. The bacA gene, which determines bacitracin susceptibility in Streptococcus pneumoniae and Staphylococcus aureus, is also required for virulence. *Microbiology (Reading, England)*, 146 (Pt 7), pp.1547–1553.

Chambers, H.F. & DeLeo, F.R., 2009. Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nature Reviews Microbiology*, 7(9), pp.629–641.

Chambert, R., Pereira, Y. & Petit-Glatron, M.-F., 2003. Purification and characterization of YfkN, a trifunctional nucleotide phosphoesterase secreted by Bacillus subtilis. *Journal of biochemistry*, 134(5), pp.655–660.

Chatterjee, C. et al., 2005. Biosynthesis and mode of action of lantibiotics. *Chemical reviews*, 105(2), pp.633–684.

Chaves, F. et al., 2005. Nosocomial Spread of a Staphylococcus hominis subsp. novobiosepticus Strain Causing Sepsis in a Neonatal Intensive Care Unit. *Journal of Clinical Microbiology*, 43(9), pp.4877–4879.

Chen, Y.-C. et al., 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE*, 8(4), pp.e62856–e62856.

Cheng, A.G. et al., 2011. A play in four acts: Staphylococcus aureus abscess formation. *Trends in Microbiology*, 19(5), pp.225–232.

Cheng, A.G. et al., 2009. Genetic requirements for Staphylococcus aureus abscess formation and persistence in host tissues. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 23(10), pp.3393–3404.

Cheng, A.G., Missiakas, D. & Schneewind, O., 2014. The giant protein Ebh is a determinant of Staphylococcus aureus cell size and complement resistance. *Journal of Bacteriology*, 196(5), pp.971–981.

Chesneau, O. et al., 2005. Molecular analysis of resistance to streptogramin A compounds conferred by the Vga proteins of staphylococci. *Antimicrobial Agents and Chemotherapy*, 49(3), pp.973–980.

Chin, C.-S. et al., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10(6), pp.563–569.

Chiu, L.S. et al., 2010. Staphylococcus aureus carriage in the anterior nares of close contacts of patients with atopic dermatitis. *Archives of dermatology*, 146(7), pp.748–752.

Chlebowicz, M.A. et al., 2010. Recombination between ccrC genes in a type V (5C2&5) staphylococcal cassette chromosome mec (SCCmec) of Staphylococcus aureus ST398 leads to conversion from methicillin resistance to methicillin susceptibility in vivo. *Antimicrobial Agents and Chemotherapy*, 54(2), pp.783–791.

Christensen, G.D. et al., 1983. Characterization of clinically significant strains of coagulase-negative staphylococci. *Journal of Clinical Microbiology*, 18(2), pp.258–269.

Christiansen, I. & Hengstenberg, W., 1999. Staphylococcal phosphoenolpyruvate-dependent phosphotransferase system--two highly similar glucose permeases in Staphylococcus carnosus with different glucoside specificity: protein engineering in vivo? *Microbiology (Reading, England)*, 145 (Pt 10), pp.2881–2889.

Christner, M. et al., 2010. The giant extracellular matrix-binding protein of Staphylococcus epidermidis mediates biofilm accumulation and attachment to fibronectin. *Molecular Microbiology*, 75(1), pp.187–207.

Clarke, S.R. & Foster, S.J., 2006. Surface adhesins of Staphylococcus aureus. *Advances in microbial physiology*, 51, pp.187–224.

Clarke, S.R. et al., 2002. Analysis of Ebh, a 1.1-megadalton cell wall-associated fibronectin-binding protein of Staphylococcus aureus. *Infection and Immunity*, 70(12), pp.6680–6687.

Clauditz, A. et al., 2006. Staphyloxanthin plays a role in the fitness of Staphylococcus aureus and its ability to cope with oxidative stress. *Infection and Immunity*, 74(8), pp.4950–4953.

Clausen, M.-L. et al., 2013. Human β-defensin-2 as a marker for disease severity and skin barrier properties in atopic dermatitis. *The British journal of dermatology*, 169(3), pp.587–593.

Coates, R., Moran, J. & Horsburgh, M.J., 2014. Staphylococci: colonizers and pathogens of human skin. *Future Microbiology*, 9(1), pp.75–91.

Cogen, A.L. et al., 2010. Staphylococcus epidermidis antimicrobial delta-toxin (phenol-soluble modulin-gamma) cooperates with host antimicrobial peptides to kill group A Streptococcus. *PLoS ONE*, 5(1), p.e8557.

Cohan, F.M., 2001. Bacterial species and speciation. *Systematic biology*.

Colomer-Lluch, M., Jofre, J. & Muniesa, M., 2011. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS ONE*, 6(3), pp.e17549–e17549.

Conlan, S. et al., 2012. Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. *Genome Biology*, 13(7), p.R64.

Conrady, D.G. et al., 2008. A zinc-dependent adhesion module is responsible for intercellular adhesion in staphylococcal biofilms. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49), pp.19456–19461.

Cork, M.J. et al., 2009. Epidermal barrier dysfunction in atopic dermatitis. *The Journal of investigative dermatology*, 129(8), pp.1892–1908.

Costello, E.K. et al., 2009. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960), pp.1694–1697.

Cotter, P.D. & Hill, C., 2003. Surviving the acid test: responses of gram-positive bacteria to low pH. *Microbiology and molecular biology reviews : MMBR*, 67(3), pp.429–contents.

Courvalin, P., 1994. Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. *Antimicrobial Agents and Chemotherapy*.

Cox, G. et al., 2012. Ribosome clearance by FusB-type proteins mediates resistance to the antibiotic fusidic acid. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6), pp.2102–2107.

Cramer, P., Bushnell, D.A. & Kornberg, R.D., 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, 292(5523), pp.1863–1876.

Cucarella, C. et al., 2001. Bap, a Staphylococcus aureus Surface Protein Involved in Biofilm Formation. *Journal of Bacteriology*, 183(9), pp.2888–2896.

Cui, L. et al., 2009. Contribution of vraSR and graSR Point Mutations to Vancomycin Resistance in Vancomycin-Intermediate Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 53(3), pp.1231–1234.

Culebras, E. & Martínez, J.L., 1999. Aminoglycoside resistance mediated by the bifunctional enzyme 6'-N-aminoglycoside acetyltransferase-2"-O-aminoglycoside phosphotransferase. *Frontiers in Bioscience*, 4, pp.D1–D8.

Darling, A.C.E. et al., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7), pp.1394–1403.

Davis, J.M. & Ramakrishnan, L., 2009. The role of the granuloma in expansion and dissemination of early tuberculous infection. *Cell*, 136(1), pp.37–49.

Davison, A.J. & Taylor, P., 1987. Genetic relations between varicella-zoster virus and Epstein-Barr virus. *The Journal of general virology*, 68 (Pt 4), pp.1067–1079.

Dawson, R.J.P. & Locher, K.P., 2006. Structure of a bacterial multidrug ABC transporter. *Nature*, 443(7108), pp.180–185.

de Goffau, M.C. et al., 2009. Bacterial pleomorphism and competition in a relative humidity gradient. *Environmental Microbiology*, 11(4), pp.809–822.

de Goffau, M.C., van Dijl, J.M. & Harmsen, H.J.M., 2011. Microbial growth on the edge of desiccation. *Environmental Microbiology*, 13(8), pp.2328–2335.

de Silva, G.D.I. et al., 2002. The ica operon and biofilm production in coagulase-negative Staphylococci associated with carriage and disease in a neonatal intensive care unit. *Journal of Clinical Microbiology*, 40(2), pp.382–388.

Delaune, A. et al., 2011. Peptidoglycan crosslinking relaxation plays an important role in Staphylococcus aureus WalKR-dependent cell viability. *PLoS ONE*, 6(2), pp.e17054–e17054.

Delaune, A. et al., 2012. The WalKR system controls major staphylococcal virulence genes and is involved in triggering the host inflammatory response. *Infection and Immunity*, 80(10), pp.3438–3453.

Delves-Broughton, J. et al., 1996. Applications of the bacteriocin, nisin. *Antonie van Leeuwenhoek*, 69(2), pp.193–202.

DeMarco, C.E. et al., 2007. Efflux-related resistance to norfloxacin, dyes, and biocides in bloodstream isolates of Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 51(9), pp.3235–3239.

Desbois, A.P. & Smith, V.J., 2009. Antibacterial free fatty acids: activities, mechanisms of action and biotechnological potential. *Applied Microbiology and Biotechnology*, 85(6), pp.1629–1642.

Desvaux, M. et al., 2006. Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS Microbiology Letters*, 256(1), pp.1–15.

Dethlefsen, L., McFall-Ngai, M. & Relman, D.A., 2007. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449(7164), pp.811–818.

Devriese, L.A., Schleifer, K.H. & Adegoke, G.O., 1985. Identification of coagulase-negative staphylococci from farm animals. *The Journal of applied bacteriology*, 58(1), pp.45–55.

Débarbouillé, M. et al., 2009. Characterization of a serine/threonine kinase involved in virulence of Staphylococcus aureus. *Journal of Bacteriology*, 191(13), pp.4070–4081.

Diep, B.A. et al., 2006. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *Lancet*, 367(9512), pp.731–739.

Diep, B.A. et al., 2008. Contribution of Panton-Valentine Leukocidin in Community-Associated Methicillin-Resistant Staphylococcus aureus Pathogenesis A. J. Ratner, ed. *PLoS ONE*, 3(9), p.e3198.

Ding, Y. et al., 2008. NorB, an efflux pump in Staphylococcus aureus strain MW2, contributes to bacterial fitness in abscesses. *Journal of Bacteriology*, 190(21), pp.7123–7129.

Donat, S. et al., 2009. Transcriptome and functional analysis of the eukaryotic-type serine/threonine kinase PknB in Staphylococcus aureus. *Journal of Bacteriology*, 191(13), pp.4056–4069.

Donati, C. et al., 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology*, 11(10), p.R107.

Donnelly, M.I. & Cooper, R.A., 1981. Two succinic semialdehyde dehydrogenases are induced when Escherichia coli K-12 Is grown on gamma-aminobutyrate. *Journal of Bacteriology*, 145(3), pp.1425–1427.

Drlica, K., 2003. The mutant selection window and antimicrobial resistance. *The Journal of antimicrobial chemotherapy*, 52(1), pp.11–17.

Eid, J. et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910), pp.133–138.

Eisenstein, M., 2015. Startups use short-read data to expand long-read sequencing market. *Nature biotechnology*, 33(5), pp.433–435.

El-Metwally, S., 2014. *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*, Springer Science & Business.

Engelke, G. et al., 1992. Biosynthesis of the lantibiotic nisin: genomic organization and membrane localization of the NisB protein. *Applied and Environmental Microbiology*, 58(11), pp.3730–3743.

English, A.C. et al., 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, 7(11), p.e47768.

Enright, M.C. et al., 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *Journal of Clinical Microbiology*, 38(3), pp.1008–1015.

Ernst, C.M. & Peschel, A., 2011. Broad-spectrum antimicrobial peptide resistance by MprF-mediated aminoacylation and flipping of phospholipids. *Molecular Microbiology*, 80(2), pp.290–299.

Faires, N. et al., 1999. The catabolite control protein CcpA controls ammonium assimilation in Bacillus subtilis. *Journal of molecular microbiology and biotechnology*, 1(1), pp.141–148.

Falord, M. et al., 2012. GraXSR proteins interact with the VraFG ABC transporter to form a five-component system required for cationic antimicrobial peptide sensing and resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 56(2), pp.1047–1058.

Falord, M. et al., 2011. Investigation of the Staphylococcus aureus GraSR regulon reveals novel links to virulence, stress response and cell wall signal transduction pathways. *PLoS ONE*, 6(7), p.e21323.

Faria, N.A. et al., 2014. Nasal carriage of methicillin resistant staphylococci. *Microbial drug resistance (Larchmont, N.Y.)*, 20(2), pp.108–117.

Feehily, C. & Karatzas, K.A.G., 2013. Role of glutamate metabolism in bacterial responses towards acid and other stresses. *Journal of applied microbiology*, 114(1), pp.11–24.

Fehér, T. et al., 2012. Competition between transposable elements and mutator genes in bacteria. *Molecular Biology and Evolution*, 29(10), pp.3153–3159.

Feingold, K.R., 2007. Thematic review series: skin lipids. The role of epidermal lipids in cutaneous permeability barrier homeostasis. *The Journal of Lipid Research*, 48(12), pp.2531–2546.

Fey, P.D. & Olson, M.E., 2010. Current concepts in biofilm formation of Staphylococcus epidermidis. *Future Microbiology*, 5(6), pp.917–933.

Fierer, N. et al., 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(46), pp.17994–17999.

Finn, R.D. et al., 2014. Pfam: the protein families database. *Nucleic Acids Research*, 42(Database issue), pp.D222–30.

Fleischmann, R.D. et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223), pp.496–512.

Floyd, J.L. et al., 2010. LmrS is a multidrug efflux pump of the major facilitator superfamily from Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 54(12), pp.5406–5412.

Flusberg, B.A. et al., 2010. Direct detection of DNA methylation during single-molecule, real-time

sequencing. *Nature methods*, 7(6), pp.461–465.

Forsgren, A. & Sjöquist, J., 1966. "Protein A" from S. aureus. I. Pseudo-immune reaction with human gamma-globulin. *Journal of immunology (Baltimore, Md. : 1950)*, 97(6), pp.822–827.

Foster, J.W., 2004. Escherichia coli acid resistance: tales of an amateur acidophile. *Nature Reviews Microbiology*, 2(11), pp.898–907.

Frank, D.N. et al., 2010. The Human Nasal Microbiota and Staphylococcus aureus Carriage R. K. Aziz, ed. *PLoS ONE*, 5(5), p.e10598.

Frankel, M.B. et al., 2010. ABI domain-containing proteins contribute to surface protein display and cell division in Staphylococcus aureus. *Molecular Microbiology*, 78(1), pp.238–252.

Fridman, M. et al., 2013. Two unique phosphorylation-driven signaling pathways crosstalk in Staphylococcus aureus to modulate the cell-wall charge: Stk1/Stp1 meets GraSR. *Biochemistry*, 52(45), pp.7975–7986.

Friedman, L. & Kolter, R., 2004. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. *Molecular Microbiology*, 51(3), pp.675–690.

Friedrich, R. et al., 2003. Staphylocoagulase is a prototype for the mechanism of cofactor-induced zymogen activation. *Nature*, 425(6957), pp.535–539.

Frirdich, E. & Gaynor, E.C., 2013. Peptidoglycan hydrolases, bacterial shape, and pathogenesis. *Current Opinion in Microbiology*, 16(6), pp.767–778.

Furubayashi, M. et al., 2014. Construction of carotenoid biosynthetic pathways using squalene synthase. *FEBS Letters*, 588(3), pp.436–442.

Galbraith, H. & Miller, T.B., 1973. Effect of long chain fatty acids on bacterial respiration and amino acid uptake. *The Journal of applied bacteriology*, 36(4), pp.659–675.

Ganz, T., 2003. Defensins: antimicrobial peptides of innate immunity. *Nature Reviews Immunology*, 3(9), pp.710–720.

Geissler, S., Gotz, F. & Kupke, T., 1996. Serine protease EpiP from Staphylococcus epidermidis catalyzes the processing of the epidermin precursor peptide. *Journal of Bacteriology*, 178(1), pp.284–288.

Ghachi, El, M. et al., 2004. The bacA gene of Escherichia coli encodes an undecaprenyl pyrophosphate phosphatase activity. *Journal of Biological Chemistry*, 279(29), pp.30106–30113.

Götz, F. et al., 2014. Epidermin and gallidermin: Staphylococcal lantibiotics. *International journal of medical microbiology : IJMM*, 304(1), pp.63–71.

Greenway, D.L. & Dyke, K.G., 1979. Mechanism of the inhibitory action of linoleic acid on the growth of Staphylococcus aureus. *Journal of general microbiology*, 115(1), pp.233–245.

Gresham, H.D. et al., 2000. Survival of Staphylococcus aureus inside neutrophils contributes to infection. *Journal of immunology (Baltimore, Md. : 1950)*, 164(7), pp.3713–3722.

Grice, E.A. & Segre, J.A., 2011. The skin microbiome. *Nature Reviews Microbiology*, 9(4), pp.244–253.

Grice, E.A. et al., 2009. Topographical and Temporal Diversity of the Human Skin Microbiome. *Science*, 324(5931), pp.1190–1192.

Grinsted, J. & Lacey, R.W., 1973. Ecological and genetic implications of pigmentation in Staphylococcus aureus. *Journal of general microbiology*, 75(2), pp.259–267.

Grohmann, E., Muth, G. & Espinosa, M., 2003. Conjugative plasmid transfer in gram-positive bacteria. *Microbiology and molecular biology reviews : MMBR*, 67(2), pp.277–contents.

Groicher, K.H. et al., 2000. The Staphylococcus aureus lrgAB operon modulates murein hydrolase activity and penicillin tolerance. *Journal of Bacteriology*, 182(7), pp.1794–1801.

Guggenberger, C. et al., 2012. Two distinct coagulase-dependent barriers protect Staphylococcus aureus from neutrophils in a three dimensional in vitro infection model. *PLoS Pathogens*, 8(1), p.e1002434.

Gullberg, E. et al., 2014. Selection of a multidrug resistance plasmid by sublethal levels of antibiotics and heavy metals. *mBio*, 5(5), pp.e01918–e01914.

Gupta, V.K. et al., 2015. Divergences in gene repertoire among the reference Prevotella genomes derived from distinct body sites of human. *BMC Genomics*, 16, p.153.

Hachem, J.-P. et al., 2005. Sustained serine proteases activity by prolonged increase in pH leads to degradation of lipid processing enzymes and profound alterations of barrier function and stratum corneum integrity. *The Journal of investigative dermatology*, 125(3), pp.510–520.

Haggar, A. et al., 2004. The extracellular adherence protein from Staphylococcus aureus inhibits neutrophil binding to endothelial cells. *Infection and Immunity*, 72(10), pp.6164–6167.

Hall-Stoodley, L. & Stoodley, P., 2005. Biofilm formation and dispersal and the transmission of human pathogens. *Trends in Microbiology*, 13(1), pp.7–10.

Harder, J. et al., 2010. Enhanced expression and secretion of antimicrobial peptides in atopic dermatitis and after superficial skin injury. *Journal of Investigative Dermatology*, 130(5), pp.1355–1364.

Harrison, E.M. et al., 2013. A Staphylococcus xylosus isolate with a new mecC allotype. *Antimicrobial Agents and Chemotherapy*, 57(3), pp.1524–1528.

Hasper, H.E. et al., 2006. An alternative bactericidal mechanism of action for lantibiotic peptides that target lipid II. *Science*, 313(5793), pp.1636–1637.

Hazenbos, W.L.W. et al., 2013. Novel staphylococcal glycosyltransferases SdgA and SdgB mediate immunogenicity and protection of virulence-associated cell wall proteins. *PLoS Pathogens*, 9(10), p.e1003653.

Hefti, M.H. et al., 2004. The PAS fold. A redefinition of the PAS domain based upon structural prediction. *European journal of biochemistry / FEBS*, 271(6), pp.1198–1208.

Heidrich, C. et al., 1998. Isolation, characterization, and heterologous expression of the novel lantibiotic epicidin 280 and analysis of its biosynthetic gene cluster. *Applied and Environmental Microbiology*, 64(9), pp.3140–3146.

Heilmann, C., 2011. Adhesion Mechanisms of Staphylococci. In *Advances in Experimental Medicine and Biology*. Advances in Experimental Medicine and Biology. Dordrecht: Springer Netherlands, pp. 105–123.

Heilmann, C. et al., 1997. Evidence for autolysin-mediated primary attachment of Staphylococcus epidermidis to a polystyrene surface. *Molecular Microbiology*, 24(5), pp.1013–1024.

Hell, W., Meyer, H.G. & Gatermann, S.G., 1998. Cloning of aas, a gene encoding a Staphylococcus saprophyticus surface protein with adhesive and autolytic properties. *Molecular Microbiology*, 29(3), pp.871–881.

Higashi, Y., Strominger, J.L. & Sweeley, C.C., 1970. Biosynthesis of the Peptidoglycan of Bacterial Cell Walls XXI. Isolation of free C55-Isoprenoid alcohol and of lipid intermediates in peptidoglycan synthesis from *Staphylococcus aureus. Journal of Biological Chemistry*.

Higgins, C.F., 2001. ABC transporters: physiology, structure and mechanism--an overview. *Research in microbiology*, 152(3-4), pp.205–210.

Hiramatsu, K. et al., 2014. Multi-drug-resistant Staphylococcus aureus and future chemotherapy. *Journal of Infection and Chemotherapy*, 20(10), pp.593–601.

Hirasawa, Y. et al., 2009. Staphylococcus aureus Extracellular Protease CausesEpidermal Barrier Dysfunction. *Journal of Investigative Dermatology*, 130(2), pp.614–617.

Hiron, A. et al., 2011. Bacitracin and nisin resistance in Staphylococcus aureus: a novel pathway involving the BraS/BraR two-component system (SA2417/SA2418) and both the BraD/BraE and VraD/VraE ABC transporters. *Molecular Microbiology*, 81(3), pp.602–622.

Hoffmann, A., Schneider, T. & Pag, U., 2004. Localization and functional analysis of PepI, the immunity peptide of Pep5-producing Staphylococcus epidermidis strain 5. *Applied and ...*.

Hooper, L.V. et al., 2003. Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nature Immunology*, 4(3), pp.269–273.

Howden, B.P. et al., 2011. Evolution of multidrug resistance during Staphylococcus aureus infection involves mutation of the essential two component regulator WalKR. *PLoS Pathogens*, 7(11), pp.e1002359–e1002359.

Howden, B.P. et al., 2008. Genomic Analysis Reveals a Point Mutation in the Two-Component Sensor Gene graS That Leads to Intermediate Vancomycin Resistance in Clinical Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 52(10), pp.3755–3762.

Howell, M.D. et al., 2006. Mechanism of HBD-3 deficiency in atopic dermatitis. *Clinical immunology (Orlando, Fla.)*, 121(3), pp.332–338.

Hsu, S.-T.D. et al., 2004. The nisin-lipid II complex reveals a pyrophosphate cage that provides a blueprint for novel antibiotics. *Nature structural & molecular biology*, 11(10), pp.963–967.

Hu, X. et al., 2014. High quality draft genome sequence of Staphylococcus cohnii subsp. cohnii strain hu-01. *Standards in genomic sciences*, 9(3), pp.755–762.

Huang, J. et al., 2004. Novel chromosomally encoded multidrug efflux transporter MdeA in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 48(3), pp.909–917.

Huang, Y.-C. et al., 2007. The flexible and clustered lysine residues of human ribonuclease 7 are critical for membrane permeability and antimicrobial activity. *Journal of Biological Chemistry*, 282(7), pp.4626–4633.

Hussain, M. et al., 2001. Teichoic acid enhances adhesion of Staphylococcus epidermidis to immobilized fibronectin. *Microbial pathogenesis*, 31(6), pp.261–270.

Høiby, N. et al., 2010. Antibiotic resistance of bacterial biofilms. *International Journal of Antimicrobial Agents*, 35(4), pp.322–332.

International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC), 2009. Classification of staphylococcal cassette chromosome mec

(SCCmec): guidelines for reporting novel SCCmec elements. *Antimicrobial Agents and Chemotherapy*, 53(12), pp.4961–4967.

Islam, M.R. et al., 2012. Ring A of nukacin ISK-1: a lipid II-binding motif for type-A(II) lantibiotic. *Journal of the American Chemical Society*, 134(8), pp.3687–3690.

Iwai, I. et al., 2012. The Human Skin Barrier Is Organized as Stacked Bilayers of Fully Extended Ceramides with Cholesterol Molecules Associated with the Ceramide Sphingoid Moiety. *Journal of Investigative Dermatology*, 132(9), pp.2215–2225.

Iwase, T. et al., 2010. Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. *Nature*, 465(7296), pp.346–349.

Jacquet, E. et al., 2008. ATP hydrolysis and pristinamycin IIA inhibition of the Staphylococcus aureus Vga(A), a dual ABC protein involved in streptogramin A resistance. *Journal of Biological Chemistry*, 283(37), pp.25332–25339.

Jensen, J.-M. et al., 2011. Differential suppression of epidermal antimicrobial protein expression in atopic dermatitis and in EFAD mice by pimecrolimus compared to corticosteroids. *Experimental Dermatology*, 20(10), pp.783–788.

Jiang, S. et al., 2012. Whole-Genome Sequence of Staphylococcus hominis, an Opportunistic Pathogen. *Journal of Bacteriology*, 194(17), pp.4761–4762.

John, J.F., Gramling, P.K. & O'Dell, N.M., 1978. Species identification of coagulase-negative staphylococci from urinary tract isolates. *Journal of Clinical Microbiology*, 8(4), pp.435–437.

Johnsborg, O., Eldholm, V. & Håvarstein, L.S., 2007. Natural genetic transformation: prevalence, mechanisms and function. *Research in microbiology*, 158(10), pp.767–778.

Jones, J.W. et al., 1992. A study of coagulase-negative staphylococci with reference to slime production, adherence, antibiotic resistance patterns and clinical significance. *The Journal of hospital infection*, 22(3), pp.217–227.

Joshi, G.S. et al., 2011. Arginine catabolic mobile element encoded speG abrogates the unique hypersensitivity of Staphylococcus aureus to exogenous polyamines. *Molecular Microbiology*, 82(1), pp.9–20.

Jönsson, K. et al., 1995. Staphylococcus aureus expresses a major histocompatibility complex class II analog. *Journal of Biological Chemistry*, 270(37), pp.21457–21460.

Jørgensen, T.S. et al., 2014. Current strategies for mobilome research. *Frontiers in microbiology*, 5, pp.750–750.

Kaatz, G.W., McAleese, F. & Seo, S.M., 2005. Multidrug resistance in Staphylococcus aureus due to overexpression of a novel multidrug and toxin extrusion (MATE) transport protein. *Antimicrobial Agents and Chemotherapy*, 49(5), pp.1857–1864.

Kanda, N. et al., 2011. Human β-defensin-2 enhances IFN-γ and IL-10 production and suppresses IL-17 production in T cells. *Journal of Leukocyte Biology*, 89(6), pp.935–944.

Kaoutari, El, A. et al., 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7), pp.497–504.

Kelly, J., 2013. Metagenomic and genomic analysis of the skin microbiota.

Kenny, J.G. et al., 2009. The Staphylococcus aureus Response to Unsaturated Long Chain Free Fatty Acids: Survival Mechanisms and Virulence Implications D. Davis, ed. *PLoS ONE*, 4(2), p.e4344.

Kersters, K. & Vancanneyt, M., 2005. *Bergey's manual of systematic bacteriology*,

Kim, P.I. et al., 2010. Characterization and structure identification of an antimicrobial peptide, hominicin, produced by Staphylococcus hominis MBBL 2-9. *Biochemical and biophysical research communications*, 399(2), pp.133–138.

King, N.P. et al., 2012. Characterisation of a cell wall-anchored protein of Staphylococcus saprophyticus associated withlinoleic acid resistance. *BMC Microbiology*, 12(1), p.8.

King, N.P. et al., 2011. UafB is a serine-rich repeat adhesin of Staphylococcus saprophyticus that mediates binding to fibronectin, fibrinogen and human uroepithelial cells. *Microbiology*, 157(Pt 4), pp.1161–1175.

Kinnevey, P.M. et al., 2014. Extensive Genetic Diversity Identified among Sporadic Methicillin-Resistant Staphylococcus aureus Isolates Recovered in Irish Hospitals between 2000 and 2012. *Antimicrobial Agents and Chemotherapy*, 58(4), pp.1907–1917.

Klappenbach, J.A., Dunbar, J.M. & Schmidt, T.M., 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, 66(4), pp.1328–1333.

Kloos, W.E., 1980. Natural populations of the genus Staphylococcus. *Annual review of microbiology*, 34, pp.559–592.

Kloos, W.E. & Schleifer, K.H., 1975. Isolation and Characterization of Staphylococci from Human Skin II. Descriptions of Four New Species: Staphylococcus warneri, Staphylococcus capitis, Staphylococcus hominis, and Staphylococcus simulans. *International Journal of Systematic Bacteriology*, 25(1), pp.62–79.

Kloos, W.E., George, C.G. & Olgiate, J.S., 1998. Staphylococcus hominis subsp. novobiosepticus subsp. nov., a novel trehalose-and N-acetyl-D-glucosamine-negative, novobiocin-and multiple-antibiotic-resistant …. *International journal …*.

Kloss, W.E. & Pattee, P.A., 1965. Transduction analysis of the histidine region in *Staphylococcus aureus. Journal of general microbiology*, 39, pp.195–207.

Koenig, R.L. et al., 2004. Staphylococcus aureus AgrA binding to the RNAIII-agr regulatory region. *Journal of Bacteriology*, 186(22), pp.7549–7555.

Kolar, S.L., 2012. The Role and Regulation of NsaRS: a Cell-Envelope Stress Sensing Two-Component System in Staphylococcus aureus.

Kolar, S.L. et al., 2011. NsaRS is a cell-envelope-stress-sensing two-component system of Staphylococcus aureus. *Microbiology*, 157(8), pp.2206–2219.

Kong, H.H. et al., 2012. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Research*, 22(5), pp.850–859.

Koonin, E.V. & Galperin, M., 2013. *Sequence — Evolution — Function*, Springer Science & Business Media.

Kopfnagel, V., Harder, J. & Werfel, T., 2013. Expression of antimicrobial peptides in atopic dermatitis and possible immunoregulatory functions. *Current opinion in allergy and clinical immunology*, 13(5), pp.531–536.

Koponen, O., 2004. *Studies Of Producer Self–Protection And Nisin Biosynthesis Of Lactococcus lactis, Doctoral issertation, Institute Of Biotech*, And Department Of Appl. Chem. Microbiol. Helsinki.

Koponen, O. et al., 2004. Distribution of the NisI immunity protein and enhancement of nisin activity by the lipid-free NisI. *FEMS Microbiology Letters*, 231(1), pp.85–90.

Kordel, M., Benz, R. & Sahl, H.G., 1988. Mode of action of the staphylococcinlike peptide Pep 5: voltage-dependent depolarization of bacterial and artificial membranes. *Journal of Bacteriology*, 170(1), pp.84–88.

Koren, S. & Phillippy, A.M., 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23, pp.110–120.

Koren, S. et al., 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, 14(9), pp.R101–R101.

Kos, V.N. et al., 2012. Comparative genomics of vancomycin-resistant Staphylococcus aureus strains and their positions within the clade most commonly associated with Methicillin-resistant S. aureus hospital-acquired infection in the United States. *mBio*, 3(3), pp.–.

Kozitskaya, S. et al., 2004. The bacterial insertion sequence element IS256 occurs preferentially in nosocomial Staphylococcus epidermidis isolates: association with biofilm formation and resistance to aminoglycosides. *Infection and Immunity*, 72(2), pp.1210–1215.

Köser, C.U. et al., 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine*, 366(24), pp.2267–2275.

Krishna, S.S., Majumdar, I. & Grishin, N.V., 2003. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Research*, 31(2), pp.532–550.

Kuhn, S., Slavetinsky, C.J. & Peschel, A., 2015. Synthesis and function of phospholipids in Staphylococcus aureus. *International journal of medical microbiology : IJMM*, 305(2), pp.196–202.

Kuipers, O.P. et al., 1995. Autoregulation of nisin biosynthesis in Lactococcus lactis by signal transduction. *Journal of Biological Chemistry*, 270(45), pp.27299–27304.

Kuroda, M. et al., 2008. Staphylococcus aureus giant protein Ebh is involved in tolerance to transient hyperosmotic pressure. *Biochemical and biophysical research communications*, 374(2), pp.237–241.

Kuroda, M. et al., 2003. Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in Staphylococcus aureus. *Molecular Microbiology*, 49(3), pp.807–821.

Kuroda, M. et al., 2004. Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in Staphylococcus aureus. *Molecular Microbiology*, 49(3), pp.807–821.

Kuroda, M. et al., 2005. Whole genome sequence of Staphylococcus saprophyticus reveals the pathogenesis of uncomplicated urinary tract infection. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37), pp.13272–13277.

Kuroda, M., Hayashi, H. & Ohta, T., 1999. Chromosome-determined zinc-responsible operon czr in Staphylococcus aureus strain 912. *Microbiology and Immunology*, 43(2), pp.115–125.

Kwok, A.Y.C. & Chow, A.W., 2003. Phylogenetic study of Staphylococcus and Macrococcus species based on partial hsp60 gene sequences. *International Journal of Systematic and Evolutionary Microbiology,* 53(Pt 1), pp.87–92.

Laarman, A.J. et al., 2011. Staphylococcus aureus metalloprotease aureolysin cleaves complement C3 to mediate immune evasion. *Journal of immunology (Baltimore, Md. : 1950)*,

186(11), pp.6445–6453.

Labandeira-Rey, M. et al., 2007. Staphylococcus aureus Panton-Valentine Leukocidin Causes Necrotizing Pneumonia. *Science*, 315(5815), pp.1130–1133.

Labrie, S.J. et al., 2014. First Complete Genome Sequence of Staphylococcus xylosus, a Meat Starter Culture and a Host to Propagate Staphylococcus aureus Phages. *Genome announcements*, 2(4).

Ladhani, S., 2003. Understanding the mechanism of action of the exfoliative toxins of Staphylococcus aureus. *FEMS Immunology & Medical Microbiology*, 39(2), pp.181–189.

Lai, Y. et al., 2010. Activation of TLR2 by a small molecule produced by Staphylococcus epidermidis increases antimicrobial defense against bacterial skin infections. *The Journal of investigative dermatology*, 130(9), pp.2211–2221.

Lai, Y. et al., 2006. The human anionic antimicrobial peptide dermcidin induces proteolytic defence mechanisms in staphylococci. *Molecular Microbiology*, 63(2), pp.497–506.

Lamers, R.P. et al., 2012. Phylogenetic relationships among Staphylococcus species and refinement of cluster groups based on multilocus data. *BMC evolutionary biology*, 12, p.171.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Lang, S. et al., 2000. Identification of a novel antigen fromStaphylococcus epidermidis. *FEMS Immunology & Medical Microbiology*, 29(3), pp.213–220.

Lasa, I. & PenadEs, J.R., 2006. Bap: a family of surface proteins involved in biofilm formation. *Research in microbiology*, 157(2), pp.99–107.

Lee, G.C. et al., 2015. Comparative whole genome sequencing of community-associated methicillin-resistant Staphylococcus aureus sequence type 8 from primary care clinics in a Texas community. *Pharmacotherapy*, 35(2), pp.220–228.

Lee, Y.H., Nam, K.H. & Helmann, J.D., 2013. A mutation of the RNA polymerase β' subunit (rpoC) confers cephalosporin resistance in Bacillus subtilis. *Antimicrobial Agents and Chemotherapy*, 57(1), pp.56–65.

Li, B. et al., 2006. Structure and mechanism of the lantibiotic cyclase involved in nisin biosynthesis. *Science*, 311(5766), pp.1464–1467.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Li, L., Stoeckert, C.J. & Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9), pp.2178–2189.

Li, M. et al., 2009. Molecular characterization of Staphylococcus epidermidis strains isolated from a teaching hospital in Shanghai, China. *Journal of Medical Microbiology*, 58(Pt 4), pp.456–461.

Li, R. et al., 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2), pp.265–272.

Li, X. et al., 2015. Overexpression of specific proton motive force-dependent transporters facilitate the export of surfactin in Bacillus subtilis. *Journal of Industrial Microbiology & Biotechnology*, 42(1), pp.93–103.

Libberton, B. et al., 2014. Evidence that intraspecific trait variation among nasal bacteria shapes the distribution of Staphylococcus aureus. *Infection and Immunity*, 82(9), pp.3811–3815.

Linke, D. & Goldman, A., 2011. *Bacterial Adhesion*, Springer.

Linnes, J.C., Ma, H. & Bryers, J.D., 2013. Giant extracellular matrix binding protein expression in Staphylococcus epidermidis is regulated by biofilm formation and osmotic pressure. *Current Microbiology*, 66(6), pp.627–633.

Liu, B. & Pop, M., 2009. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Research*, 37(Database issue), pp.D443–D447.

Liu, C.-I. et al., 2008. A cholesterol biosynthesis inhibitor blocks Staphylococcus aureus virulence. *Science*, 319(5868), pp.1391–1394.

Liu, G.Y. & Nizet, V., 2009. Color me bad: microbial pigments as virulence factors. *Trends in Microbiology*, 17(9), pp.406–413.

Liu, G.Y. et al., 2005. Staphylococcus aureus golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. *The Journal of …*.

Liu, L. et al., 2012. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012, p.251364.

Loman, N.J., Quick, J. & Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*.

Long, J.P. et al., 1992. The production of fatty acid modifying enzyme (FAME) and lipase by various staphylococcal species. *Journal of Medical Microbiology*, 37(4), pp.232–234.

Lowder, B.V. & Fitzgerald, J.R., 2010. Human origin for avian pathogenic Staphylococcus aureus. *Virulence*, 1(4), pp.283–284.

Lowder, B.V. et al., 2009. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, 106(46), pp.19545–19550.

Lower, S.K. et al., 2011. Polymorphisms in fibronectin binding protein A of Staphylococcus aureus are associated with infection of cardiovascular devices. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45), pp.18372–18377.

Löffler, B. et al., 2010. Staphylococcus aureus Panton-Valentine Leukocidin Is a Very Potent Cytotoxic Factor for Human Neutrophils A. Cheung, ed. *PLoS Pathogens*, 6(1), p.e1000715.

Ma, X.X. et al., 2002. Novel type of staphylococcal cassette chromosome mec identified in community-acquired methicillin-resistant Staphylococcus aureus strains. *Antimicrobial Agents and Chemotherapy*, 46(4), pp.1147–1152.

Macintosh, R.L. et al., 2009. The terminal A domain of the fibrillar accumulation-associated protein (Aap) of Staphylococcus epidermidis mediates adhesion to human corneocytes. *Journal of Bacteriology*, 191(22), pp.7007–7016.

Macía-Heras, M. & Donate-Correa, J., 2013. Daptomycin, Methicillin Resistant Staphylococcus hominis Catheter-Related Bacteraemia in a Hemodialysis Patient. *Open Journal of Medical …*.

Mahillon, J. & Chandler, M., 1998. Insertion sequences. *Microbiology and molecular biology reviews : MMBR*, 62(3), pp.725–774.

Manganelli, R. & van de Rijn, I., 1999. Characterization of emb, a gene encoding the major

adhesin of Streptococcus defectivus. *Infection and Immunity*, 67(1), pp.50–56.

Manichanh, C. et al., 2010. Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome Research*, 20(10), pp.1411–1419.

Mann, E.E. et al., 2009. Modulation of eDNA release and degradation affects Staphylococcus aureus biofilm maturation. *PLoS ONE*, 4(6), p.e5822.

Mann, R.A. et al., 2013. Comparative genomics of 12 strains of Erwinia amylovora identifies a pan-genome with a large conserved core. *PLoS ONE*, 8(2), p.e55644.

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3), pp.133–141.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–380.

Mascher, T., 2014. Bacterial (intramembrane-sensing) histidine kinases: signal transfer rather than stimulus perception. *Trends in Microbiology*, 22(10), pp.559–565.

Mascher, T., 2006. Intramembrane-sensing histidine kinases: a new family of cell envelope stress sensors in Firmicutes bacteria. *FEMS Microbiology Letters*, 264(2), pp.133–144.

Mazmanian, S.K. et al., 2003. Passage of heme-iron across the envelope of Staphylococcus aureus. *Science*, 299(5608), pp.906–909.

Mazmanian, S.K. et al., 1999. Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall. *Science*, 285(5428), pp.760–763.

Mazmanian, S.K., Ton-That, H. & Schneewind, O., 2001. Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. *Molecular Microbiology*, 40(5), pp.1049–1057.

McAleese, F. et al., 2006. Overexpression of genes of the cell wall stimulon in clinical isolates of Staphylococcus aureus exhibiting vancomycin-intermediate- S. aureus-type resistance to vancomycin. *Journal of Bacteriology*, 188(3), pp.1120–1133.

McAuliffe, O., Ross, R.P. & Hill, C., 2001. Lantibiotics: structure, biosynthesis and mode of action. *FEMS microbiology reviews*, 25(3), pp.285–308.

McCarthy, A.J. & Lindsay, J.A., 2010. Genetic variation in Staphylococcus aureus surface and immune evasion genes is lineage associated: implications for vaccine design and host-pathogen interactions. *BMC Microbiology*, 10, p.173.

McDevitt, D. et al., 1994. Molecular characterization of the clumping factor (fibrinogen receptor) of Staphylococcus aureus. *Molecular Microbiology*, 11(2), pp.237–248.

Medini, D. et al., 2005. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6), pp.589–594.

Meehl, M. et al., 2007. Interaction of the GraRS Two-Component System with the VraFG ABC Transporter To Support Vancomycin-Intermediate Resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 51(8), pp.2679–2689.

Mendoza-Olazarán, S. et al., 2013. Microbiological and molecular characterization of Staphylococcus hominis isolates from blood. *PLoS ONE*, 8(4), p.e61161.

Meyer, C. et al., 1995. Nucleotide sequence of the lantibiotic Pep5 biosynthetic gene cluster and functional analysis of PepP and PepC. Evidence for a role of PepC in thioether formation.

European journal of biochemistry / FEBS, 232(2), pp.478–489.

Méric, G. et al., 2015. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. *Genome biology and evolution*, 7(5), pp.1313–1328.

Mishra, N.N. et al., 2011. Carotenoid-related alteration of cell membrane fluidity impacts Staphylococcus aureus susceptibility to host defense peptides. *Antimicrobial Agents and Chemotherapy*, 55(2), pp.526–531.

Mohammadi, T. et al., 2011. Identification of FtsW as a transporter of lipid-linked cell wall precursors across the membrane. *EMBO Journal*, 30(8), pp.1425–1432.

Montanaro, L. et al., 2007. Antibiotic multiresistance strictly associated with IS256 and ica genes in Staphylococcus epidermidis strains from implant orthopedic infections. *Journal of Biomedical Materials Research Part A*, 83(3), pp.813–818.

Montgomery, C.P. et al., 2012. CodY deletion enhances in vivo virulence of community-associated methicillin-resistant Staphylococcus aureus clone USA300. *Infection and Immunity*, 80(7), pp.2382–2389.

Moon, B.Y. et al., 2015. Phage-mediated horizontal transfer of a Staphylococcus aureus virulence-associated genomic island. *Scientific reports*, 5, p.9784.

Moreillon, P. et al., 1995. Role of Staphylococcus aureus coagulase and clumping factor in pathogenesis of experimental endocarditis. *Infection and Immunity*, 63(12), pp.4738–4743.

Morikawa, K. et al., 2012. Expression of a cryptic secondary sigma factor gene unveils natural competence for DNA transformation in Staphylococcus aureus. *PLoS Pathogens*, 8(11), p.e1003003.

Morikawa, K. et al., 2001. Overexpression of Sigma Factor, σB, Urges Staphylococcus aureus to Thicken the Cell Wall and to Resist β-Lactams. *Biochemical and biophysical research communications*, 288(2), pp.385–389.

Morizane, S. et al., 2010. Kallikrein Expression and Cathelicidin Processing Are Independently Controlled in Keratinocytes by Calcium, Vitamin D3, and Retinoic Acid. *Journal of Investigative Dermatology*, 130(5), pp.1297–1306.

Mulcahy, M.E. et al., 2012. Nasal colonisation by Staphylococcus aureus depends upon clumping factor B binding to the squamous epithelial cell envelope protein loricrin. *PLoS Pathogens*, 8(12), pp.e1003092–e1003092.

Mullikin, J.C., 2014. The evolution of comparative genomics. *Molecular Genetics & Genomic Medicine*, 2(5), pp.363–368.

Murakami, M. et al., 2002. Cathelicidin anti-microbial peptide expression in sweat, an innate defense system for the skin. *The Journal of investigative dermatology*, 119(5), pp.1090–1095.

Müller, A. et al., 2012. Interaction of type A lantibiotics with undecaprenol-bound cell envelope precursors. *Microbial drug resistance (Larchmont, N.Y.)*, 18(3), pp.261–270.

Mwangi, M.M. et al., 2007. Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 104(22), pp.9451–9456.

Myers, E.W. et al., 2000. A whole-genome assembly of Drosophila. *Science*, 287(5461), pp.2196–2204.

Nagao, J.-I. et al., 2005. Lanthionine introduction into nukacin ISK-1 prepeptide by co-expression with modification enzyme NukM in Escherichia coli. *Biochemical and biophysical research communications*, 336(2), pp.507–513.

Nagase, N. et al., 2002. Isolation and species distribution of staphylococci from animal and human skin. *The Journal of veterinary medical science / the Japanese Society of Veterinary Science*, 64(3), pp.245–250.

Nakamura, Y. et al., 2013. Staphylococcus δ-toxin induces allergic skin disease by activating mast cells. *Nature*, 503(7476), pp.397–401.

Narzisi, G. & Mishra, B., 2011. Comparing de novo genome assembly: the long and short of it. *PLoS ONE*, 6(4), p.e19175.

Navarre, W.W. & Schneewind, O., 1999. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiology and molecular biology reviews : MMBR*, 63(1), pp.174–229.

Noble, W.C., 1977. Variation in the prevalence of antibiotic resistance of Staphylococcus aureus from human skin and nares. *Journal of general microbiology*, 98(1), pp.125–132.

Nomura, I. et al., 2003. Cytokine milieu of atopic dermatitis, as compared to psoriasis, skin prevents induction of innate immune response genes. *Journal of immunology (Baltimore, Md. : 1950)*, 171(6), pp.3262–3269.

Novick, R.P., 2003. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands ofStaphylococcus aureus. *Plasmid*, 49(2), pp.93–105.

Nurk, S. et al., 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of computational biology : a journal of computational molecular cell biology*, 20(10), pp.714–737.

Obolski, U. et al., 2014. Resistance profiles of coagulase-negative staphylococci contaminating blood cultures predict pathogen resistance and patient mortality. *Journal of Antimicrobial Chemotherapy*.

Oh, J. et al., 2014. Biogeography and individuality shape function in the human skin metagenome. *Nature*, 514(7520), pp.59–64.

Ohniwa, R.L., Kitabayashi, K. & Morikawa, K., 2013. Alternative cardiolipin synthase Cls1 compensates for stalled Cls2 function in Staphylococcus aureus under conditions of acute acid stress. *FEMS Microbiology Letters*, 338(2), pp.141–146.

Olivier, A.C. et al., 2009. Role of rsbU and staphyloxanthin in phagocytosis and intracellular growth of Staphylococcus aureus in human macrophages and endothelial cells. *The Journal of Infectious Diseases*, 200(9), pp.1367–1370.

Orrett, F.A. & Shurland, S.M., 1998. Significance of coagulase-negative staphylococci in urinary tract infections in a developing country. *Connecticut medicine*.

Otto, M., 2013. Staphylococcal infections: mechanisms of biofilm maturation and detachment as critical determinants of pathogenicity. *Medicine*, 64, pp.175–188.

Otto, M., 2009. Staphylococcus epidermidis — the “accidental” pathogen. *Nature Reviews Microbiology*, 7(8), pp.555–567.

Overton, I.M. et al., 2011. Global network analysis of drug tolerance, mode of action and virulence in methicillin-resistant S. aureus. *BMC Systems Biology*, 5(1), pp.68–68.

Pallen, M.J. et al., 2001. An embarrassment of sortases - a richness of substrates? *Trends in Microbiology*, 9(3), pp.97–102.

Palmqvist, N. et al., 2004. Expression of staphylococcal clumping factor A impedes macrophage phagocytosis. *Microbes and Infection*, 6(2), pp.188–195.

Palys, T. et al., 2000. Protein-coding genes as molecular markers for ecologically distinct populations: the case of two Bacillus species. *International Journal of Systematic and Evolutionary Microbiology*, 50 Pt 3, pp.1021–1028.

Papadimitriou, K. et al., 2014. Comparative genomics of the dairy isolate Streptococcus macedonicus ACA-DC 198 against related members of the Streptococcus bovis/Streptococcus equinus complex. *BMC Genomics*, 15, p.272.

Paul, W.E. & Zhu, J., 2010. How are TH2-type immune responses initiated and amplified? *Nature Reviews Immunology*.

Paulsen, I.T. et al., 1995. Molecular characterization of the staphylococcal multidrug resistance export protein QacC. *Journal of Bacteriology*, 177(10), pp.2827–2833.

Pál, C. & Papp, B., 2013. From passengers to drivers: Impact of bacterial transposable elements on evolvability. *Mobile Genetic Elements*, 3(1), pp.–.

Pelz, A. et al., 2005. Structure and biosynthesis of staphyloxanthin from Staphylococcus aureus. *The Journal of biological chemistry*, 280(37), pp.32493–32498.

Periasamy, S. et al., 2012. Phenol-soluble modulins in staphylococci: What are they originally for? *Communicative & Integrative Biology*, 5(3), pp.275–277.

Peschel, A. et al., 1999. Inactivation of the dlt operon in Staphylococcus aureus confers sensitivity to defensins, protegrins, and other antimicrobial peptides. *The Journal of biological chemistry*, 274(13), pp.8405–8410.

Pevzner, P.A., Tang, H. & Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp.9748–9753.

Poelarends et al., 2000. An ABC-type multidrug transporter of Lactococcus lactis possesses an exceptionally broad substrate specificity. *Drug Resistance Updates*, 3(6), pp.5–5.

Ponnuraj, K. et al., 2003. A “dock, lock, and latch” structural model for a staphylococcal adhesin binding to fibrinogen. *Cell*, 115(2), pp.217–228.

Poston, S.M. & Hee, F., 1991. Genetic characterisation of resistance to metal ions in methicillin-resistant Staphylococcus aureus: elimination of resistance to cadmium, mercury and tetracycline with …. *Journal of Medical Microbiology*.

Prober, J.M. et al., 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*, 238(4825), pp.336–341.

Probst, A.J. et al., 1998. Staphylococcus condimenti sp. nov., from soy sauce mash, and Staphylococcus carnosus (Schleifer and Fischer 1982) subsp. utilis subsp. nov. *International Journal of Systematic Bacteriology*, 48 Pt 3, pp.651–658.

Pruitt, K.D. et al., 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(Database issue), pp.D130–5.

Punj, V. et al., 2000. Phagocytic cell killing mediated by secreted cytotoxic factors of Vibrio cholerae. *Infection and Immunity*, 68(9), pp.4930–4937.

Quax, T.E.F. et al., 2015. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, 59(2), pp.149–161.

Quiberoni, A. et al., 2001. Distinctive features of homologous recombination in an "old" microorganism, Lactococcus lactis. *Research in microbiology*, 152(2), pp.131–139.

Quiblier, C. et al., 2013. The Staphylococcus aureus Membrane Protein SA2056 Interacts with Peptidoglycan Synthesis Enzymes. *Antibiotics*, 2(1), pp.11–27.

Quirós, P. et al., 2014. Antibiotic resistance genes in the bacteriophage DNA fraction of human fecal samples. *Antimicrobial Agents and Chemotherapy*, 58(1), pp.606–609.

Rani, S.A. et al., 2007. Spatial patterns of DNA replication, protein synthesis, and oxygen concentration within bacterial biofilms reveal diverse physiological states. *Journal of Bacteriology*, 189(11), pp.4223–4233.

Rasko, D.A. et al., 2005. Genomics of the Bacillus cereus group of organisms. *FEMS microbiology reviews*, 29(2), pp.303–329.

Ravaioli, S. et al., 2012. Staphylococcus lugdunensis, an aggressive coagulase-negative pathogen not to be underestimated. *International Journal of Artificial Organs*, 35(10), pp.742–753.

Reagan, D.R. et al., 1991. Elimination of coincident Staphylococcus aureus nasal and hand carriage with intranasal application of mupirocin calcium ointment. *Annals of internal medicine*, 114(2), pp.101–106.

Reizer, J. et al., 1999. Novel phosphotransferase system genes revealed by genome analysis - the complete complement of PTS proteins encoded within the genome of Bacillus subtilis. *Microbiology (Reading, England)*, 145 (Pt 12), pp.3419–3429.

Renzoni, A. et al., 2011. Whole genome sequencing and complete genetic analysis reveals novel pathways to glycopeptide resistance in Staphylococcus aureus. *PLoS ONE*, 6(6), p.e21577.

Ribeiro, F.J. et al., 2012. Finished bacterial genomes from shotgun sequence data. *Genome Research*, 22(11), pp.2270–2277.

Rice, K.C. et al., 2007. The cidA murein hydrolase regulator contributes to DNA release and biofilm development in Staphylococcus aureus. *Proceedings of the National Academy of Sciences of the United States of America*, 104(19), pp.8113–8118.

Rice, R.H. & Green, H., 1977. The cornified envelope of terminally differentiated human epidermal keratinocytes consists of cross-linked protein. *Cell*, 11(2), pp.417–422.

Rieg, S. et al., 2005. Deficiency of dermcidin-derived antimicrobial peptides in sweat of patients with atopic dermatitis correlates with an impaired innate defense of human skin in vivo. *The Journal of …*.

Rietkötter, E., Hoyer, D. & Mascher, T., 2008. Bacitracin sensing in Bacillus subtilis. *Molecular Microbiology*, 68(3), pp.768–785.

Riley, M.A. & Wertz, J.E., 2002. Bacteriocins: evolution, ecology, and application. *Annual review of microbiology*, 56, pp.117–137.

Roberson, J.R. et al., 1992. Evaluation of methods for differentiation of coagulase-positive staphylococci. *Journal of Clinical Microbiology*, 30(12), pp.3217–3219.

Roberts, A.P. et al., 2008. Revised nomenclature for transposable genetic elements. *Plasmid*, 60(3), pp.167–173.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature biotechnology*, 29(1), pp.24–26.

Roche, F.M., Massey, R., et al., 2003. Characterization of novel LPXTG-containing proteins of Staphylococcus aureus identified from genome sequences. *Microbiology (Reading, England)*, 149(Pt 3), pp.643–654.

Roche, F.M., Meehan, M. & Foster, T.J., 2003. The Staphylococcus aureus surface protein SasG and its homologues promote bacterial adherence to human desquamated nasal epithelial cells. *Microbiology (Reading, England)*, 149(Pt 10), pp.2759–2767.

Romantsov, T., Guan, Z. & Wood, J.M., 2009. Cardiolipin and the osmotic stress responses of bacteria. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1788(10), pp.2092–2100.

Romero-Martinez, R. et al., 2000. Biosynthesis and functions of melanin in Sporothrix schenckii. *Infection and Immunity*, 68(6), pp.3696–3703.

Rouch, D.A. et al., 1987. The aacA-aphD gentamicin and kanamycin resistance determinant of Tn4001 from Staphylococcus aureus: expression and nucleotide sequence analysis. *Journal of general microbiology*, 133(11), pp.3039–3052.

Röhrl, J. et al., 2010. Human beta-defensin 2 and 3 and their mouse orthologs induce chemotaxis through interaction with CCR2. *Journal of immunology (Baltimore, Md. : 1950)*, 184(12), pp.6688–6694.

Römling, U. et al., 2000. AgfD, the checkpoint of multicellular and aggregative behaviour in Salmonella typhimurium regulates at least two independent pathways. *Molecular Microbiology*, 36(1), pp.10–23.

Rutherford, K. et al., 2000. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10), pp.944–945.

Sabala, I. et al., 2014. Crystal structure of the antimicrobial peptidase lysostaphin from Staphylococcus simulans. *FEBS Journal*, 281(18), pp.4112–4122.

Sadykov, M.R. et al., 2010. Using NMR metabolomics to investigate tricarboxylic acid cycle-dependent signal transduction in Staphylococcus epidermidis. *The Journal of biological chemistry*, 285(47), pp.36616–36624.

Sakinç, T. et al., 2009. SdrI of Staphylococcus saprophyticus is a multifunctional protein: localization of the fibronectin-binding site. *FEMS Microbiology Letters*, 301(1), pp.28–34.

Sakinç, T. et al., 2005. The surface-associated protein of Staphylococcus saprophyticus is a lipase. *Infection and Immunity*, 73(10), pp.6419–6428.

Sandiford, S. & Upton, M., 2012. Identification, Characterization, and Recombinant Expression of Epidermicin NI01, a Novel Unmodified Bacteriocin Produced by Staphylococcus epidermidis That Displays Potent Activity against Staphylococci. *Antimicrobial Agents and Chemotherapy*, 56(3), pp.1539–1547.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Santos Nascimento, dos, J. et al., 2005. Production of bacteriocins by coagulase-negative staphylococci involved in bovine mastitis. *Veterinary microbiology*, 106(1-2), pp.61–71.

Sashihara, T. et al., 2000. A novel lantibiotic, nukacin ISK-1, of Staphylococcus warneri ISK-1: cloning of the structural gene and identification of the structure. *Bioscience, biotechnology, and biochemistry*, 64(11), pp.2420–2428.

Schaffer, A.C. et al., 2006. Immunization with Staphylococcus aureus clumping factor B, a major determinant in nasal carriage, reduces nasal colonization in a murine model. *Infection and Immunity*, 74(4), pp.2145–2153.

Scharn, C.R., Tenover, F.C. & Goering, R.V., 2013. Transduction of staphylococcal cassette chromosome mec elements between strains of Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 57(11), pp.5233–5238.

Schattner, P., Brooks, A.N. & Lowe, T.M., 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33(Web Server issue), pp.W686–9.

Schechter-Perkins, E.M. et al., 2011. Prevalence and predictors of nasal and extranasal staphylococcal colonization in patients presenting to the emergency department. *Annals of emergency medicine*, 57(5), pp.492–499.

Schleifer, K.H. & Fischer, U., 1982. Description of a new species of the genus Staphylococcus: Staphylococcus carnosus. *International Journal of Systematic …*.

Schleifer, K.H., Kilpper-Bälz, R. & Devriese, L.A., 1984. Staphylococcus arlettaesp. nov.,S. equorumsp. nov. and S. k1oosii sp. nov.: Three New Coagulase-Negative, Novobiocin-Resistant Species from Animals. *Systematic and Applied Microbiology*, 5(4), pp.501–509.

Schloss, P.D., 2014. Microbiology: An integrated view of the skin microbiome. *Nature*, 514(7520), pp.44–45.

Schneewind, O. & Missiakas, D.M., 2012. Protein secretion and surface display in Gram-positive bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1592), pp.1123–1139.

Schroeder, K. et al., 2009. Molecular characterization of a novel Staphylococcus aureus surface protein (SasC) involved in cell aggregation and biofilm accumulation. *PLoS ONE*, 4(10), p.e7567.

Schwartz, S. et al., 2003. Human-mouse alignments with BLASTZ. *Genome Research*, 13(1), pp.103–107.

Schwartz, S. et al., 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Research*, 10(4), pp.577–586.

Schwarz, S. et al., 2014. Plasmid-mediated antimicrobial resistance in staphylococci and other Firmicutes.

Schwarz, S. et al., 2011. Plasmid-mediated resistance to protein biosynthesis inhibitors in staphylococci. *Annals of the New York Academy of Sciences*, 1241(1), pp.82–103.

Scriba, T.J. et al., 2008. The Staphyloccous aureus Eap protein activates expression of proinflammatory cytokines. *Infection and Immunity*, 76(5), pp.2164–2168.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.

Segre, J.A., 2006. Epidermal barrier formation and recovery in skin disorders. *Journal of Clinical Investigation*, 116(5), pp.1150–1158.

Sellman, B.R. et al., 2008. Expression of Staphylococcus epidermidis SdrG increases following exposure to an in vivo environment. *Infection and Immunity*, 76(7), pp.2950–2957.

Sen, A.K. et al., 1999. Post-translational modification of nisin. The involvement of NisB in the dehydration process. *European journal of biochemistry / FEBS*, 261(2), pp.524–532.

Senyurek, I. et al., 2009. Dermcidin-Derived Peptides Show a Different Mode of Action than the Cathelicidin LL-37 against Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 53(6), pp.2499–2509.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature biotechnology*, 26(10), pp.1135–1145.

Sidhu, M.S. et al., 2002. Frequency of disinfectant resistance genes and genetic linkage with beta-lactamase transposon Tn552 among clinical staphylococci. *Antimicrobial Agents and Chemotherapy*, 46(9), pp.2797–2803.

Siegers, K., Heinzmann, S. & Entian, K.D., 1996. Biosynthesis of lantibiotic nisin. Posttranslational modification of its prepeptide occurs at a multimeric membrane-associated lanthionine synthetase complex. *Journal of Biological Chemistry*, 271(21), pp.12294–12301.

Siguier, P. et al., 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, 34(Database issue), pp.D32–6.

Simanski, M. et al., 2010. RNase 7 Protects Healthy Skin from Staphylococcus aureusColonization. *Journal of Investigative Dermatology*, 130(12), pp.2836–2838.

Simor, A.E., 2011. Staphylococcal decolonisation: an effective strategy for prevention of infection? *The Lancet Infectious Diseases*, 11(12), pp.952–962.

Simpson, A.E., Skurray, R.A. & Firth, N., 2000. An IS257-derived hybrid promoter directs transcription of a tetA(K) tetracycline resistance gene in the Staphylococcus aureus chromosomal mec region. *Journal of Bacteriology*, 182(12), pp.3345–3352.

Slettemeås, J.S., Mikalsen, J. & Sunde, M., 2010. Further diversity of the Staphylococcus intermedius group and heterogeneity in the MboI restriction site used for Staphylococcus pseudintermedius species identification. *Journal of Veterinary Diagnostic Investigation*, 22(5), pp.756–759.

Slootweg, J.C. et al., 2014. Semi-synthesis of biologically active nisin hybrids composed of the native lanthionine ABC-fragment and a cross-stapled synthetic DE-fragment. *Bioorganic & medicinal chemistry*, 22(19), pp.5345–5353.

Smith, L.M. et al., 1986. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071), pp.674–679.

Soares, S.C. et al., 2013. The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. *PLoS ONE*, 8(1), pp.e53818–e53818.

Somerville, G.A. & Proctor, R.A., 2009. At the crossroads of bacterial metabolism and virulence factor synthesis in Staphylococci. *Microbiology and molecular biology reviews : MMBR*, 73(2), pp.233–248.

Song, Y. et al., 2013. Additional routes to Staphylococcus aureus daptomycin resistance as revealed by comparative genome sequencing, transcriptional profiling, and phenotypic studies. *PLoS ONE*, 8(3), p.e58469.

Söding, J., Biegert, A. & Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(Web Server issue), pp.W244–8.

Spoor, L.E. et al., 2013. Livestock origin for a human pandemic clone of community-associated methicillin-resistant Staphylococcus aureus. *mBio*, 4(4).

Stauff, D.L., Torres, V.J. & Skaar, E.P., 2007. Signaling and DNA-binding activities of the Staphylococcus aureus HssR-HssS two-component system required for heme sensing. *Journal of Biological Chemistry*, 282(36), pp.26111–26121.

Steinig, E.J. et al., 2015. Single-molecule sequencing reveals the molecular basis of multidrug-resistance in ST772 methicillin-resistant Staphylococcus aureus. *BMC Genomics*, 16(1), p.388.

Stepanovic, S. et al., 2003. Isolation of members of the Staphylococcus sciuri group from urine and their relationship to urinary tract infections. *Journal of Clinical Microbiology*, 41(11), pp.5262–5264.

Steward, C.D. et al., 2005. Testing for induction of clindamycin resistance in erythromycin-resistant isolates of Staphylococcus aureus 2. *Journal of Clinical Microbiology*, 43(4), pp.1716–1721.

Stock, A.M., Robinson, V.L. & Goudreau, P.N., 2000. Two-component signal transduction. *Annual review of biochemistry*, 69, pp.183–215.

Strasheim, W. et al., 2013. Molecular markers of resistance in coagulase-negative staphylococci implicated in catheter-related bloodstream infections. *mortality*.

Stull, J.W. et al., 2014. Staphylococcus delphini and methicillin-resistant S. pseudintermedius in horses, Canada. *Emerging Infectious Diseases*, 20(3), pp.485–487.

Suzuki, H. et al., 2012. Comparative genomic analysis of the genus Staphylococcus including Staphylococcus aureus and its newly described sister species Staphylococcus simiae. *BMC Genomics*, 13, p.38.

Takahashi, T., Satoh, I. & Kikuchi, N., 1999. Phylogenetic relationships of 38 taxa of the genus Staphylococcus based on 16S rRNA gene sequence analysis. *International Journal of Systematic Bacteriology*, 49 Pt 2, pp.725–728.

Takala, T.M. & Saris, P.E.J., 2006. C terminus of NisI provides specificity to nisin. *Microbiology (Reading, England)*, 152(Pt 12), pp.3543–3549.

Takala, T.M. et al., 2004. Lipid-free NisI: interaction with nisin and contribution to nisin immunity via secretion. *FEMS Microbiology Letters*, 237(1), pp.171–177.

Takeuchi, F. et al., 2005. Whole-Genome Sequencing of Staphylococcus haemolyticus Uncovers the Extreme Plasticity of Its Genome and the Evolution of Human-Colonizing Staphylococcal Species. *Journal of Bacteriology*, 187(21), pp.7292–7308.

Tao, L. et al., 2005. Novel carotenoid oxidase involved in biosynthesis of 4,4'-diapolycopene dialdehyde. *Applied and Environmental Microbiology*, 71(6), pp.3294–3301.

Tart, R.C. & van de Rijn, I., 1993. Identification of the surface component of Streptococcus defectivus that mediates extracellular matrix adherence. *Infection and Immunity*, 61(12), pp.4994–5000.

Tennent, J.M. et al., 1989. Physical and biochemical characterization of the qacA gene encoding antiseptic and disinfectant resistance in Staphylococcus aureus. *Journal of general microbiology*, 135(1), pp.1–10.

Tenover, F.C. & Goering, R.V., 2009. Methicillin-resistant Staphylococcus aureus strain USA300: origin and epidemiology. *Journal of Antimicrobial Chemotherapy*, 64(3), pp.441–446.

Tettelin, H. et al., 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5), pp.472–477.

Thammavongsa, V. et al., 2009. Staphylococcus aureus synthesizes adenosine to escape host immune responses. *Journal of Experimental Medicine*, 206(11), pp.2417–2427.

Thammavongsa, V., Schneewind, O. & Missiakas, D.M., 2011. Enzymatic properties of Staphylococcus aureus adenosine synthase (AdsA). *BMC Biochemistry*, 12, pp.56–56.

Thiel, M., Caldwell, C.C. & Sitkovsky, M.V., 2003. The critical role of adenosine A2A receptors in downregulation of inflammation and immunity in the pathogenesis of infectious diseases. *Microbes and Infection*, 5(6), pp.515–526.

Thomas, C.M. & Nielsen, K.M., 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9), pp.711–721.

Thomas, V.C. et al., 2013. A dysfunctional tricarboxylic acid cycle enhances fitness of Staphylococcus epidermidis during β-lactam stress. *mBio*, 4(4).

Thurlow, L.R. et al., 2013. Functional modularity of the arginine catabolic mobile element contributes to the success of USA300 methicillin-resistant Staphylococcus aureus. *Cell Host and Microbe*, 13(1), pp.100–107.

Thurlow, L.R., Joshi, G.S. & Richardson, A.R., 2012. Virulence strategies of the dominant USA300 lineage of community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA). *FEMS Immunology & Medical Microbiology*, 65(1), pp.5–22.

Tomida, S. et al., 2013. Pan-genome and comparative genome analyses of propionibacterium acnes reveal its genomic diversity in the healthy and diseased human skin microbiome. *mBio*, 4(3), pp.e00003–e00013.

Tong, S.Y.C. et al., 2015. Novel staphylococcal species that form part of a Staphylococcus aureus-related complex: the non-pigmented Staphylococcus argenteus sp. nov. and the non-human primate-associated Staphylococcus schweitzeri sp. nov. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 65(Pt 1), pp.15–22.

Toprak, E. et al., 2012. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature genetics*, 44(1), pp.101–105.

Tormo, M.A. et al., 2005. Bap-dependent biofilm formation by pathogenic species of Staphylococcus: evidence of horizontal gene transfer? *Microbiology (Reading, England)*, 151(Pt 7), pp.2465–2475.

Torres, V.J. et al., 2007. A Staphylococcus aureus regulatory system that responds to host heme and modulates virulence. *Cell Host and Microbe*, 1(2), pp.109–119.

Touchon, M. & Rocha, E.P.C., 2007. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution*, 24(4), pp.969–981.

Truong-Bolduc, Q.C. et al., 2013. Native Efflux Pumps Contribute Resistance to Antimicrobials of Skin and the Ability of Staphylococcus aureus to Colonize Skin. *The Journal of Infectious Diseases*.

Truong-Bolduc, Q.C., Strahilevitz, J. & Hooper, D.C., 2006. NorC, a new efflux pump regulated by MgrA of Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 50(3), pp.1104–1107.

Tsai, M. et al., 2011. Staphylococcus aureus requires cardiolipin for survival under conditions of high salinity. *BMC Microbiology*, 11(1), p.13.

Turcatti, G. et al., 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids*

Research, 36(4), pp.e25–e25.

Turnbaugh, P.J. et al., 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122), pp.1027–1031.

Turner, K.M.E. & Feil, E.J., 2007. The secret life of the multilocus sequence type. *International Journal of Antimicrobial Agents*, 29(2), pp.129–135.

Ubukata, K., Itoh-Yamashita, N. & Konno, M., 1989. Cloning and expression of the norA gene for fluoroquinolone resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 33(9), pp.1535–1539.

Uehara, Y. et al., 2000. Bacterial interference among nasal inhabitants: eradication of Staphylococcus aureus from nasal cavities by artificial implantation of Corynebacterium sp. *The Journal of hospital infection*, 44(2), pp.127–133.

Utaida, S. et al., 2003. Genome-wide transcriptional profiling of the response of Staphylococcus aureus to cell-wall-active antibiotics reveals a cell-wall-stress stimulon. *Microbiology (Reading, England)*, 149(Pt 10), pp.2719–2732.

Vacheethasanee, K. et al., 1998. Bacterial surface properties of clinically isolated Staphylococcus epidermidis strains determine adhesion on polyethylene. *Journal of Biomedical Materials Research Part A*, 42(3), pp.425–432.

van de Kamp, M. et al., 1995. Elucidation of the primary structure of the lantibiotic epilancin K7 from Staphylococcus epidermidis K7. Cloning and characterisation of the epilancin-K7-encoding gene and NMR analysis of mature epilancin K7. *European journal of biochemistry / FEBS*, 230(2), pp.587–600.

Varaldo, P.E. & Kilpper-Bälz, R., 1988. Staphylococcus delphini sp. nov., a coagulase-positive species isolated from dolphins. *International ….*

Velamakanni, S. et al., 2008. Multidrug transport by the ABC transporter Sav1866 from Staphylococcus aureus. *Biochemistry*, 47(35), pp.9300–9308.

Verbist, L., 1990. The antimicrobial activity of fusidic acid. *Journal of Antimicrobial Chemotherapy*.

Viana, D. et al., 2015. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nature genetics*, 47(4), pp.361–366.

Viana, D. et al., 2010. Adaptation of Staphylococcus aureus to ruminant and equine hosts involves SaPI-carried variants of von Willebrand factor-binding protein. *Molecular Microbiology*, 77(6), pp.1583–1594.

Voyich, J.M. et al., 2005. Insights into mechanisms used by Staphylococcus aureus to avoid destruction by human neutrophils. *Journal of immunology (Baltimore, Md. : 1950)*, 175(6), pp.3907–3919.

Voyich, J.M. et al., 2006. Is Panton-Valentine leukocidin the major virulence determinant in community-associated methicillin-resistant Staphylococcus aureus disease? *The Journal of Infectious Diseases*, 194(12), pp.1761–1770.

Walkenhorst, W.F. et al., 2013. pH Dependence of microbe sterilization by cationic antimicrobial peptides. *Antimicrobial Agents and Chemotherapy*, 57(7), pp.3312–3320.

Walters, W.A., Xu, Z. & Knight, R., 2014. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Letters*, 588(22), pp.4223–4233.

Wang, R. et al., 2007. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. *Nature Medicine*, 13(12), pp.1510–1514.

Wang, Y. et al., 2014. Staphylococcus epidermidis in the human skin microbiome mediates fermentation to inhibit the growth of Propionibacterium acnes: implications of probiotics in acne vulgaris. *Applied Microbiology and Biotechnology*, 98(1), pp.411–424.

Weidenmaier, C. et al., 2008. Differential roles of sortase-anchored surface proteins and wall teichoic acid in Staphylococcus aureus nasal colonization. *International journal of medical microbiology : IJMM*, 298(5-6), pp.505–513.

Weidenmaier, C. et al., 2004. Role of teichoic acids in Staphylococcus aureus nasal colonization, a major risk factor in nosocomial infections. *Nature Medicine*, 10(3), pp.243–245.

Weinstein, M.P. et al., 1998. Clinical importance of identifying coagulase-negative staphylococci isolated from blood cultures: evaluation of MicroScan Rapid and Dried Overnight Gram-Positive panels versus a conventional reference method. *Journal of Clinical Microbiology*, 36(7), pp.2089–2092.

Wiedemann, I., Benz, R. & Sahl, H.-G., 2004. Lipid II-mediated pore formation by the peptide antibiotic nisin: a black lipid membrane study. *Journal of Bacteriology*, 186(10), pp.3259–3261.

Wilaipun, P. et al., 2008. Identification of the nukacin KQU-131, a new type-A(II) lantibiotic produced by Staphylococcus hominis KQU-131 isolated from Thai fermented fish product (Pla-ra). *Bioscience, biotechnology, and biochemistry*, 72(8), pp.2232–2235.

Wilson-Stanford, S. et al., 2009. Oxidation of lanthionines renders the lantibiotic nisin inactive. *Applied and Environmental Microbiology*, 75(5), pp.1381–1387.

Winstel, V. et al., 2014. Biosynthesis of the unique wall teichoic acid of Staphylococcus aureus lineage ST395. *mBio*, 5(2), p.e00869.

Wu, S. et al., 2011. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics*, 12, p.444.

Wu, S.W. & De Lencastre, H., 1999. Mrp--a new auxiliary gene essential for optimal expression of methicillin resistance in Staphylococcus aureus. *Microbial drug resistance (Larchmont, N.Y.)*, 5(1), pp.9–18.

Yamamoto, H., Uchiyama, S. & Sekiguchi, J., 1996. Cloning and sequencing of a 40.6 kb segment in the 73 degrees-76 degrees region of the Bacillus subtilis chromosome containing genes for trehalose metabolism and acetoin utilization. *Microbiology (Reading, England)*, 142 (Pt 11), pp.3057–3065.

Yang, D. et al., 1999. Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. *Science*, 286(5439), pp.525–528.

Yao, Y. et al., 2005. Factors characterizing Staphylococcus epidermidis invasiveness determined by comparative genomics. *Infection and Immunity*, 73(3), pp.1856–1860.

Ye, J. et al., 2005. Crystal structure of the Staphylococcus aureus pI258 CadC Cd(II)/Pb(II)/Zn(II)-responsive repressor. *Journal of Bacteriology*, 187(12), pp.4214–4221.

Yoshida, H. et al., 1990. Nucleotide sequence and characterization of the Staphylococcus aureus norA gene, which confers resistance to quinolones. *Journal of Bacteriology*, 172(12), pp.6942–6949.

Yoshida, Y. et al., 2011. Bacitracin sensing and resistance in Staphylococcus aureus. *FEMS*

Microbiology Letters, 320(1), pp.33–39.

Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

Zhang, G. et al., 1999. Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. *Cell*, 98(6), pp.811–824.

Zhang, L. et al., 2013. Multilocus sequence typing and further genetic characterization of the enigmatic pathogen, Staphylococcus hominis. *PLoS ONE*, 8(6), p.e66496.

Zhang, Y. & Sievert, S.M., 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. *Frontiers in microbiology*, 5, pp.1–13.

Zheng, T. et al., 2011. The atopic march: progression from atopic dermatitis to allergic rhinitis and asthma. *Allergy*.

Zhou, Y. et al., 2011. PHAST: a fast phage search tool. *Nucleic Acids Research*, 39(Web Server issue), pp.W347–W352.

Zhou, Y.N. et al., 2013. Isolation and characterization of RNA polymerase rpoB mutations that alter transcription slippage during elongation in Escherichia coli. *The Journal of biological chemistry*, 288(4), pp.2700–2710.

Zhu, L. et al., 2010. Structure and regulation of the gab gene cluster, involved in the gamma-aminobutyric acid shunt, are controlled by a sigma54 factor in Bacillus thuringiensis. *Journal of Bacteriology*, 192(1), pp.346–355.

Ziebuhr, W. et al., 1999. A novel mechanism of phase variation of virulence in Staphylococcus epidermidis: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. *Molecular Microbiology*, 32(2), pp.345–356.

Łeski, T.A. & Tomasz, A., 2005. Role of penicillin-binding protein 2 (PBP2) in the antibiotic susceptibility and cell wall cross-linking of Staphylococcus aureus: evidence for the cooperative functioning of PBP2, PBP4, and PBP2A. *Journal of Bacteriology*, 187(5), pp.1815–1824.