Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Nora Berg Blomseth

Molecular characterization and whole genome epidemiology of *Staphylococcus aureus* strains

Master's thesis in Biotechnology
Supervisor: Anuradha Ravi
Co-supervisor: Christina Gabrielsen Ås, Jan Egil Afset
May 2023

NTNU
Norwegian University of
Science and Technology

Nora Berg Blomseth

Molecular characterization and whole genome epidemiology of *Staphylococcus aureus* strains

Master's thesis in Biotechnology
Supervisor: Anuradha Ravi
Co-supervisor: Christina Gabrielsen Ås, Jan Egil Afset
May 2023

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Sammendrag

Staphylococcus aureus er både en kommensal bakterie som finnes på huden og i nesehulen, og en patogen bakterie kan forårsake en rekke alvorlige infeksjoner. Det er den nest vanligste årsaken til blodstrømsinfeksjon (BSI), som kan føre til sepsis, et fryktet syndrom på grunn av den høye dødeligheten. Formålet med studien er å utføre komparative genomanalyser av *S. aureus* stammer isolert fra blodstrømsinfeksjoner, bærerstammer av *S. aureus* og methicillin-resistante *S. aureus* bærerstammer. Hundre og sekstito *S. aureus* stammer fra Tromsøundersøkelsen (TSSS), 62 fra Nasjonalt referanselaboratorium MRSA og 63 fra sepsisregisteret i Helse Nord-Trøndelag samlet inn i perioden 2007-2008 er inkludert i studien. Sekvensene med 300 basepar produsert med Illumina MiSeq teknologi i MRSA kohorten prodeserte bedre rekonstrurte genomer (assembly) og færre kontigs enn sekvensene med 150 basepar produsert med Illumina HiSeq teknologi i bærerstamme- og BSI kohortene. Sekvensspesifikk analyse av *sdrC* genet indikerer et ufullstendig rekonstruert genom (assembly) på grunn av repetisjonsrike regioner i genet. Dette er en viktig faktor å vurdere ved valg av sekvensmetode og i studier om utbredelsen av gener. Kohortene hadde stort sett de samme klonalkompleksene (CC). MRSA kohorten hadde færrest klonalkomplekser, mens kohorten med bærerstammene hadde flest. De største klonalkompleksene (CC) observert var i bærerstamme kohorten CC30, CC45 og CC15, i BSI kohorten CC45, CC1 og CC15 og i MRSA kohorten CC30, CC45, CC8, CC5 og CC22. Tilstedeværelsen av virulensgener og resistensgener i kohortene ble identifisert. Alle MRSA positive stammer bar *mecA* genet som er relatert til meticillinresistens. Ingen av de andre kohortene hadde meticillinresistens. Frekvensen av de analyserte virulensgenene var likere mellom stammer i samme klonalkompleks eller med samme sekvenstype enn mellom stammer fra samme kohort. De samme virulensgenene funnet i BSI og MRSA stammene ble også funnet i bærerstammene. Genene assosiert med unnvikelse av immunsystemet var til stede i de fleste av stammene i alle kohortene. Tilstedeværelsen av gener involvert i koagulering og aggregering var generelt lavere i bærerstammene. En mikrobiell genomvid assosiasjonsstudie viste at *clfB* i CC45 var assosiert med stammer som har forårsaket BSI. Det ble ikke funnet noen significante forskjeller mellom stammer som har forårsaket BSI, bærerstammene og MRSA stammene som kan indikere hvorfor noen stammer forårsaker BSI mens andre forblir bærerstammer. Studien identifiserte at stammer fra ulike kohorter inneholder de samme virulensgenene, og andre faktorer som ikke-annoterte virulensgener, andre ikke-essensielle gener og interaksjoner mellom vert og patogen kan være mulige forklaringer på BSI-stammer sin patogenitet.

Abstract

Staphylococcus aureus is both a commensal and a pathogen that can reside on skin or nasal cavity as a commensal or can cause multiple serious infections. It is the second most common cause for bloodstream infections (BSIs). BSI can lead to sepsis, a feared syndrome due to its high mortality rate. The aim of study for this master thesis is comparative genomics of *S. aureus* strains isolated from bloodstream infections, carriage *S. aureus* strains and methicillin-resistant *S. aureus* strains. One hundred and sixty two *S. aureus* strains from The Tromsø Staph and Skin Study (TSSS), 62 from the National reference laboratory for MRSA and 63 from the Nord-Trøndelag Hospital Trust (HNT) Sepsis registry collected in the period 2007-2008 was included in the study. The sequencing reads with 300 base pairs obtained with Illumina MiSeq technology for the MRSA cohort produced much better assemblies and lesser contigs compared to assemblies from 150 base pairs reads obtained with Illumina HiSeq technology in the carriage and BSI cohorts. Sequence-specific analysis of the *sdrC* gene in the carriage strains indicated incomplete assemblies due to repeat-rich regions. This is an important factor to consider when choosing sequencing techniques and studying the prevalence of genes. Most clonal complexes were shared between the cohorts. While the MRSA cohort had the lowest diversity of clonal complexes, carriage strains cohort had the highest. The major CCs detected in carriage S. aureus were CC30, CC45 and CC15, in the BSI strains, CC45, CC1 and CC15 and in the MRSA strains CC30, CC45, CC8, CC5 and CC22. Virulence genes and resistance genes present in the strains were also identified. All the MRSA positive strains carried the *mecA* gene in relation to methicillin resistance. None of the other cohorts contained methicillin resistance. The prevalence of the analysed virulence genes was more similar between strains in the same CCs and STs rather than from the same cohort. The same virulence genes found in BSI-causing strains and MRSA was also found in the carriage strains. The genes associated with immune system evasion were present in most of the strains from all three cohorts. The prevalence of the genes involved in coagulation and aggregation was generally lower for the carriage strains. Microbial Genome-Wide Association Study (GWAS) showed clumping factor B (*clfB*) in CC45 was associated with BSI-causing strains. It was not found any significant differences between the BSI-causing strains and the carriage and MRSA strains that could indicate why some strains cause BSI while others remain carriage strains. Overall, the thesis identified that strains from different cohorts share similar virulence genes and other factors such as unannotated virulence genes, other accessory genes and host-pathogen interactions could be possible explanations for the BSI strains pathogenicity.

Preface

This master thesis was carried out from fall 2022 to spring 2023 as a final part of the Master of Science degree in Biotechnology at the Department of Biotechnology and Food Science at the Norwegian University of Science and Technology (NTNU).

I would like to thank my fantastic supervisors Anuradha Ravi, Christina Gabrielsen Ås and Jan Egil Afset. Anu, for guiding me through this year of master thesis work, all the meetings, all the questions answered and the multiple read-throughs and comments on my drafts. Christina, for helping and guiding me in the lab, answering any question I had and for commenting on and reading through my drafts. Jan Egil, for helping me with practicalities and feedback on the thesis.

Nora Berg Blomseth
Trondheim, May 2023

Contents

List of Figures

List of Tables

Abbreviations

BSI = Bloodstream infection
PBP = Penicillin-binding protein
bp = base pair
GC = Guanine-cytosine
MGE = Mobile genetic element
SNP = Single nucleotide polymorphism
MSCRAMM = Microbial Surface Components Recognizing Adhesive Matrix Molecule
fnb = Fibronectin-binding protein
clf = Clumping factor
sdr = Serine-aspartate repeat protein
hla = α-toxin
hlg = Hemolysin
pvl = Panton-Valentine leucosidin
chp = Chemotaxis inhibitory protein
scn = Staphylococcal complement inhibitor
aur = Aureolysin
spa = Staphylococcal protein A
coa = Coagulase
vWbp = Von Willebrand factor binding protein
MRSA = Methicillin-Resistant Staphylococcus Aureus
MSSA = Methicillin-Sensitive Staphylococcus Aureus
HA-MRSA = Healthcare-associated MRSA
CA-MRSA = Community-associated MRSA
LA-MRSA = Livestock MRSA
SCC*mec* = Staphylococcal cassette chromosome *mec*
MLST = Multi-Locus Sequence Typing
arcC = Carbamate kinase
aroE = Shikimate dehydrogenase
glp = Glycerol kinase
gmk = Guanylate kinase
pta = Phosphate acetyltransferase
tpi = Triosephosphate isomerase
yqiL = Acetyl coenzyme A acetyltransferase
ST = Sequence type
CC = Clonal complex
NGS = Next generation sequencing
PCR = polymerase chain reaction
SBS = Sequencing by synthesis
GWAS = Genome-Wide Association Study
CNVs = Copy number variations
SI = Sequence inversions
TSSS = The Tromsø Staph and Skin Study
HNT = The Nord-Trøndelag Hospital Trust
iTOL = Interactive tree of life
MEGA = Molecular Evolutionary Genetics Analysis

Part I

Introduction

1 Bloodstream infection and sepsis

Bloodstream infections (BSIs) are described as infections in the bloodstream where a viable bacterium or fungus associated with infection is present, and where the presence of these microorganisms have lead to an inflammatory response[1,2]. The origin of the BSI can often be unknown[3,4]. There are several categories of infection acquirement related to BSIs. They can be categorized as either community- or hospital-onset, where the community-onset BSIs are contracted outside of hospital while hospital-onset BSIs are contracted in the hospital. Community-onset BSI can again be categorized as either community- or healthcare-acquired, where the BSI is categorized as healthcare-acquired when the patient recently have been significantly exposed to a healthcare setting. A community-acquired BSI is not associated with healthcare settings[1]. BSI is an infection type with high mortality worldwide, with an incidence rate of community-onset BSI of 40 to 154/100.000 population every year[1,5].

Sepsis can be a response to BSI and is according to the Sepsis 3 definition from 2016 described as a clinical syndrome where there is an infection and a dysregulated host response and acute organ dysfuntion[6]. There are two stages of sepsis, classified as sepsis without and with septic shock, where the latter is characterized by severe drop in blood pressure (hypotension) and comes together with multiorgan dysfunction, which gives a higher possibility of death compared to sepsis without shock[7,4]. Not all BSIs lead to sepsis, but BSI is detectable in about 20-30% of sepsis cases[8,9]. Studies aimed to acquire more information about BSI is important despite not all cases leading to sepsis, as effective therapies against BSI could decrease the incidence of sepsis caused by a BSI. Sepsis is one of the leading causes of death in the world, as it is estimated to account for 20% of all deaths worldwide[10]. Focus on acquiring more knowledge about the syndrome is therefore very relevant, but there are challenges associated with the study of sepsis and the development of therapies. One of them is that sepsis is not defined or divided into groups based on the underlying cause, which can make it more difficult to find suited treatments[3]. The identification of sepsis in patients can also be difficult as the criteria for sepsis is mostly based on responses in the host, which can also be observed in patients with other complications[10,11].

2 *Staphylococcus aureus*

Staphylococcus aureus belongs to the genus *Staphylococcus* as part of the phylum Firmicutes[12,13], and was first described as *Staphylococcus* in 1880 by the Sottish surgeon Sir Alexander Ogston[13]. Its spherical shape and tendency to cluster gave *Staphylococcus* its name, with *staphyle* being greek for "bunch of grapes" and *kokkos* meaning "berry". A few years later, in 1884, physician Friedrich Julius Rosenbach differentiated *S. aureus* from *S. albus* (now *S. epidermis*) based on their colours, as *S. aureus* appears yellow, while *S. albus* is white[13]. *Aureus* is derived from the latin word for gold, *aurum*, while *albus* is latin for white[14]. Biochemically *S. aureus* can be distinguished from other *Staphylococci*

due to its coagulase positivity[12,13].

S. aureus is a Gram positive bacterium referring to what type of cell wall it has. The term "Gram positive/negative" comes from the technique Gram-staining, where bacteria in three steps are stained, and due to differences in the cell wall, leaving the positive and negative bacteria differently coloured. The major difference between Gram-positive and -negative bacteria is the number of layers and thickness of the cell wall. Gram-negative cell wall consists of two or more layers, while the Gram-positive has one, which is usually thicker than the individual layers of the Gram-negative bacteria. The cell wall mainly consists of strands of peptidoglycan, which consists of *N-acetylglucosamine* and *N-acetylmuramic acid* in an alternating repeated pattern. The strands are connected to each other by a peptide cross-link, which in *S. aureus* consists of five glycines. The formation of cross-links is dependent on penicillin-binding proteins (PBPs), which are as the name implies enzymes that binds to, and is inhibited by, the antibiotic penicillin. If the formation of the cross-links is inhibited, the cell wall will get weaker and burst[12]. An illustration of the cell wall structure is shown in figure 2.1.

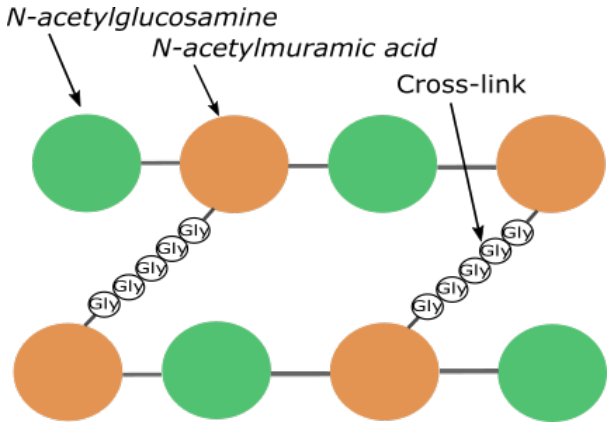


Figure 2.1: Illustration of the cell wall structure of *Staphylococcus aureus*. Strands consisting of *N-acetylglucosamine* and *N-acetylmuramic acid* repeated in an alternating pattern are connected to each other through cross-links of glycines. The figure was illustrated for this thesis.

S. aureus is a pathogen that can cause a number of serious diseases and infections, such as endocarditis, skin and soft tissue infections, osteomyelitis and BSI leading to sepsis[15]. Skin and soft tissue infections are the most frequent kind of *S. aureus* infection, and *S. aureus* is the most common pathogen isolated from surgical site infections and cutaneous abscesses[16,17]. BSI leading to sepsis is the most feared consequence of *S. aureus* infection due to the high mortality rates of sepsis[16]. Worldwide, 20-50 per 100 000 develops *S. aureus* BSI each year, making *S. aureus* the second most common pathogenic cause for BSI, after *Escherichia coli*[18,19,20]. The mortality rate of *S. aureus* BSI before the introduction of antibiotics was 75%-80%. This has decreased a lot, and the rate appears to have stabilized at around 20%[20], which means 2-10 per 100 000 dies from *S. aureus* BSI every year. There is still need for more research and development of treatments against *S. aureus* BSI with the evolution of multidrug resistant infections.

S. aureus is also a commensal bacterium, meaning it is present in a host's microbiota without causing disease. Around 12-30% of the population are persistent carriers, and has a carrier-index of 0.8-1.0[21]. The carrier index is calculated by dividing the number of positive swabs on the number of total swabs taken from an individual[22,23]. 30% are

intermittent carriers, and has a carrier-index of 0.5-0.8[21,23]. Carriage *S. aureus* is usually found in the upper respiratory tract and on the skin. The bacterium transmits by direct contact with a carrier or an infected individual[24,13]. The detection of *S. aureus* infection in a patient will lead to a search of the source of infection. A frequent source of infection in hospitals are vascular catheters, so these are always removed even if the infection is suspected to have another source[25,26]. An antibiotic treatment will also be started[25]. It is important to act fast after detection of *S. aureus* infection to prevent it from developing into sepsis.

2.1 Genome

The genome of *S. aureus* has a size of 2.7-2.8 million base pairs (bp) with a guanine-cytosine (GC) content of around 33%, and it primarily consist of a circular chromosome[27,28]. The genome can be divided into core and accessory genome. The core genome of *S. aureus* consists of approximately 1300-2000 genes[29,30,31,32]. The combination of core and accessory genome present in all strains, or a collection of strains, of a particular species is called the pangenome[12]. The core genome consists of conserved genetic material, meaning that it is conserved in different clonal lineages. Essential genes needed for the survival of the bacteria are usually in the core genome[16]. Accessory genome is genetic material that can differ between lineages or strains. Non-essential genes can be part of the accessory genome, and it is often located on extrachromosomal genetic elements like plasmids, prophages and pathogenicity islands[16,27]. Prophages are DNA from bacteriophages, while plasmids are DNA molecules that can replicate their own DNA independently from the chromosome. The plasmids are usually smaller than the chromsome and circular, but linear plasmids also exist[12]. A pathogeneicity island is a type of genomic island, which are clusters of non-essential genes located on the bacterial chromosome[12,33]. These elements are also called mobile genetic elements (MGEs). They all contain DNA and can move intra- or intercellularly[34]. Intercellular transfer is the transfer of MGEs to other bacterial cells. This happens through horizontal gene transfer, which is divided into three different mechanisms called transformation, conjugation and transduction. Transformation is the uptake of free DNA, while conjugation is the direct transfer from one cell to another, meaning that cell-cell contact is required. Transduction is the transfer of DNA via bacteriophages[34,12].

2.1.1 Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are single base positions in the genome that differ between two organisms. SNPs can occur due to transition, transversion, deletion or insertion[35]. Transition is the change from one purine to another (A↔G), or one pyrimidine to another (T↔C). Purines have a two-ring structure and are more similar to each other than to pyrimidines with a one-ring structure. Transversion is the change from a purine to a pyrimidine, or visa versa[36]. The change in a single base position could lead to a change in amino acid composition, and then also the structure of the protein product. SNPs can, in a microbiological perspective, be used to investigate differences in resistance and virulence genes between isolates, and look at the epidemiological evolution or outbreak of pathogenic bacteria[35].

2.2 Virulence

Virulence factors are molecules such as proteins, enzymes or toxins that contribute to a bacteria's pathogenicity by invasion of the host, causing disease and evasion of the host's immune response[37]. Virulence factors can be found on MGEs like plasmids and bacteriophages, however a large portion are found on pathogenicity islands[12,33]. Multiple virulence factors can be detected in *S. aureus*, and they contribute to the pathogenicity of the bacteria in different ways. During BSI, and potentially sepsis, certain *S. aureus* virulence factors are more relevant than others are. The most relevant ones are listed in table 2.1[26,38,39].

Table 2.1: *Staphylococcus aureus* virulence factors (genes) important in bloodstream infection, with focus on immune system evasion, coagulation and aggregation and barrier breaching.

Function	**Gene**	**Name**
Immune system evasion	*pvl(lukF/S-PV)*	Panton-Valentine leucocidin
	hlgA	Gamma-hemolysin chain II precursor
	hlgB/C	Gamma-hemolysin component B/C
	chp	CHIPS/chemotaxis inhibitory protein
	scn	Staphylococcal complement inhibitor
	spa	Staphylococcal protein A
	aur	Aureolysin
Coagulation and aggregation	*coa*	Coagulase
	vWbp	Von Willebrand factor-binding protein
	clfA/B	Clumping factor A/B
	sdrC/D/E	Serine-aspartate repeat protein C/D/E
	fnbpA/B	Fibronectin-bindin protein A/B
Pore-forming	*hla*	α-toxin

Adherence and damage to host cells is essential for the bacteria to cause BSI and potentially spread to tissue and organs. Most virulence factors involved in adhesion belongs to the Microbial Surface Components Recognizing Adhesive Matrix Molecules (MSCRAMM) family, which is a collection of surface proteins[40,41]. Two of these are the fibronectin-binding proteins A and B (*fnbpA/B*), which are essential in adherence as they bind to epithelial cells, as well as endothelial cells in the blood vessels[41]. Other members of the MSCRAMM family are clumping factors A and B (*clfA/B*) and serine-aspartate repeat proteins C, D and E (*sdrC*, *sdrD* and *sdrE*)[38,40]. The MSCRAMM genes *sdr*, *fnbp* and *clf* all have very similar main structure, as illustrated in figure 2.2. At the N- and C-terminal there is a signal sequence and a sorting sequence respectively. They all have A domains that binds to ligands, but only the *sdr* genes have two to five repeats of a B domain after the A domain. The *clf* and *sdr* genes then have a region consisting of serine-aspartate repeats, while the repeat region of the *fnbp* gene consists of fibronectin-binding repeats[40,42,41].

For the spread of *S. aureus* to other host cells after adhesion, α-toxin (*hla*) is central as it is pore-forming and can lyse epithelial, endothelial and immune cells[26]. In BSI specifically, *hla* can increase the chance of spreading through damaging of the endothelial barrier[38]. It is not enough for *S. aureus* to adhere and spread to other host cells to cause BSI. It also must survive in the bloodstream, and evasion of the hosts immune cells is crucial for that. Multiple virulence factors are involved in the survival process by inhibiting and destroying the components of the immune system. This involves targeting of leukocytes by hemolysin (*hlg*) and the leukocidin Panton-Valentine leucosidin (*pvl*) composed

of the two components *lukF-PV* and *lukS-PV* [26,38,43]. It also involves the inhibition of the complement pathway with chemotaxis inhibitory protein (*chp*), staphylococcal complement inhibitor (*scn*) and aureolysin (*aur*). Staphylococcal protein A (*spa*) will bind to antibodies to restrict the function of the immune cell [26,38].

Figure 2.2: The general structure of the Microbial Surface Components Recognizing Adhesive Matrix Molecules (MSCRAMM) family genes serine-aspartate repeat protein (*sdr*), clumping factor (*clf*) and fibronectin-binding protein (*fnb*). They all have signal sequences (S and S*) on the N- and C-terminals, and an A domain (A). The *sdr* genes additionally have two to five repeats of a B domain (B). The repeat region consists for the *sdr* and *clf* genes of serine-aspartate repeats, while the *fnbp* gene has fibronection-binding repeats. After the repeat domain is a cell wall-spanning domain (M). The illustration is inspired by [40,42,41].

When *S. aureus* is able to survive in the blood, virulence factors coagulase (*coa*) and von Willebrand factor binding protein (*vWbp*) causes fibrin to clot, which will promote staphylococcal aggregation by adhering to the surface proteins *fnbpA/B*, *clfA/B* and *sdrC/D/E* [26,38]. The clotting of fibrin and *S. aureus* promotes its survival in the blood while also being a central part of sepsis as it blocks blood flow and therefore decreases the oxygen supply to the organs possibly leading to organ failure [44,39]. The virulence genes and their involvement in BSI is illustrated in figure 2.3.

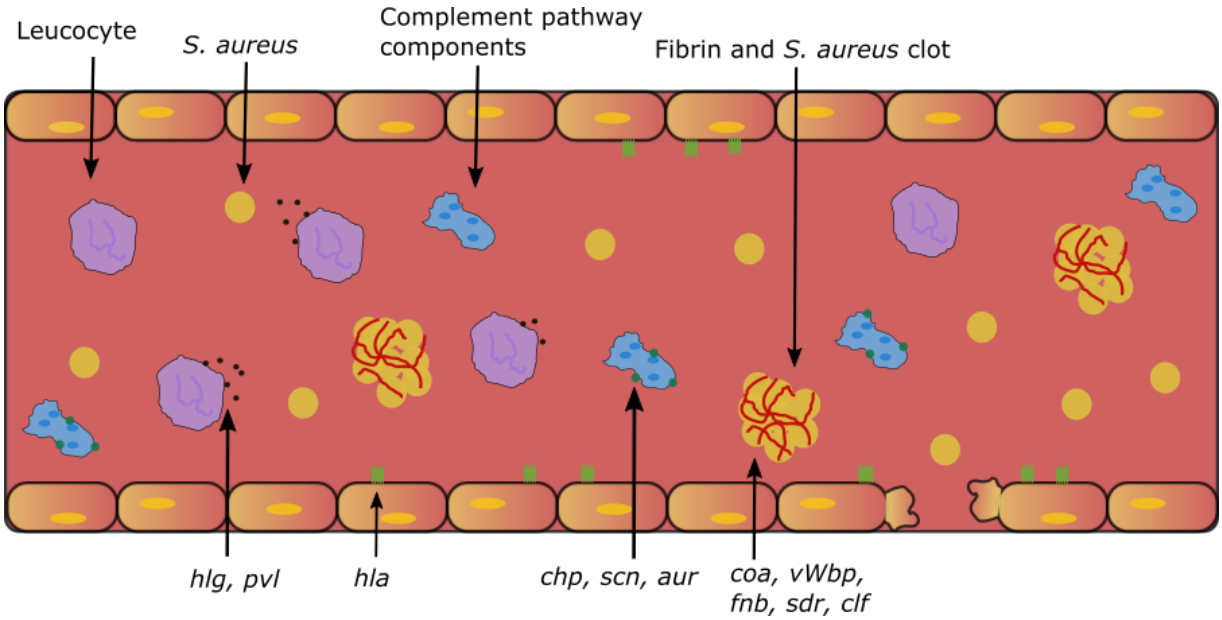


Figure 2.3: An illustration of *Staphylococcus aureus* virulence genes and their involvement in bloodstream infection. Hemolysin (*hlg*), Panton-Valentine leucosidin (*pvl*), chemotaxis inhibitory protein (*chp*), staphylococcal complement inhibitor (*scn*) and aureolysin (*aur*) are all involved in immune system evasion by inhibiting and destroying components of the immune system. Coagulase (*coa*), von Willebrand factor binding protein (*vWbp*), fibronectin-binding protein (*fnbp*), clumping factor (*clf*) and serine-aspartate repeat protein (*sdr*) are involved in coagulation and aggregation of fibrin and *S. aureus*, which blocks blood flow. The pore forming gene α-toxin (*hla*) can lyse epithelial, endothelial and immune cells. The illustration is inspired by [26,38,39].

2.3 Antibiotic resistance

Antibiotic resistance genes are genes encoding molecules that either protect the bacterial cell from or directly target antibiotics, and can be located on the chromosome or on MGEs. The resistance genes are often obtained through horizontal gene transfer, but can also occur due to mutations in the genome, such as SNPs[12,45,24]. Antibiotic resistance in *S. aureus* was first seen in the 1940s, not long after the introduction of penicillin in the hospitals[24]. The resistant strains produced the enzyme penicillinase, which hydrolyses the β-lactam ring in penicillin. This obstructs the penicillins ability to target the PBPs involved in cell-wall formation, which means that it no longer has antimicrobial activity[24].

Methicillin-resistance in *S. aureus* was first described in 1961, and the strains were given the name methicillin-resistant *Staphylococcus aureus* (MRSA). It was first observed as healthcare-associated MRSA (HA-MRSA), but was later observed as community-associated (CA-MRSA) and in livestock (LA-MRSA)[46]. The resistance to methicillin, as well as most other β-lactam antibiotics, is due to the *mecA* gene, which is found on a MGE called staphylococcal cassette chromosome *mec* (SCC*mec*)[46]. Due to horizontal gene transfer, the SCC*mec* has spread to a variety of genetically different *S. aureus* clones[46]. *MecA* encodes an alternative to the PBPs, which are present in methicillin-sensitive *S. aureus* (MSSA). The alternative protein encoded by *mecA* is not recognized by most β-lactams, making it resistant to β-lactam based antibiotics[12]. Resistance to other kinds of antibiotics have also been observed in MRSA on multiple occasions, making it a feared pathogen as it limits the treatment options[47].

An outbreak of MRSA in the 1970s lead to increased use of the antibiotic vancomycin, which as methicillin and penicillin targets the cell wall synthesis. It binds to the terminal of a percursor of peptidoglycan, which will inhibit it from being incorporated into the growing cell wall[24,48]. Resistance against vancomycin was first detected in 2002, where the terminus of peptidoglycan was altered so that the binding of vancomycin was decreased[48]. Increased antibiotic resistance leading to limited treatment options is one of the challenges with antibiotic resistance. Failed treatment will increase the risk of death or long term illness, that again can increase the chance of other individuals being affected by the bacterium[49]. If a bacterium develops resistance towards the main form of antibiotic treatment, an alternative has to be used, which is often more expensive[49].

2.4 Characterization of *S. aureus* genome

2.4.1 Multi-Locus Sequence Typing

Multi-Locus Sequence Typing (MLST) is a method used to characterize bacterial species[50]. The characterization is based on six to seven essential housekeeping genes, as their gene products are involved in processes required for an organism's survival and therefore conserved within the species[51]. The housekeeping genes used to characterize *S. aureus* isolates are carbamate kinase (*arcC*), shikimate dehydrogenase (*aroE*), glycerol kinase (*glp*), guanylate kinase (*gmk*), phosphate acetyltransferase (*pta*), triosephosphate isomerase (*tpi*) and acetyl coenzyme A acetyltransferase (*yqiL*)[52]. Each housekeeping gene has different alleles based on an internal fragment of approximately 450 bp. If the nucleotide sequence of that fragment differ between two isolates, that specific housekeeping gene will give the isolates different alleles. Each unique allele is given a number, and all seven alleles combined defines a unique allelic profile called sequence type (ST)[50]. This is illustrated in figure 2.4.

2.4.2 Clonal Complex

STs can be further categorized into different clonal complexes (CC). CCs are groups containing one predominant ST as well as genotypes that are common close relatives to the predominant ST[53]. What CC a ST belongs to is dependent on its evolutionary relationship to the predominant ST in a particular CC. The emergence of a CC begins with the increase of a genotype's frequency in the population. Over time, the genotype will diversify and eventually a number of genotypes will together with the founding ST constitute a CC. As the founding genotype diversifies, the allelic profile of the new variant will eventually change as well. One of the housekeeping genes will change by either point mutation or recombination, resulting in a new genotype. This new genotype will eventually diversify again, which will lead to the change of another housekeeping gene and so on[53].

2.4.3 *Spa* Typing

Spa Typing is another way to characterize *S. aureus*. The method is based on differences in the Xr region of the protein A gene (*spa*)[54]. Protein A is a surface and cell wall anchored protein of *S. aureus*, and the Xr region on the protein is located between the cell surface and a ligand binding region[54,41,16]. The DNA sequence encoding the Xr region is composed of small repeats that varies between 3 and 15. Each base composition of each repeat is assigned a unique number. The numbered repeats combined gives a unique *spa* type, as illustrated in figure 2.4. Variation in number and composition of small repeats gives different *spa* types[54]. *Spa* Typing can as MLST be used to investigate long-term genetic changes and analyse the phylogenetic relationship between strains, but it can also be used to analyse outbreaks with a much shorter time-span. Genetic variation over a long period of time can be termed as macrovariation, while variation over a short period of time can be termed as microvariation[55].

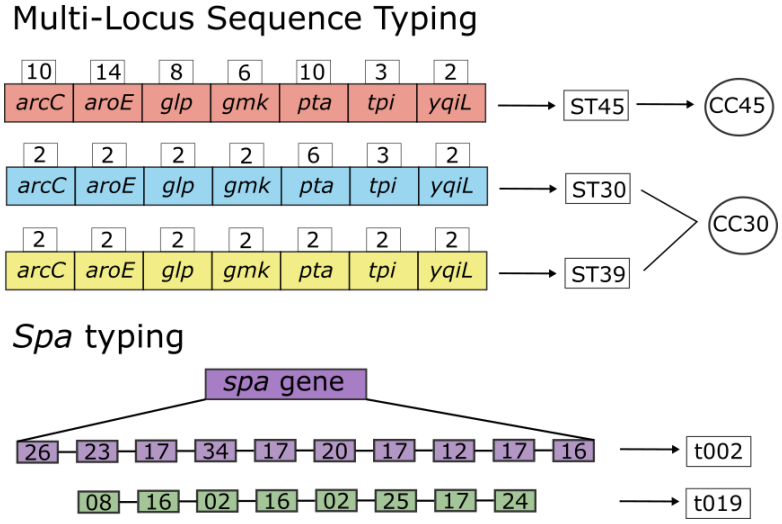


Figure 2.4: During Multi-Locus Sequence Typing of *Staphylococcus aureus*, housekeeping genes are assigned a specific number based on the allele. The combination of alleles defines a unique sequence type. The sequence types can be further associated with a clonal complex. *Spa* typing is based on the Xr region of the protein A gene (*spa*) gene, which is composed of small repeats with varying numbers between 3 and 15. Each unique repeat is assigned a number, and the repeats combined creates a unique *spa* type.

3 Microbial genomics

Microbial genomics is the study of microorganisms' entire genomes, which involves characterization of components such as virulence factors and resistance genes. Genome sequencing makes this possible and was first performed on prokaryotic genome in 1995. Since then, sequencing methods have evolved so that multiple organisms can be sequenced at once in a matter of hours. Computational genomics is also necessary to analyse the sequenced DNA, and has together with sequencing made it possible to map the diversity of microbes and their components[56].

Microbial genomics is used in clinical and hospital settings for identification of pathogens to give the correct treatment, as well as mapping and monitoring of outbreaks. Identification of pathogens is traditionally done by biochemical tests on growth culture, detection of specific biomarkers of the pathogen, like antibodies and antigens, or detection of specific nucleic acid sequences with polymerase chain reaction (PCR). With the emergence of microbial genomics, the entire genome of the pathogens is available for analysis giving potentially important information about all components of the pathogen. This can help with providing the correct diagnosis, for instance the cause of infection, as well as the best treatment options, like what antibiotics could be the most effective. Another advantage is the possibility of surveillance of outbreaks and the discovery of potential mutations, as well as discovery of new targets for vaccines or other kinds of treatment[57]. Traditional methods are still cheaper than NGS, but microbial genomics has a lot of potential in health care and the cost of sequencing is decreasing. NGS and computational genomics gives information beyond the identification of the pathogen, making it possible to analyse and acquire more knowledge about components, such as virulence factors, and the effect they could have on the pathogenicity of the bacterium and the host it is infecting[57].

3.1 Next Generation Sequencing: Illumina

Next generation sequencing (NGS) is a term used to describe various sequencing technologies developed to sequence whole or parts of genomes effectively. The techniques use different approaches to sequence DNA, however they all determine the nucleotides of millions of small DNA fragments at once, which is why NGS is also referred to as massively parallel sequencing. After sequencing, the short DNA sequences about the size of 150bp to 300bp are referred to as "reads", which can be assembled into longer sequences called contigs[58].

Illumina sequencing is a broadly used NGS method developed in the beginning of the 2000's, and consists of three main steps as well as analysis of the sequence data. Illumina has developed multiple machines with slightly different performances. Miseq is a sequencing machine that can perform paired-end sequencing where the maximum read length is 2x300 bp, while the HiSeq sequencer can perform paired-end sequencing but with a read length of 2X150 bp[59,60].

Library preparation is the first step performed, where extracted and purified DNA is randomly fragmented and adapters are added to the end of the fragmented sequences, as shown in figure 3.1. The adapters differentiate between the 3' and 5' end on the sequences and contain indexes referring to a specific sample so that multiple samples can be sequenced at once. They also contain sequencing binding sites, which are regions that can bind to complementary sequences. The sequences are then amplified with PCR[59,60].

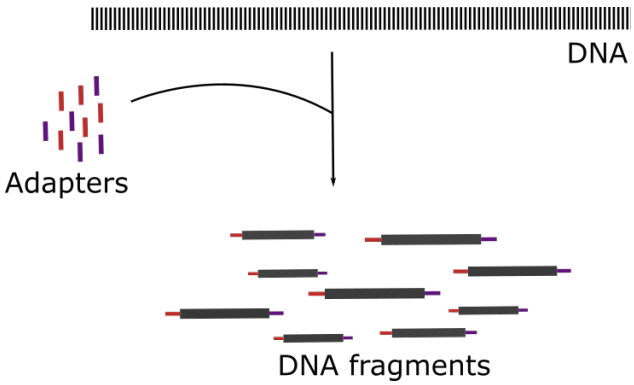


Figure 3.1: Illustration of library preparation, where purified DNA is fragmented and adapters are added. The illustration was made for this thesis.

The next main step is called cluster generation, which is illustrated in figure 3.2. The library prepared samples are loaded to a flow cell, which is a glass slide coated with two kinds of oligonucleotides (short strands of DNA) complementary to the adapters attached to each end of the single stranded DNA fragments, enabling them to bind to the flow cell. Only one of the adapters attached to the DNA fragments binds to an oligo on the flow cell at this point. Polymerase will produce a complementary reverse strand originating from the oligo bound to the flow cell before the template strand is removed. The DNA fragments now attached to the flow cell are then amplified with bridge amplification. The adapter sequences on the non-attached end of the DNA fragments will bind to a complementary oligonucleotide, resulting in the fragment being attached in both ends, creating a "bridge". Polymerase will produce complementary strands, giving both forward and reverse strands, and the process repeats multiple times producing clusters of strands[59,60].

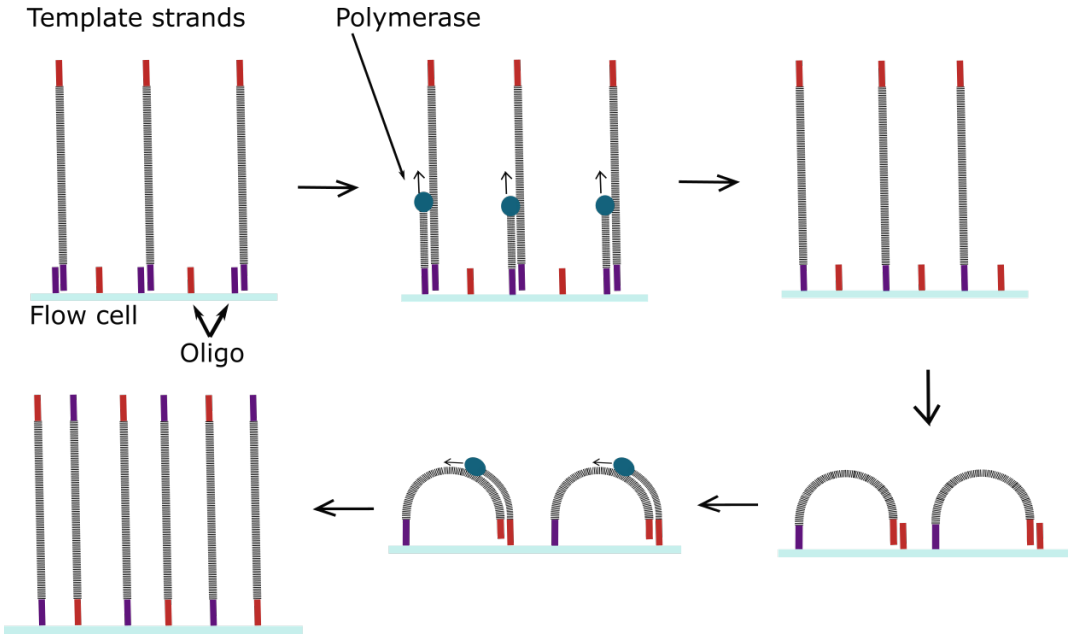


Figure 3.2: Illustration of cluster generation during Illumina sequencing. The adapter sequences attached to the template strands binds to oligos attached to the flow cell. Polymerase produces a complementary strand before the template strand is removed. The unattached end of the complementary strands then binds to another oligo on the flow cell, and polymerase produces another complementary strand. This process is repeated multiple times. The illustration was made for this thesis.

The last step is sequencing, where each base in the sequences are read with sequencing by synthesis (SBS). Reverse strands are removed, and the forward strands are sequenced first. A primer is attached to the adapter before nucleotides, each with a specific fluorescent tag, are added to the flow cell. The first nucleotide attaches to the primer, making the start of the read. Only one nucleotide is attached to each forward strand at a time, and an image is taken of the fluorescent colour emitted from each cluster. The process is repeated until the end of the fragment resulting in a read complementary to the forward strand. For paired-end sequencing, the reverse strands are synthesized again and sequenced after the removal of the forward strands[59,60].

3.2 Assembly

Assembly of reads obtained from DNA sequencing is an important step in the analysis of genomes, as the reads will be connected and form longer sequences called contigs. *De novo* assembly is the process of connecting the reads without any knowledge about the organism they stem from. Meaning that the method does not use a reference to reconstruct the genome[61]. Different approaches can be used to assemble reads, and one of them is *de Bruijn* graphs. The principle behind *de Bruijn* graphs is that reads are divided into k-mers. When k-mers in the beginning or end of a read overlaps with k-mers of another read, these are connected[62].

Different challenges can occur during assembly, with one of them being sequencing errors. During sequencing, single bases can be replaced with another one (substitution), removed (deletion) or added to a sequence (insertion). This can make the assembly process harder and lead to complex *de Bruijn* graphs[61]. Another challenge is repeats in the sequenced genome, which are stretches of DNA with repeats of two or more bases[63,64]. Longer repeat sequences can be hard to assemble especially if they are longer than the short-reads ($<$300 bp) produced with Illumina technology[64]. If the repeat section is longer than the read, it is also longer than the k-mer length used by the assembling tool, and it will have difficulties connecting the reads properly. Multiple genes with similar structure and repeat regions makes it even harder to assemble them correctly, and could lead to local assemble collapse where the tool thinks there is one gene instead of multiple very similar ones[64].

Tandem repeats can not only lead to errors in the assembly, but also cause challenges during annotation. Previously assembled and annotated genomes are used as reference during annotation, and the collection of genes that vary a lot in repeat sequences could be limited. A gene could then not be annotated as the repeat sequence is too different from the reference database[64].

3.3 Sequence alignment

Sequence alignment is when the sequences of two or more organisms are compared[12]. The comparison of two sequences is called pairwise alignment, while multiple alignment is the comparison of multiple sequences. During the alignment, gaps will be added to the sequences so that similar stretches of DNA in each organism will align vertically. Either parts of, or the entire genome of organisms can be aligned[12]. The sequences, or genes, compared are homologous if they have the same common ancestor. These genes can be either orthologs or paralogs. Orthologous genes originate from the same gene in a common ancestor, and have the same function. Paralogous genes does also have the

same common ancestor, but they have evolved differently so that they no longer have the same function[12]. Sequence alignment can be used to compare a sequence of a particular organism to reference sequences from that same organism. The similarity to the reference sequence can be presented by query coverage and percent identity. The query coverage is the percentage of the sequence (query) that aligns with the reference sequence. The percent identity is the percentage of bases in the aligned part of the query sequence that is identical to the reference[65].

3.4 Phylogenetic analyses

The evolutionary relationship between organisms within the same or different species can be presented with a phylogenetic tree. Availability of DNA sequences makes it possible to look at evolution at a molecular level[66]. There are however limitations associated with the attempt of reconstructing the evolutionary history of a species or sample collection. One of them is convergent evolution[67], which is when similar traits have been independently developed in organisms that are not closely related[68]. This is a challenge as in the reconstruction of the relationship between organisms it is assumed that organisms with the same traits are related[67]. Recombination, the exchange of genetic material between two DNA molecules, is another challenge as the parts of the new sequence could have different evolutionary histories[69]. Methods used to circumvent these kinds of limitations are the use of core gene alignment and robustness estimation. Bootstrapping is an example of robustness estimation method and it investigates if small changes in the gene alignment results in the same clades[70].

Before constructing a phylogenetic tree, DNA sequences obtained from the organisms must be aligned to maximize the similarity between the sequences[12]. When comparing the entire genome of organisms, core gene alignment can be used. Core gene alignment is an alignment of orthologous genes present in all samples of the collection of organisms[12]. Phylogenetic trees consists of terminals, nodes and branches. Terminals represent the organisms used to construct the phylogenetic tree. Nodes show previous ancestors dividing into two new linages, while branches connects the nodes as well as the terminals. The branch length represents the amount of genetic change between nodes and terminals. Longer branches indicate more genetic change than for the shorter branches[71,12,72]. Phylogenetic trees can be built using different methods, which can be divided into two main groups, distance-based methods and character-based methods. When the construction of the phylogenetic tree is based on the differences between the sequences, it is distance-based. In a character-based phylogenetic tree, each column of characters in the multiple alignment is considered one at the time[73,72]. The evolutionary change and relationship between the sequences in the tree is described by a nucleotide substitution model. It will estimate the amount of nucleotide substitutions that have occurred in two sequences since they shared a common ancestor[74].

3.4.1 Maximum likelihood

Maximum likelihood is a character-based method, which estimates unknown parameters of a specific model so that the probability of the observed data is maximized[75]. In phylogeny, the tree is considered the model and the observed data is the multiple alignment of DNA sequences of different samples, and the goal is to estimate certain parameters so that the phylogenetic tree best represents the evolutionary relationship between the samples. Parameters that can be estimated are branch lengths, substitution rates, frequency of the

different nucleotides and tree topology, which is how the samples are placed in relation to each other. Each column in the multiple alignment of the sequences of length n is considered separately. The parameters are estimated for each sequence position in the alignment by maximizing the likelihood function L(S|M), where S is the sequences and M is the model, which is the phylogenetic tree. The likelihood for the entire alignment is the product of the likelihood of each sequence position (i) as shown below [76];

$$L(S|M) = \prod_{i=1}^{n} L(S_i|M)$$

3.5 Microbial GWAS

Genome-Wide Association Study (GWAS) is a method used to find genetic variants that are statistically associated with a specific trait [77]. It was first developed to find genetic variants that could be disease causing factors in humans. Human GWAS is primarily based on SNPs as the genetic variant in the genome, and the method looks at the correlation between a specific trait and the SNPs in a population [78].

Human GWAS was later adapted to use on microbial DNA to detect possible genetic factors that could affect the host during microbial diseases, or explain different phenotypes of pathogens. The method is referred to as microbial GWAS (mGWAS) [79]. An adaption from GWAS to mGWAS was necessary because of differences between the human and microbial genome, in the sense that microbial genome has more forms of genetic variation [79,80]. The main types of genetic variations in microbial genomes are SNPs, copy number variations (CNVs), sequence inversions (SIs) and the presence and absence of genes. CNVs are stretches of DNA that are present in various numbers of copies in different individuals due to major deletions or duplications, while SIs are reversed stretches of DNA [79,80]. Entire genes can be lost or gained in bacteria and lead to the presence or absence of genes as a genetic trait. One reason is horizontal transfer, where MGEs are transferred to the bacteria from the environment, another cell or by bacteriophages [80,81].

3.5.1 Scoary

Multiple mGWAS performing tools with differences in methods and output have been developed, but further development is still needed to obtain a tool considering all aspects of the microbial genome and its behavior [79,80]. Scoary is a tool that is based on the presence and absence of genes of the pangenome, which are all genes present in a sample set. It will find and present genes that could be associated with a specific chosen trait [80,82]. The statistical association between the genes and the trait is presented with a p-value, which is the likelihood of the results to be by chance. A p-value of 0.05 implies that the presence of a specific gene for a specific phenotype has less than 5% chance of being a random result.

When multiple independent t-tests are performed to get a p-value, a multiple testing problem will occur. For a single t-test with a cut off value of 0.05, there will be an error rate of that same value. More tests increases the probability of getting an error, making it more likely that the result is by chance. The probability of getting at least one error with an error rate α for n independent tests is [83];

$$1 - (1 - \alpha)^n$$

Bonferroni corrections or the Benjamini-Hochberg procedure can be used to account for the increased probability of error in multiple testing. The Benjamini-Hochberg procedure will account for the multiple testing problem by limiting the number of false positives, meaning non-significant cases that are reported as significant [84]. Bonferroni corrections accounts for the multiple testing problem by saying that the p-value needs to be smaller than the error rate α divided by all independent tests n to be significant; [83]

$$p < \frac{\alpha}{n}$$

Sensitivity and specificity also give a measurement of the mGWAS results. Sensitivity when talking about mGWAS is the probability of a gene being correctly identified as present in a specific group of microbes with or without the specific trait, and is calculated as shown below [85].

$$\text{Sensitivity} = \frac{\#\text{ true positive}}{\#\text{ true positive} + \#\text{ false negative}}$$

Specificity is the probability of a gene being correctly identified as not present, and it is calculated as shown below [85];

$$\text{Specificity } = \frac{\#\text{ true negative}}{\#\text{ true negative} + \#\text{ false positive}}$$

Odds ratio is a measurement of the relationship between two cases, like the relationship between a trait and a specific gene. It is calculated by dividing the odds off having the trait and specific gene by the odds of having the trait but not the specific gene. The result would then indicate if the odds is higher of having the trait if the specific gene is present. An odds ratio of more than 1 indicates that the odds of having the trait is higher if the specific gene is present, while an odds ratio of less than one indicates that the odds is lower [86,87].

$$\text{Odds ratio} = \frac{\#\text{gene and trait positive}/\#\text{gene positive, trait negative}}{\#\text{gene negative, trait positive}/\#\text{gene and trait negative}}$$

Part II

Aim of study

The aim of study for this master thesis is comparative genomics of *S. aureus* strains isolated from bloodstream infection, carriage *S. aureus* strains and methicillin-resistant *S. aureus* strains.

The main objectives will be to study the distribution of sequence types, *spa* types and clonal complexes within the different cohorts of *S. aureus* available from the Nord-Trøndelag Hospital Trust Sepsis Registry, Tromsø Staph and Skin Study and the National reference laboratory for MRSA. The secondary objectives will be to sequence the MRSA strains using Illumina technology (MiSeq) and perform quality control on the sequences from all three cohorts to determine the proportion of samples with pure isolates and good quality sequences. The genomic diversity of the strains will be studied by in silico multi-locus sequence typing (MLST), *spa* typing and clustering in clonal complexes of *S. aureus* from different cohorts. Different virulence factors and genes associated with reduced antimicrobial susceptibility will be identified. A phylogenetic reconstruction will be made and associations of the strains between and within the cohorts will be studied, and a microbial Genome-Wide Association Study (mGWAS) will be performed to detect any genes associated with BSI-causing strains.

Part III

Materials and methods

4 Workflow

All materials and methods used in this master thesis is illustrated in the workflow in figure 4.1, and will be described in more detailed in the following sections.

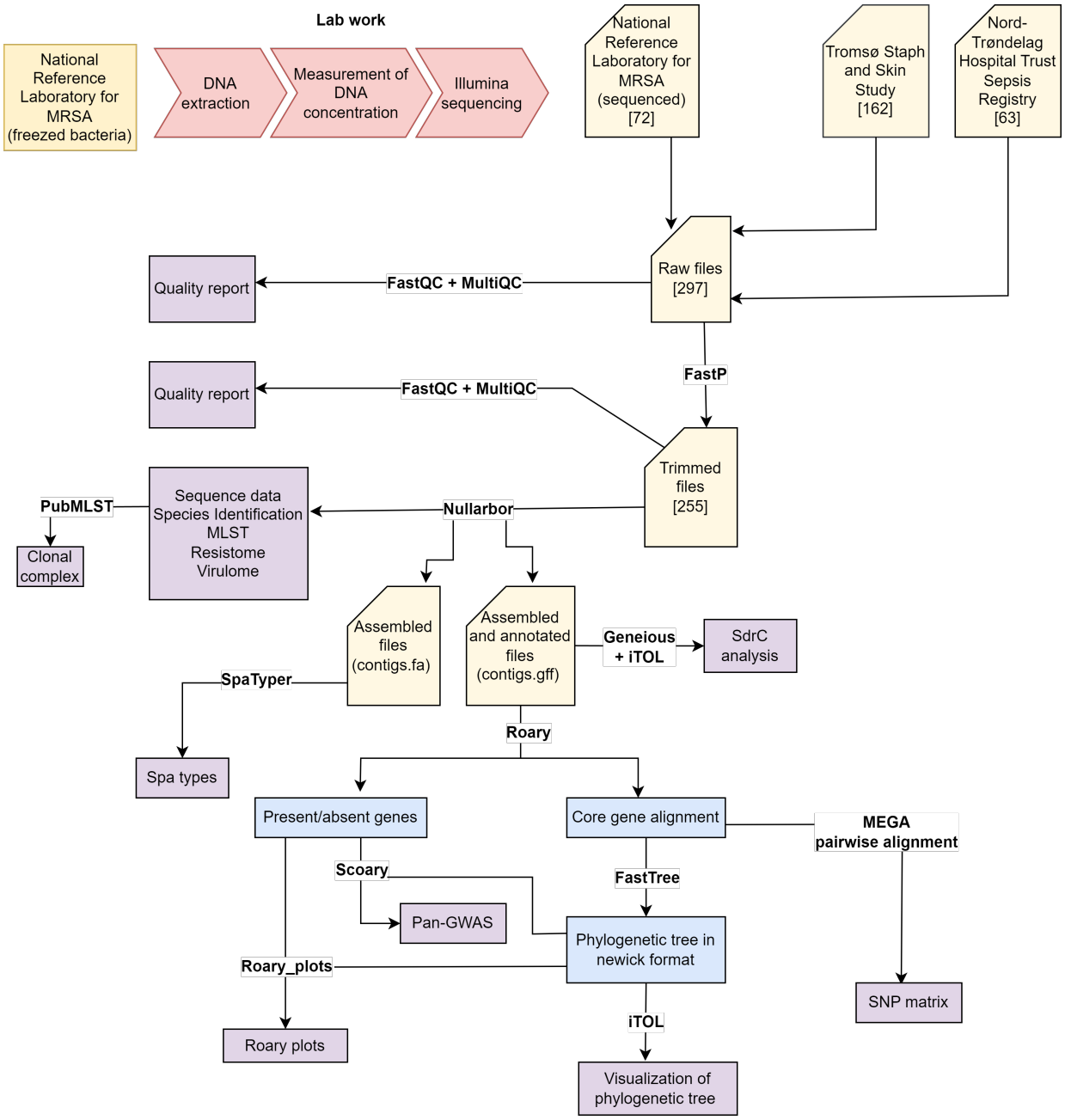


Figure 4.1: Workflow showing the methods, and their output, used in this master thesis. The yellow boxes represent the sequenced strains from the three cohorts TSSS, BSI and MRSA. The numbers in brackets represents the number of strains. The blue coloured boxes represent output used in further analysis to get the results, which are represented by the purple coloured boxes.

5 Strain collection

S. aureus strains from three different cohorts were used to answer the objectives of this thesis. Carriage strains were obtained from The Tromsø Staph and Skin Study (TSSS) [88] and The National Reference Laboratory for MRSA and bloodstream infection (BSI) causing strains were obtained from The Nord-Trøndelag Hospital Trust (HNT) Sepsis registry. The strains from the TSSS is referred to as TSSS strains, the ones from The National Reference Laboratory for MRSA is referred to as MRSA strains and the strains from the HNT sepsis registry is referred to as BSI strains. All strains were collected in Norway during the period 2007-2008.

5.1 Tromsø Staph and Skin Study

TSSS was initiated in 1974 to study and determine causes of the high cardiovascular mortality seen in middle-aged men in Norway at the time, particularly in Northern Norway. Over the years, it developed into a larger study also focusing on other chronic diseases [88]. TSSS consists of seven surveys. The strains used in this master thesis is from the sixth survey, called Tromsø 6, which was conducted in 2007 and 2008. The attendees in Tromsø 6 were between 30 and 87 years old and a total of 12 984 persons participated. Nose and throat swabs were taken from 4026 participants between October 2007 and July 2008 for the detection of *S. aureus*. The participants were mostly between 30 and 49 years old, but there were also some older participants [89]. 162 randomly selected and sequenced strains of detected *S. aureus* were used in this master thesis. The genomes of these strains were previously sequenced with Illumina HiSeq by Genomics Core Facility, NTNU with 2x150 bp read configuration.

5.2 Nord-Trøndelag Hospital Trust Sepsis Registry

The Nord-Trøndelag Hospital Trust Sepsis registry (HNT sepsis registry) is a registry with microbiological and clinical data of patients with bloodstream infection admitted to HNT since the registry was initiated in 1994 at Levanger hospital, Norway. The registry is used to control the quality of treatment of sepsis and the occurrence of antibiotic-resistant bacteria, as well as for research on sepsis and antibiotic-resistant bacteria [90]. In this master thesis, 63 sequenced strains of *S. aureus* that caused BSI during the period 2007-2008 were included. The genomes of these strains were previously sequenced with Illumina HiSeq with 2x150 bp read configuration.

5.3 National Reference Laboratory for MRSA

The national reference laboratory for MRSA was initiated in 2006 at St. Olavs Hospital in Trondheim, Norway. It contains all MRSA positive strains detected in Norway from 1. January 2008, as well as strains from an MRSA strain bank initiated in 2005 [91]. 72 randomly selected carriage strains from the national reference laboratory for MRSA collected in the period 2007-2008 were used in this master thesis. The strains were anonymized before being made available from the MRSA-lab. The DNA of each strain was extracted and sequenced at the Department of Medical Microbiology at St. Olavs hospital as part of this master thesis.

6 DNA extraction and Next Generation Sequencing of MRSA strains

The DNA of the selected MRSA strains from the national reference laboratory at St. Olavs Hospital was extracted and its concentration measured. Next generation sequencing, specifically Illumina MiSeq sequencing, was then performed with 2x300 bp read configuration that was used in further analysis of the strains.

6.1 DNA extraction

The MRSA strains at St. Olavs are stored in pure culture in Greaves solution at -80 °C. The selected MRSA strains were collected from the freezers and each strain was streaked out on a blood agar plate. The plates were incubated at 35 °C over night.

A master mix was prepared and mixed with each strain sample for DNA extraction. The master mix consisted of 200 µL TE-buffer, 20 µL proteinase K (20 mg/mL) and 10 µL lysostaphin (2 mg/mL) per sample. For each sample, bacterial cells were collected with a loop and resuspended in 230 µL master mix in a 2 mL tube. The samples were incubated at 37 °C for 15 minutes, then at 65 °C for 15 minutes. The samples were then cooled down to room temperature before 4 µL of RNAse A (100 mg/mL) was added to each sample tube and then vortexed.

The extraction instrument QIAGEN EZ1 Advanced XL and the EZ1 DNA Tissue kit (QIAGEN) was used for DNA extraction. Elution volume was set to 50 µL, and sample volume to 200 µL. Reagent cartridges, consisting of prefilled reagents from QIAGEN, were placed in the cartridge racks and inverted 10 times to mix the magnetic particles in the reagent. Eluation tubes (1,5 mL), tip holders, filter-tips and sample tubes were loaded as described by the instrument display.

6.2 Measurement of DNA concentration

After extraction of bacterial DNA, concentrations were measured using a Qubit 3.0 Fluorometer (Thermo Fischer). The measurement was prepared by mixing 1 µL of Qubit dsDNA HS Reagent (Thermo Fischer) with 199 µL Qubit dsDNA HS Buffer (Thermo Fischer) per sample, making up the working solution. Two standard tubes were prepared by adding 190 µL of working solution to each tube. 10 µL of Qubit dsDNA HS Standard #1 (Thermo Fischer) was added to one of the tubes, and 10 µL of Qubit dsDNA HS Standard #2 (Thermo Fischer) to the other. The sample tubes and control tube were prepared by adding 198 µL of working solution to each tube. 2 µL of each sample was added to one tube each. 2 µL of Qubit dsDNA HS Standard #2 was added to the control tube. All tubes were vortexed for 2-3 seconds and incubated for 2 minutes at room temperature.

A Qubit 3.0 Fluorometer was then used to measure the concentration of each sample. The analysis program "dsDNA High Sensitivity" was used. The option "Read standards" was chosen and the standard solutions were read. The option "Read sample" was then chosen, "sample volume" was set to 2 µL and the control sample was read. The concentration of all the samples were then measured individually.

6.3 Illumina Sequencing

An Illumina MiSeq instrument was used to sequence the extracted DNA of each sample. A DNA library was prepared by using Illumina DNA prep kit according to the manufacturer´s instructions[92]. Quality control of libraries was done by measuring DNA concentration using Qubit dsDNA HS assay kit. The libraries were pooled, denatured and added into the reagent cassette for 300 bp paired-end sequencing with a MiSeq V3 reagent Kit[93].

7 Bioinformatic analysis

All bioinformatic methods used in this master thesis are described in the following sections. The commands run during the bioinformatic analysis are shown in appendix F.

7.1 Quality control of raw data

Quality control of reads is important to determine whether the sample is eligible for further analysis. It involves control of per read quality, trimming and taxonomic classification of the reads. The quality can be measured and evaluated based on different terms. Phred score is a quality measurement based on the error probability of the bases being correctly called during sequencing, and is calculated as $Q=-10\log_{10}(p)$, where p is the error probability. A higher phred score indicates that the probability of a false base call is lower. If $Q>20$, the accuracy of the base call is more than 99% and the probability of an incorrect base call is less than $1/100$[94].

7.1.1 FastQC, FastP and MultiQC

FastQC[95] (version v0.11.9) was used to quality check the raw reads from the TSSS, BSI and MRSA cohorts. FastQC is a tool providing different analyses of the sequence data giving an impression of the quality of the reads. The raw reads were in fastq format. Each strain from the TSSS, BSI and MRSA cohorts have two reads files, as they were sequenced with paired-end sequencing. FastQC was run separately on reads files for all strains.

FastP[96] (version 0.22.0) was then used to trim the reads to remove the adapters attached during Illumina sequencing. FastQC was then run again for the trimmed sequences. FastP for paired end data was run for all three cohorts separately.The input was the raw reads files, while the output was divided into trimmed reads, unpaired reads and failed reads, as well as a report in json and html format.

MultiQC[97] (version 0.9) was used to combine the FastQC output analyses, and the FastP data. The tool combines analysis results from each strain and assembles them into one HTML report. MultiQC was run with the FastQC output for both raw reads files and both trimmed reads files, for all three cohorts TSSS, BSI and MRSA. This gave a total of 12 reports. The parameter –interactive was added to be able to get interactive plots. The parameter is needed when the strain number is above 100. MultiQC was run with the FastP data for all three cohorts BSI, TSSS and MRSA. The report shows the number of forward and reverse reads before and after trimming, the number of reads which passed the quality filter and the number of reads which did not pass.

7.2 Nullarbor

Nullarbor[98] (version 2.0.20191013) is a pipeline performing a number of operations on Illumina paired-end sequenced data and presenting the results in a report. The operations, described in subsections 7.2.2-7.2.4, include assembly of the reads into contigs and annotation of these, which is used in further analysis. The pipeline also provides taxonomic assignment of reads for species identification, and information about sequence data quality and genotyping. Resistance and virulence genes are also detected, and information about the core and pangenome of the provided strains is given, as well as a phylogenetic tree showing the relationship between them. Nullarbor was used to assemble reads into contigs later used in other pipelines, as well as obtaining information about the strains from the BSI, TSSS and MRSA cohorts.

Nullarbor was run three times, once for each cohort of strains. Input were the reads in fastq format trimmed with FastP, a specification of *S. aureus* as the strains organism and a *S. aureus* reference strain. The reference strain was *Staphylococcus aureu*s subsp. *aureus* strain ATCC 25923 (GenBank code: CP009361.1) obtained from European Nucleotide Archive[99]. Output was assembled and annotated contigs in .gff and .gbk (GenBank) format, as well as a report (one for each cohort) with results from the multiple operations run by Nullarbor.

7.2.1 Assembly and annotation

Each sequenced genome consists of a number of reads that are assembled into contigs and then annotated. Nullarbor uses SKESA[100] (version 2.4.0) to assemble the reads into contigs. SKESA is based on DeBruijn graphs. Prokka[101] (version 1.14.6) was used by Nullarbor to add features to the assembled contigs. Prokka will annotate the contigs by comparing them to a number of databases.

Information about the assembled and annotated genomes are presented in the Nullarbor report. For each strain, number of contigs, number of basepairs, average length of contigs are some of the values presented. N50, which is the length of the shortest read in the sequenced genome that together with other reads make up at least 50% of the entire genome of the strain, is also presented. These values can give an impression of the quality of the assembly.

7.2.2 Identification of species

Nullarbor uses Kraken 2 for identification of species, and Kraken[102] (version 1.1.1), Kraken 2[103] (version 2.1.2) and Centrifuge[104] (version 1.0.4) had to be installed for this purpose. Kraken 2 is a database containing k-mers and the lowest common ancestors of all organisms having that specific k-mer in their genome is reported. A k-mer is a nucleotide sequence consisting of k number of nucleotides. Kraken 2 will classify a strain by looking for specific k-mers in the strain sequence and compare it to the database[102]. Kraken 2 is an improved version of Kraken, using less memory and faster classification[103].

7.2.3 Multi-Locus Sequence Typing

MLST is used to characterize *S. aureus* based on seven housekeeping genes. Nullarbor uses an MLST pipeline[105] that will scan the assembled contigs and determine the sequence

type based on PubMLST schemes. The clonal complexes of each strain was found by using a scheme from PubMLST[106].

7.2.4 Resistome and virulome

Nullarbor uses the pipeline Abricate[107] (version 1.0.1) to find resistance and virulence genes present in each strain. Abricate compares contigs to the databases Resfinder[108] and VFDB[109] to find resistance and virulence genes respectively. Resfinder is a resource for identification of microbial resistance genes, while VFDB is a virulence factor database. Abricate sets identity and coverage percentage to 80% as default.

Among the virulence genes found by Abricate, the ones mentioned frequently in literature describing the pathogenesis of *S. aureus* BSI were selected for further analysis. The genes selected were *aur*, *hlgA/B/C*, *scn*, *chp*, *pvl*, *sdrC/D/E*, *fnbpA/B*, *clfA/B*, *vWbp* and *hla*. These are described in section 2.2 in the introduction.

7.3 Characterization of strains with SpaTyper

SpaTyper[110] (version 0.2.1) was used to characterize the *S. aureus* strains from the three cohorts, in addition to MLST. The characterization is based on differences in *Staphylococcus aureus* protein A (*spa*). Each strain is assigned a specific *spa* type based on what *spa* allele they have. The input for SpaTyper was one file in fasta format from each strain containing contigs obtained from running Nullarbor.

7.4 Pangenome and phylogenetic analysis

Roary[111] (version 3.13.0) is a pangenome pipeline that was run to calculate the pangenome of all the strains from the three cohorts. The input was the annotated genomes from all three cohorts obtained from Nullarbor. Two of the output files from Roary were used for further analysis. One file is the presence/absence of genes in all genomes, while the other is a core gene alignment. The parameters -e and -n was added to create the core gene alignment, which is needed to build a phylogenetic tree.

To utilize the output from Roary in further analysis, a phylogenetic tree in newick format was generated using FastTree[112,113] (version 2.1.11). The method used by FastTree to generate the tree is maximum likelihood. The input was the core gene alignment file generated by Roary. The parameters -nt and -gtr specifies that it is a nucleotide alignment using the GTR+CAT model, and that the tree is constructed based on this substitution model. Roary plots was then used to visualize the core genome calculated by Roary. The presence/absence file from Roary, as well as the phylogenetic tree from FastTree, was used as input. To visualize the phylogenetic tree generated by FastTree, interactive tree of life (iTOL)[114] was used. iTOL is an online tool where one can visualize, annotate and manipulate phylogenetic trees.

7.5 Microbial GWAS

A mGWAS using Scoary[82] (version 1.6.16) was performed to detect if any genes found in the *S. aureus* genomes could be significant to BSI-causing strains (BSI strains). Scoary is a software tool developed to perform genome-wide association studies (GWAS) on bacterial genomes. It is however referred to as pan-GWAS, or mGWAS, to distinguish it from

eukaryotic GWAS. The method is used to detect genetic variants that are statistically associated with a specific trait. The trait in this case was BSI-causing strains, as the aim was to detect genes that could be statistically associated with BSI-causing strains opposed to carriage strains.

The presence/absence file from Roary, the phylogenetic tree generated by FastTree and a file determining the trait to be BSI causing strains were used as input. Scoary was run separately for strains from different CCs, where an input file indicated what CC would be looked at in that specific run. The output is a list of genes that could be associated with the set trait. The association is measured with sensitivity, specificity and odds ratio, and the significance of the association is presented with a p-value.

7.6 Making a SNP matrix with MEGA

Molecular Evolutionary Genetics Analysis (MEGA)[115] was used to make a SNP matrix, which is a matrix presenting the number of SNPs between two strains. MEGA is a software that can perform a variety of operations related to computational molecular evolution. The GUI version of MEGAX was downloaded on a windows computer.

The SNP matrix was made based on the core genome alignment obtained from running Roary. The alignment file was transformed to mega files (.meg), and the pairwise alignment tool was used to make the SNP matrix. Parameters chosen were nucleotide sequences, non protein-coding nucleotide sequence data, pairs of taxa, no variance estimation method, no. of differences, transitions + transversions, same (homogeneous) pattern among lineages and complete deletion.

Strains belonging to different cohorts and with less than 110 SNPs between them were selected in order to identify possible virulence genes unique to the BSI-causing strains or carriage strains. The frequency of *S. aureus* virulence genes relevant in BSI, and detected with Abricate, was compared between the selected TSSS, BSI and MRSA strains. The *sdrC* gene was analysed further due to the difference in frequency between the BSI and TSSS strains.

7.7 Detection of *sdrC* with Geneious

Geneious[116] (version 2022.2.2) is a software containing tools that can be used to analyse genomes, and was used to further analyse the *sdrC* gene by detecting the gene in the genomes from the cohorts TSSS, BSI and MRSA. Contig files obtained with Nullarbor for all genomes in the three cohorts were loaded to Geneious. For each genome, the contigs were concatenated so that each genome consisted of a single coherent sequence. A local database containing four *sdrC* reference sequences was made, and used to run a BLAST on all genomes. The *sdrC* reference sequences were from the reference genomes *Staphylococcus aureus* subsp. *aureus* HO 5096 0412 (RefSeq: NC_017763), *Staphylococcus aureus* subsp. *aureus* str. Newman (RefSeq: NC_009641), *Staphylococcus aureus* subsp. *aureus* N315 (RefSeq: NC_002745) and *Staphylococcus aureus* subsp. *aureus* MW2 (RefSeq: NC_003923). Results was shown as a hit table with a maximum hit of 1. Hits with a sequence length of more than 1500 (average sequence length for the reference sequences is 2871) and pairwise identity above 80%. The *sdrC* sequence from each genome with a BLAST hit was extracted. The extracted *sdrC* sequences were aligned with MAFFT alignment using default settings, and a tree was built with FastTree in Geneious. The

tree was visualized in iTOL.

The frequency of strains with *sdrC* genes detected with Nullarbor and then detected with Geneious was calculated. Fisher's exact test was performed in RStudio[117] (version 2023.03.0+386) to determine if there could be a significant relationship between BSI causing strains and the presence of the *sdrC* gene.

Part IV

Results

8 Quality control

Quality control of reads was done in order to determine what strains were suited for further analysis. The number of reverse and forward reads and base pairs (bp) before and after trimming, number of reads passing the quality filter and number of reads failing to pass the quality filter is shown for all three cohort in appendix B. The average number of bp for both forward and reverse sequences after filtering for each cohort was MRSA=361 Mbp (± 52 Mbp), BSI=1071 Mbp (± 354 Mbp), TSSS=620 Mbp bp (± 192 Mbp). The average number of bp was almost three times higher in the BSI cohort than in the MRSA cohort, and almost twice as high as in the TSSS cohort. The average number of reads for both forward and reverse sequences after filtering for each cohort was MRSA=0.67×10^6 reads (± 0.10×10^6 reads), BSI=3.66×10^6 reads (± 1.12×10^6 reads), TSSS=2.11×10^6 reads (± 0.67×10^6 reads). The BSI cohort had the largest amount of reads with a big difference between strains, as the number of reads is ranging from 1.50-6.1$\times10^6$ reads. In comparison, the MRSA cohort strains had a lot fewer reads ranging from 0.47-0.90$\times10^6$ reads. This is likely due to the strains in the MRSA cohort being sequenced with a MiSeq machine giving read lengths of 300 bp, while the strains in the other two cohorts were sequenced with a HiSeq machine giving 150 bp long reads.

The mean quality scores of the forward and reverse strains before and after trimming with FastP was analysed. Plots are shown in appendix C. For the TSSS and BSI cohorts, the mean quality score for each base position for both the forward and reverse reads were above 30, before a small decrease at the end of the reads. The same trend was seen before and after trimming. For the MRSA cohort, that had 300 bp reads, the quality score started to decrease around base position 230. There was a slight improvement in the quality score for the end positions after trimming for all three cohorts.

A difference in adapter content for the forward and reverse reads before and after trimming was observed for all three cohorts. Plots provided by MultiQC in shown in appendix D. The adapter content increased for each base position in the forward and reverse reads for all three cohorts before trimming. After trimming with FastP, the adapter content had decreased in all three cohorts.

8.1 Criteria for exclusion of strains from further analysis

Quality control of raw and trimmed reads was done with both FastQC and Nullarbor. Strains that equally had contigs>200 and N50<15 000 (both conditions had to be present) was excluded from further analysis. The TSSS cohort consisted of 162 strains, but 40 strains were excluded. In addition, two more strains were excluded from the TSSS cohort as they did not have any of the housekeeping genes of *S. aureus*. Further analysis was based on the remaining 120 TSSS strains, as well as 63 strains from the BSI cohort and 72 strains from the MRSA cohort, where none of the strains met the exclusion criteria.

9 Identification and assembly of *S. aureus*

In order to determine whether the strains actually are *S. aureus*, taxonomy classification was run through the Kraken 2 database within the Nullarbor run. The percentage of identity to *S. aureus* reference genomes from the Kraken 2 database for the strains in the three cohorts is presented with box plots in figure 9.1a. All the 255 strains remaining after quality control had a percentage of identity to *S. aureus* of more than 80%. The strains in the TSSS cohort mainly ranges from approximately 85%-94% identity to *S. aureus*, with an average of 90.32% (±2.49%) and median of 91.07%. One of the strains (Tromso9105) has a noticeably lower percentage of identity, compared to the rest of the strains in that cohort, of 82.52%. In the BSI cohort, the percentage of identity to *S. aureus* for the strains mainly ranges from approximately 84%-90%, with an average of 86.80% (± 1.47%) and median of 86.73%. The strains in the MRSA cohort has percentage of identity ranging from 85%-92%, with an average of 89.06% (± 1.81%) and median of 89.33%. One strain (SO-SAU7-16) has a percentage of identity of 82.22%. The strains with lower percentage of identity have less genome identical to the reference *S. aureus* genomes in the Kraken 2 database, than the rest of the strains. Genes present in both the references and the strains are not necessarily 100% identical, for instance due to mutations. They could also have additional genes not found in the references.

The GC-content was checked for each strain to see if their values were similar to values reported for reference *S. aureus* strains. The GC-content of the strains in each cohort is presented by box plots in figure 9.1b. The TSSS cohort has an average and median GC-content of 34.9% (± 0.7%), with the values ranging from 33.3%-36.7%. The GC-content of strains in the BSI cohort ranges from 34.6%-36.9%, with an average and median of 35.8% (± 0.4%). The strains in the MRSA cohort has little difference in their GC-content as it ranges from 32.5%-33%, with an average and median of 32.8% (±0.1%). The strains in the MRSA cohort has a lower GC-content than the strains in the TSSS and MRSA cohorts. This could be due to longer reads and better contig assembly for the MRSA strains.

Assembly of the reads was necessary to do further analysis of the genomes, and it resulted in various numbers of contigs for each strain, as shown in figure 9.1c. The average number of contigs in each cohort is TSSS=109 (± 83), BSI=78 (± 58) and MRSA=31 (± 14). The median in each cohort is TSSS=76, BSI=60 and MRSA=27. The number of contigs for the strains in the TSSS cohort has the largest difference, as it ranges from 18-343 contigs, however the majority of the strains has a number of contigs between approximately 18 and 135. In the BSI cohort, the majority of the strains in the cohort has a number of contigs ranging from 14-162. The MRSA cohort has the overall lesser numbers of contigs ranging from 13-93.

The N50 value, which is the length of the shortest read out of a group up reads that make up at least 50% of the entire genome of the strain, is shown in figure 9.1d. The average N50 value in each cohort is TSSS=10.3×10^4 (± 10.1×10^4), BSI=12.8×10^4 (± 12.9×10^4) and MRSA=31.4×10^4 (± 14.5×10^4). The average N50 value is similar for the TSSS and BSI cohorts, while the average value is larger for the MRSA cohort. The median N50 value in each cohort is TSSS=7.6×10^4, BSI=8.9×10^4 and MRSA=30.3×10^4. As most of the strains in the MRSA cohort has low numbers of contigs and high values of N50 compared to the other cohorts, it indicates that the assembly of these have been good. The other cohorts, especially TSSS, has more strains with higher numbers of contigs, which indicates

poorer assembly. The more contigs the genome of the strain is distributed on, the harder it can be to detect and annotate genes. The fact that the MRSA strains in average has a lower number of contigs and a higher N50 value also indicates that longer reads such as MiSeq sequencing gives a better basis for assembly than shorter reads such as HiSeq.

The genome size for each strain in each cohort is presented with box-plots in figure 9.1e, where the values are quite similar for the strains in the TSSS and BSI cohorts, while some of the strains in the MRSA cohort has larger genomes. The median number of bp in each cohort is TSSS=2.73 Mbp, BSI=2.73 Mbp and MRSA=2.80 Mbp. The average number of bp in the TSSS cohort is 2.73 Mbp (± 0.046 Mbp), for the BSI cohort the average is 2.73 Mbp (± 0.035 Mbp) and for the MRSA cohort the average is 2.81 Mbp (± 0.057 Mbp). The higher average observed for the MRSA cohort could be due to longer reads (300 bp vs 150 bp) resulting in better assembly. Two of the strains in the MRSA cohort (SO-SAU7-16 and SO-SAU7-9) has higher numbers of bp (3.00 Mbp and 2.98 Mbp) than the others, which are ranging from 2.72-2.91 Mbp. These strains could have more genes than the others, giving them a larger genome. The reference strain used to run Nullarbor has 2.78 Mbp, while other reported numbers of bp in the *S. aureus* genome are 2.7-2.8 Mbp. All three cohorts has an average bp within the previously reported values, which adds to the likelihood of the strains being *S. aureus* and that most of their genomes have been sequenced.

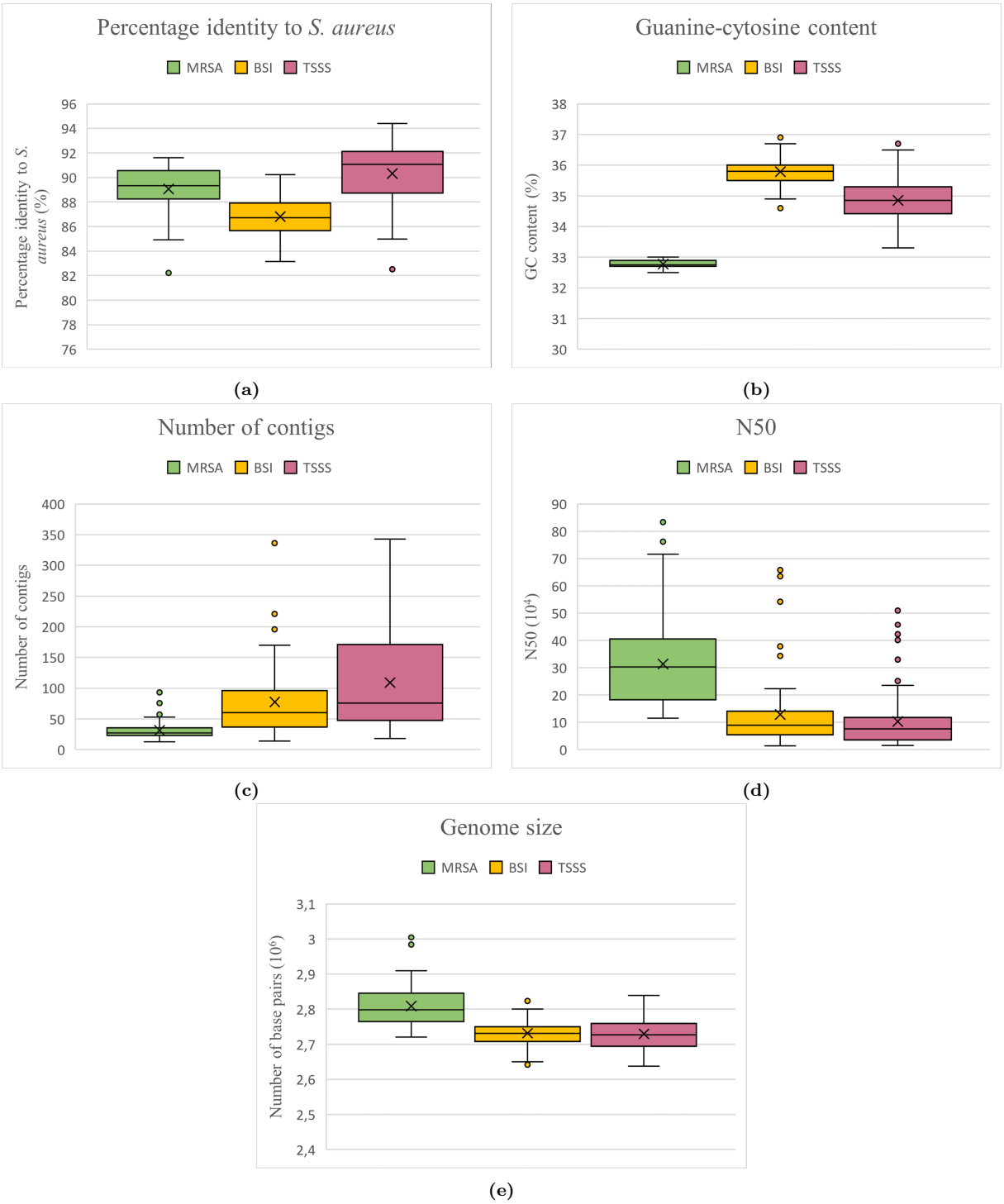


Figure 9.1: Results from Nullarbor presented with box-plots for each cohort TSSS, BSI and MRSA. The median is represented by a line inside the box, the mean is represented by an X and the whiskers represents the minimum and maximum values. Outliers are represented by dots. The plots show; a) The percentage identity to *Staphylococcus aureus*, b) the guanine-cytosine content, c) number of contigs, d) the N50 value and e) the genome size presented as number of base pairs.

10 Characterization of strains

10.1 Sequence types and *spa* types

The distribution of genotypes (STs, CCs and *spa* types) for each cohort BSI, MRSA and TSSS were analysed to compare similarly characterized strains to each other. The consolidated result from the molecular typing of the strains for each cohort are shown in tables 10.1, 10.3 and 10.2. The STs accounting for more than 10% of all the samples in the TSSS cohort are ST30 (27/120) and ST45 (25/120). The STs accounting for more than 10 % of all the samples in the BSI cohort are ST45 (22/63) and ST15 (9/63), and the STs accounting for more than 10% of all the samples in the MRSA cohort are ST5 (11/72), ST8 (10/72), ST30 (9/72) and ST45 (8/72). The largest ST groups when combining all three cohorts are ST45 (55/255) and ST30 (40/255). There is however a difference in the frequency of ST45 and ST30 between the three cohort. The highest frequency of ST45 can be seen in the BSI cohort, where 35% of the strains are ST45. In the TSSS cohort 21% of the strains are ST45, while the frequency is 11% in the MRSA cohort. It is also a difference in the frequency of ST30 between the three cohorts. 23% of the TSSS strains are ST30, while the frequency is 6% (4/63) in the BSI cohort and 13% in the MRSA cohort. A difference in the frequency of ST5 is also seen between the cohorts, as the MRSA cohort has a higher frequency of ST5 than the other two. The frequency of ST5 is 2% (2/120) and 3% (2/63) in the TSSS and BSI cohorts respectively, while it is 15% in the MRSA cohort.

The TSSS cohort has the highest diversity of STs (37), while both the BSI (17) and MRSA (21) cohorts have fewer STs. This could be because the TSSS has more strains than the other cohorts, and/or it could indicate that commensal carriage strains have more diversity than infection-associated and MRSA strains. Six strains from the TSSS cohort, four strains from the BSI cohort and four strains from the MRSA cohort did not have a ST. All the six strains from the TSSS cohort, two from the BSI cohort and three from the MRSA cohort had a novel full length allele of a housekeeping gene similar to the corresponding allele in an existing ST. The remaining one strain from MRSA and two from BSI without ST had a full length novel allele not similar to a corresponding allele in an existing ST. The diversity in *spa* types is greater than the STs. There was detected 80 unique *spa* types, with t015 (7/120) being the most prevalent, and 37 unique STs in the TSSS cohort. In the BSI cohort, 39 unique *spa* types, with t015 (6/63) being the most prevalent, and 17 unique STs was detected. This was similar to the MRSA cohort, where 42 unique *spa* types, with t002 (10/72) being the most prevalent, and 21 unique STs was detected.

Table 10.1: Sequence types (STs) present in the BSI cohort and what clonal complex they belong to. 4 strains do not have a ST. The table also shows *spa* types and in what ST they are observed. "-" are strains with no *spa* type.

Sequence Type (number of strains > 1)	***Spa*-type** (number of strains > 1)	**Clonal Complex**
1 (3) 9 188 (2) 2418	t127(3) t800 t189, t7099 t591	CC1
5 (2)	t002, t6267	CC5
8 (5)	t008(3), t024, t064	CC8
15 (9)	t084, t16383, t2603(2), t346(3), t6121, -	CC15
22	t2183	CC22
30 (4) 39	t021(3), t1414 t275	CC30
45 (22)	t015(6), t026(3), t050, t061, t065(4), t1231, t1248, t230, t2884, t330, t333, t344	CC45
97	t267	CC97
12 25 (2) 50 (2) 130 152	t160 t2471, - t246(1) - t4690	No CC
No ST (4)	t002, t008, t096, t840	No CC

Table 10.2: Sequence types (STs) present in the MRSA strains and what clonal complex they belong to. 4 strains do not have a ST. The table also shows *spa* types and in what ST they are observed.

Sequence Type (number of strains> 1)	***Spa* type** (number of strains > 1)	**Clonal Complex**
1 (3) 772	t127(3) t345	CC1
5 (11) 105 149 (2) 225 (2) 2626	t002(7), t088, t306, t688(2) t002 t002, t4382 t003(2) t002	CC5
8 (10) 72 239 343 1324 (2)	t008(3), t059, t1627, t2384, t304(3), t723 t324 t030 t037 t324(2)	CC8
22 (7) 1326	t020, t032, t032, t15806, t310(2), t718 t223	CC22
30 (9)	t012(2), t019(3), t021(2), t1202, t1434	CC30
45 (8) 1330 497	t015, t026(2), t1081, t3084, t3090, t333, t362 t015 t2015	CC45
80 (2) 88 859 338	t044(2) t690 t325 t437	No CC
No ST (4)	t019(2), t064, t2952	No CC

Table 10.3: Sequence types (STs) present in the TSSS cohort and what clonal complex they belong to. 6 strains do not have a ST. The table also shows *spa* types and in what ST they are observed. "-" are strains with no *spa* type.

Sequence Type (number of strains> 1)	***Spa* type** (number of strains> 1)	**Clonal Complex**
188	t189	CC1
1218	t189	
5 (2)	t548, t2595	CC5
6	t701	
8 (4)	t1476, t008, t197, t024	CC8
15 (9)	t084(5), t144(4), t346, t5232, t605, t360, t416	CC15
1876	t5314	
1882	t605	
22 (2)	t005, t192	CC22
30 (27)	t012(4), t017, t018(2), t019(2), t021(5), t037(2), t1135, t122, t2018, t2303(2), t318, t363, t5250, t5255, t6134, t700	CC30
34 (2)	t4244, t884	
39 (2)	t129, t275	
1879	t318	
1884	t840	
1889	t138	
1890	t021	
45 (25)	t015(6), t026, t050, t065(3), t073, t1248, t1402, t180, t1826, t2045, t230(2), t282, t3219, t5211, t6137, t6149, t630	CC45
47	t2383	
455 (2)	t065(2)	
1877	t116	
1878	t065	
1881	t230	
1891	t026	
3043	t050	
3177	t015	
97	t359	CC97
121	t4390	CC21
1693	t812	
7	t091	No CC
10	t240	
25 (4)	t167, t5242, t5449, t759	
50 (3)	t1269, t246(2)	
59	t216	
101 (3)	t056(3)	
182 (3)	t364(2), t493	
207	t375	
395	t5243	
No ST (6)	-, t1027, t160, t164, t840, t8416	No CC

10.2 Clonal Complex

The frequency of CCs in the three cohorts BSI, TSSS and MRSA is shown in figure 10.1, while the number of strains in each CC is shown in table 10.4. The CCs associated to more than 10% of the strains in the TSSS cohort are CC30 (35/120), CC45 (34/120) and CC15 (13/120). In the BSI cohort, CC45 (22/63), CC1 (7/63) and CC15 (9/63) accounts for more than 10% of all the strains, while in the MRSA cohort it is CC30 (9/72), CC45 (10/73), CC8 (15/73), CC5 (17/73) and CC22 (8/73). The TSSS and BSI cohorts do also have strains belonging to CC8 (3%/8%) and CC5 (2%/3%), but with much lower frequencies. CC15 is considered a major CC in both the TSSS and BSI cohort, but was not associated with any of the strains in the MRSA cohort. The number of STs not associated with a CC in each cohort is TSSS=9, BSI=5 and MRSA=4. When combining the strains from all three cohorts, CC45 is the largest CC group representing 26% (66/255) of all the strains. 85% of the strains in CC45 belongs to the TSSS and BSI cohort, while 15% belongs to the MRSA cohort.

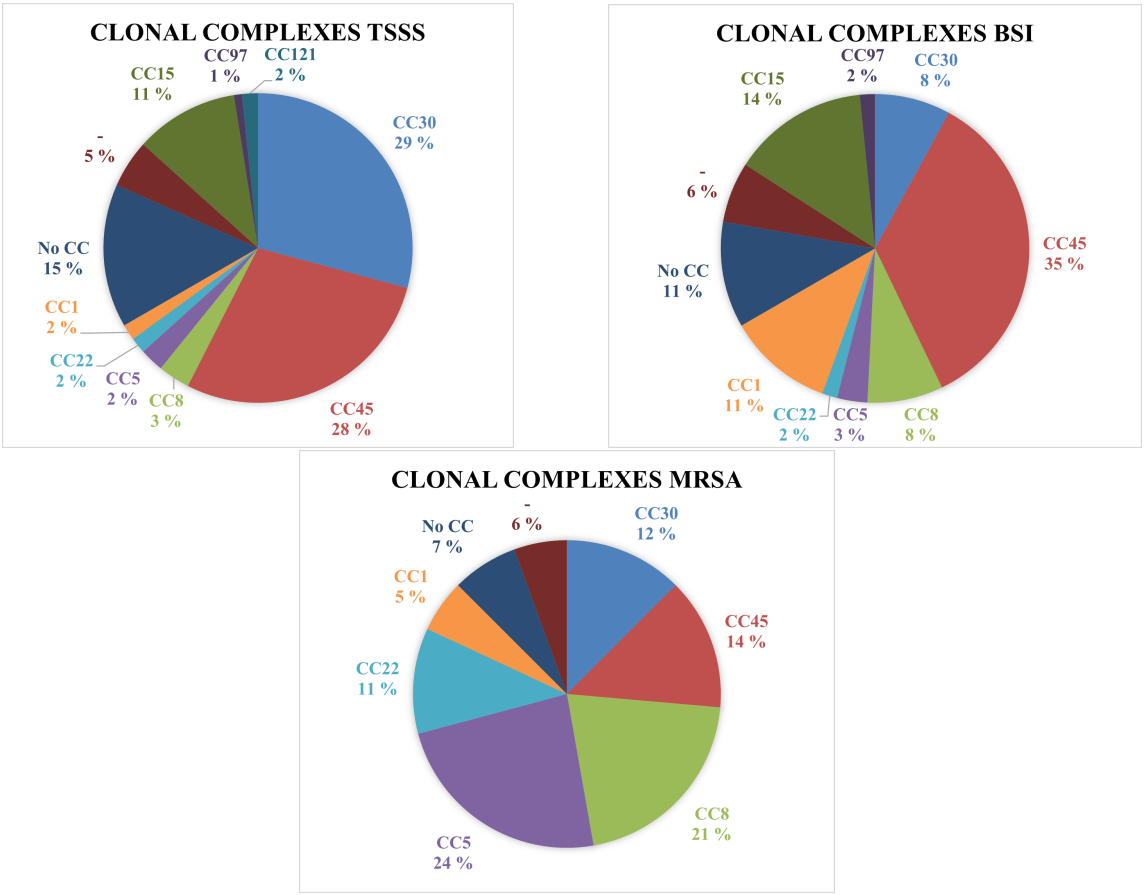


Figure 10.1: The frequency of clonal complexes in the TSSS, BSI and MRSA cohorts. Group "-" are strains with no sequence type. "No CC" are strains with sequence types not associated with a clonal complex.

Table 10.4: The table shows number of strains in each clonal complex group for each cohort TSSS, BSI and MRSA.

	CC30	CC45	CC8	CC5	CC22	CC1	No CC	-	CC15	CC97	CC121
TSSS (n=120)	35	34	4	3	2	2	18	6	13	1	2
BSI (n=63)	5	22	5	2	1	7	7	4	9	1	-
MRSA (n=72)	9	10	15	17	8	4	5	4	-	-	-

11 Resistance genes

Resistance genes were detected in the three cohorts to characterise the diversity of resistance genes in commensal carriage, BSI-associated and carriage MRSA strains. The resistance genes, and the frequency detected in the three cohorts MRSA, TSSS and BSI are shown in table 11.1. The cutoff for percentage identify was 80% identity and 95% gene coverage for the detection of the resistance genes from the database. The biggest difference in prevalence of resistance genes was seen between the MRSA cohort and the two others. A total of 27 resistance genes were present in the MRSA cohort, 9 in the TSSS cohort and 10 in the BSI cohort. The median of resistance genes detected in each cohort was TSSS=3, BSI=3 and MRSA=6. This indicates that MRSA strains are resistant to more types of antibiotics than MSSA strains. The frequencies of the genes detected in both the BSI cohort and the TSSS cohort were relatively similar and the median exactly the same. This indicates that resistance genes not necessarily is a factor in the development of BSI.

Ant(4')-Ia, *lnu(A)* and *fus(B)*, associated with resistance to aminoglycoside, lincosamide and fusidic acid respectively, were each detected in 1% of the carriage TSSS strains, but not in the BSI-causing strains. The resistance genes *str* (2%), *erm(C)* (2%), *fus(C)* (5%) and *fosD* (2%), associated with resistance to streptomycin, marcolide, fusidic acid and fosfomycin respectively, were detected in some of the BSI-causing strains, but not in any of the carriage TSSS strains. The frequencies of these genes were very low, indicating that they are not relevant factors in the development of BSI by BSI-causing strains. All MRSA strains, and none of the strains from the BSI and TSSS cohorts, contained the *mecA* gene associated with methicillin-resistance in *S. aureus*. The most common resistance genes observed in both the TSSS and BSI cohorts were *blaZ*, which is associated with penicillin resistance, and *fosB-Saur*, which is associated with fosfomycin resistance. These genes were also the most observed in the MRSA cohort after *mecA*. An additional 13 resistance genes were observed in 10% or more of the MRSA strains.

Table 11.1: Resistance genes observed in the MRSA, TSSS and BSI cohorts. Number of strains with specific resistance gene is given, with the percentage in parentheses.

Resistant to	**Resistance gene**	**MRSA** n=72	**TSSS** n=120	**BSI** n=63
Methicillin	*mecA*	72 (100%)	-	-
β-lactam	*blaZ*	38 (53%)	21 (18%)	12 (19%)
Tetracycline	*tet(K)*	11 (15%)	4 (3%)	2 (3%)
	tet(M)	8 (11%)	-	-
Aminoglycoside	*aph(3')-IIIa* (*aphA3*)	15 (21%)	-	-
	ant(4')-Ia (*aadD*)	8 (11%)	1 (1%)	-
	ant(6)-Ia (*aadE*)	3 (4%)	-	-
	ant(9)-Ia (*spc*)	8 (11%)	2 (2%)	2 (3%)
Streptomycin	*str*	-	-	1 (2%)
Marcolide	*erm(C)*	8 (11%)	-	1 (2%)
	erm(A)	7 (10%)	2 (2%)	2 (3%)
	erm(B)	3 (4%)	-	-
	msr(A)	9 (13%)	-	-
	mph(C)	9 (13%)	-	-
Bleomycin	*bleO*	8 (11%)	-	-
Phenicol	*fexA*	2 (3%)	-	-
	catA7	1 (1%)	-	-
	catA8	1 (1%)	-	-
	cat-TC	1 (1%)	-	-
Streptothricin	*sat4*	12 (17%)	-	-
Lincosamide	*lnu(A)*	1 (1%)	1 (1%)	-
	vga(A)	1 (1%)	-	-
Trimethoprim	*dfrG*	7 (10%)	1 (1%)	2 (3%)
	dfrC	12 (17%)	-	-
Fusidic acid	*fusB*	1 (1%)	1 (1%)	-
	fusC	6 (8%)	-	3 (5%)
Fosfomycin	*fosB-Saur*	46 (64%)	68 (57%)	29 (46%)
	fosD	1 (1%)	-	1 (2%)

12 Pangenome

The pan and core genome of the 255 *S. aureus* strains from all three cohorts TSSS, BSI and MRSA was defined by Roary. This was done in order to obtain a core genome alignment from all the genomes in the cohort. This could further be used to identify and compare closely related strains from different cohorts, in addition to comparing the core genome from this project to previously reported core genome sizes of *S. aureus*. A pie chart, shown in figure 12.1, represents the distribution of genes in the pangenome of the 255 strains by categorizing them as core, soft-core, shell and cloud genes. Roary defined 1593 genes as core genes as they were present in 99%-100% of all strains. This is within the span of *S. aureus* core genome sizes. 221 genes were present in 95%-98% of the strains, and were defined as soft-core genes. There were 1458 shell genes and 5151 cloud genes, which were present in 15%-94% and less than 15% of the strains respectively.

The pangenome matrix in figure 12.2 shows all genes in the pangenome of the 255 strains from all three cohorts, and whether the genes are present or absent in each strain. The strains are presented with a phylogenetic tree and the presence or absence of genes are shown in the matrix to the right of the strains. The core genome is observed to the left in the matrix, where all genes are present in approximately all strains, while additional soft-core, shell and cloud genes present in a selection of strains are shown to the right of

the core genome in the matrix. A figure showing the pangenome frequency can be seen in appendix E.

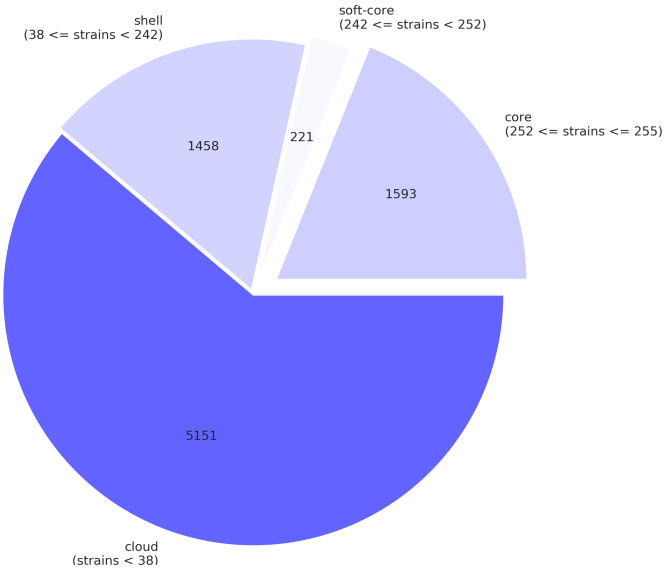


Figure 12.1: Pangenome pie showing the distribution of core, soft-core, shell and cloud genes for all strains in the cohorts BSI, TSSS and MRSA. (Core genes are present in more than 99% of all the genomes, soft-core genes are present in 95%-99% of the genomes, shell genes are present in 15%-95% of genomes and cloud genes are found in less than 15%.)

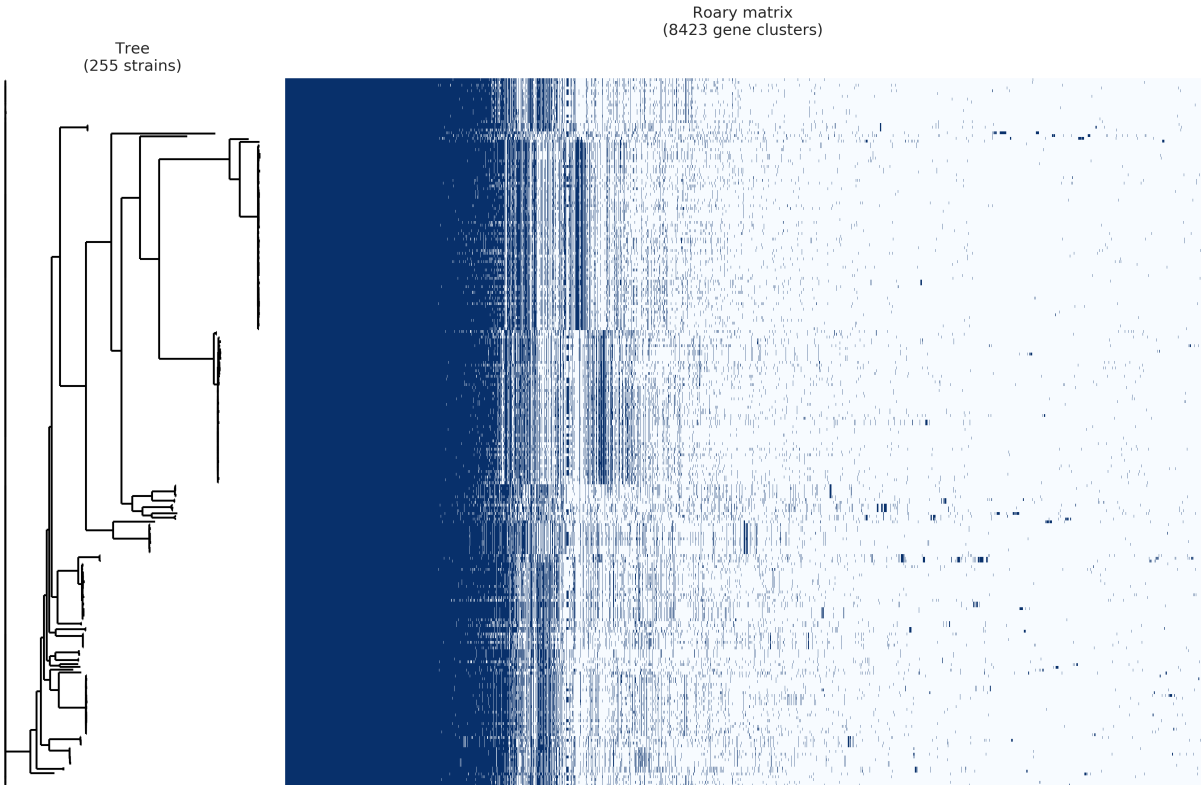


Figure 12.2: Pangenome matrix showing the presence and absence of all genes in the pangenome of all 255 strains in the three cohorts TSSS, BSI and MRSA.

13 Phylogenetic analysis and virulence genes

13.1 Phylogenetic analysis

The full phylogenetic tree containing all strains from all three cohorts TSSS, BSI and MRSA was constructed with the core genome alignment from Roary and visualized in iTOL (appendix F). The strains cluster into two main groups at the first node. One of the clusters contains CC45, CC30 and a couple of additional STs not associated to a CC, while the other cluster contains the rest of the strains. Strains in the same CC cluster together, and strains with the same ST not associated with a CC cluster together. Most of the strains from all cohorts cluster by CCs irrespective of their pathogenecity. This means that strains from different cohorts in some cases are closely related to each other. One strain from CC5, two from CC1 and three from CC8 do not cluster together with the strains in their assigned CC. This could indicate that these strains do not necessarily belong to their assigned CCs based on their core genome. A strain can be assigned to the wrong CC as the assignment of CC is based on a few housekeeping genes rather than the entire core genome.

Subtrees of the two largest CCs, CC30 and CC45, are shown in figures 13.1 and 13.2 respectively. CC30 consists of the predominant ST30 and six additional STs. ST34 is separated from the rest of the strains at the first node. The rest of the CC30 strains cluster into to groups. The upper half cluster consists of ST1889, ST1879, ST39 and 10 predominant ST30 strains. The lower half consists of predominant ST30, ST1890 and ST1884. CC45 consists of the predominant ST45 and ten additional STs. One predominant ST45 (SO-SAU8-21) is separated from the rest of the strains at the first node. The rest of the CC45 strains are closely related, according to the phylogenetic tree.

13.2 Virulence genes

In order to look for factors that could be more significant in BSI-causing strains, the virulence genes (factors) most associated with BSI from literature was characterised in all the strains. The presence/absence of these genes is shown next to each strain in the full phylogenetic tree (appendix F). The prevalence of the virulence genes in each cohort is shown in table 13.1. The median prevalence of the 17 analysed virulence genes in each cohort was TSSS=12, BSI=14 and MRSA=14.

The virulence genes *aur*, *hlgC*, *hlgB*, *hlgA* and *scn*, all relevant in immune system evasion, are present in all or almost all of the strains in all three cohorts. The pore forming gene *hla* is also present in all strains, making it likely that all the strains in the three cohorts can lyse epithelial, endothelial and immune cells. The remaining virulence genes are present in various degree depending on cohort. The TSSS cohort has the highest, and the MRSA cohort the lowest, prevalence of *chp*, which is involved in immune evasion. The *pvl* component *lukS-PV* is almost exclusively seen in the MRSA cohort. Only one of the BSI strains has the gene. All the *sdr* genes and both *clf* genes, all associated with coagulation and aggregation, are most prevalent in the MRSA cohort and least prevalent in the TSSS cohort. The BSI cohort has the highest, and the TSSS cohort the lowest, prevalence of the *fnbp* genes. *VWbp* is also most prevalent in the BSI cohort. The TSSS strains has the lowest frequency for 11/17 virulence genes, which could indicate that these genes are relevant in the development of BSI.

Table 13.1: The prevalence of the important virulence genes associated with bloodstream infection-causing *Staphylococcus aureus* detected in the three cohorts TSSS, BSI and MRSA. The genes are grouped according to their relevance during BSI.

	Virulence gene	**TSSS** n=120	**BSI** n=63	**MRSA** n=72
Immune system evasion	*aur*	119 (99%)	63 (100%)	72 (100%)
	hlgC	119 (99%)	63 (100%)	72 (100%)
	hlgB	120 (100%)	63 (100%)	72 (100%)
	hlgA	120 (100%)	63 (100%)	72 (100%)
	scn	113 (94%)	59 (94%)	70 (97%)
	chp	95 (79%)	48 (76%)	49 (68%)
	lukS-PV	0 (0%)	1 (2%)	15 (21%)
Coagulation and aggregation	*sdrE*	91 (76%)	56 (89%)	70 (97%)
	sdrD	41 (34%)	29 (46%)	54 (75%)
	sdrC	46 (38%)	56 (89%)	70 (97%)
	fnbpB	66 (55%)	49 (78%)	45 (63%)
	fnbpA	82 (68%)	58 (92%)	64 (89%)
	clfB	75 (63%)	54 (86%)	65 (90%)
	clfA	98 (82%)	59 (94%)	70 (97%)
	vWbp	82 (68%)	46 (73%)	40 (56%)
Pore forming	*hla*	120 (100%)	63 (100%)	72 (100%)

Prevalence of the virulence genes were CC-specific. None of the virulence genes is associated with only a single CC or ST group, but *vWbp* is the virulence gene detected in the fewest CCs and ST groups. Examples of other virulence genes that are absent from all of the strains in specific CCs or STs are *chp*, which is absent from CC97, CC121 and ST101, and *fnbpB* which is absent in CC30, ST182 and ST25. This indicates that the presence of these virulence genes could be determined by what CC or ST the strain belongs to. The group missing the most of the virulence genes compared to the other CCs and STs (not associated with a CC) is ST182, which consists of three carriage TSSS strains. *FnbpB*, *fnbpA*, *sdrE* and *vWbp* was not detected in any of the ST182 strains. This could indicate that strains with ST182 are less likely to be pathogenic.

In the CC30 subtree, the frequency of virulence genes is the lowest in ST1879 and ST1884, where the median is 9. The highest frequency is observed for ST39 and ST1889, where the median is 12. All strains within these STs belongs to the TSSS cohort, expect one strain from ST39 which belongs to the BSI cohort. This indicates that the ST, rather than the cohort, is more likely to determine the frequency of virulence genes. A bigger difference is observed between the strains clustered together in the top half of the CC30 subtree and the strains clustered together at the bottom half. In the top half, 16/18 strains (excluding ST34) contains *fnbpA*, while it is observed in only one of the strains in the other cluster. In addition, only the MRSA strains in the top cluster contains the *pvl* genes (7/8). These differences indicates that the strains in the top cluster have evolved differently than the strains in the bottom cluster, as assumed by the phylogenetic tree. Another big difference can be seen in the *sdrC* frequency. The gene is present in all of the BSI and MRSA strains, but only detected in 17% of the TSSS strains.

There are differences in virulence genes presence between different pairs of BSI and carriage strains in the CC30 subtree. This indicates that it is not a consistent difference

between the strains in the cohorts. One example is Tromso9100 from the TSSS cohort and STAU275 from the BSI cohort, which are both ST30 and each others closest relative according to the tree. The only difference in virulence gene presence is that *sdrD* was only detected in Tromso9100 and *sdrC* was only detected in STAU275. Other closely related BSI and TSSS strains have exactly the same virulence genes, while some has a bigger difference in virulence genes presence. For instance the *clf* genes, *fnbpA* and *sdrC* was detected in STAU279 (BSI cohort), but not in the closely related Tromso9092 (TSSS cohort).

In the CC45 subtree, the frequency of virulence genes is the lowest in ST455, where the median is 10.5. The highest frequency is observed for ST45, where the median is 15. There is however only two strains representing ST455 so the results might not be accurate for ST455 strain in general. As for CC30, a difference can be observed in the presence of certain virulence genes between the cohorts. *SdrC* is also detected in few TSSS strains (26%) compared to the BSI (91%) and MRSA (100%) strains. Another gene present in few TSSS strains compared to the other cohorts is *clfB*. It was detected in 29% of the TSSS strains, 77% of the BSI strains and 80% of the MRSA strains. Other genes that had differences in frequency between the three cohorts were *clfA* and *sdrD*. The frequency of *clfA* in the three cohorts was TSSS=71%, BSI=95% and MRSA=100%. This could indicate that the BSI and MRSA strains are more likely to be involved in aggregation. On the other hand, *sdrD*, which is also involved in aggregation, was observed with a higher frequency in the TSSS strains. The frequencies of *sdrD* was TSSS=38%, BSI=9% and MRSA=10%. As for CC30, there is no consistent difference between closely related TSSS and BSI strains. For instance is *clfA* and *sdrD* present in STAU277 (BSI cohort) and absent in the closest relative Tromso9105 (TSSS cohort), while *clfB* and *sdrC* is present in STAU284 (BSI cohort) and absent in Tromso9063 (TSSS cohort). Tromso9063 does additionally have *sdrD*.

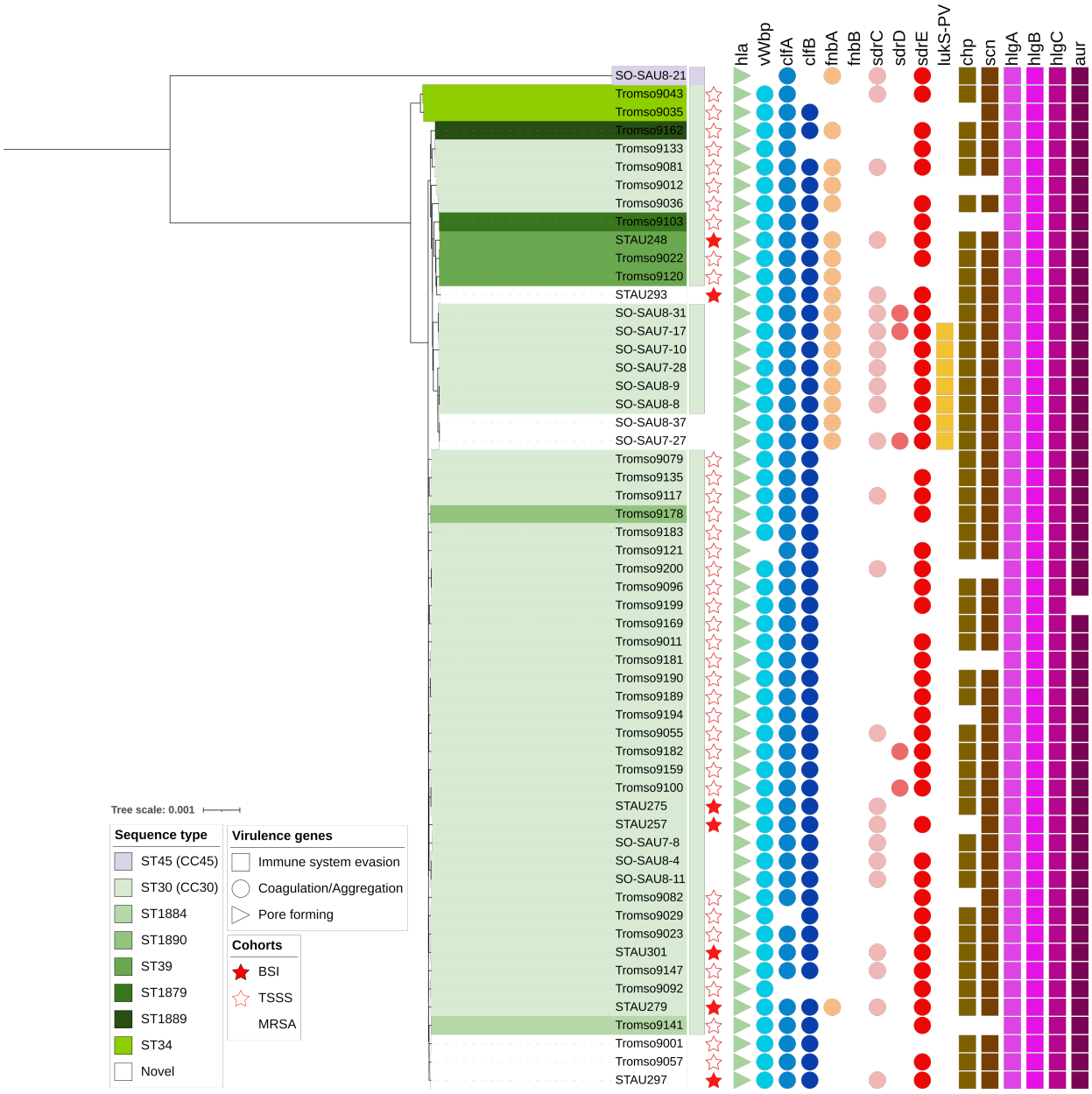


Figure 13.1: Phylogenetic tree showing all strains in CC30 from all three cohorts TSSS, BSI and MRSA, with a ST45 strain as an outgroup. The strains from the BSI cohort are marked with a filled red star, the TSSS strains are marked with a white star and the MRSA strains have no marking. The strains are marked in colours according to their sequence type (ST). The strains with no colour marking did not get a ST during Multi-Locus Sequence Typing. The presence and absence of virulence genes relevant in bloodstream infection-causing *Staphylococcus aureus*, are displayed to the right of the tree. Virulence genes marked with a rectangle plays a role in immune system evasion, the circular plays a role in coagulation and aggregation of blood and the pore forming gene *hla* is marked with a triangle.

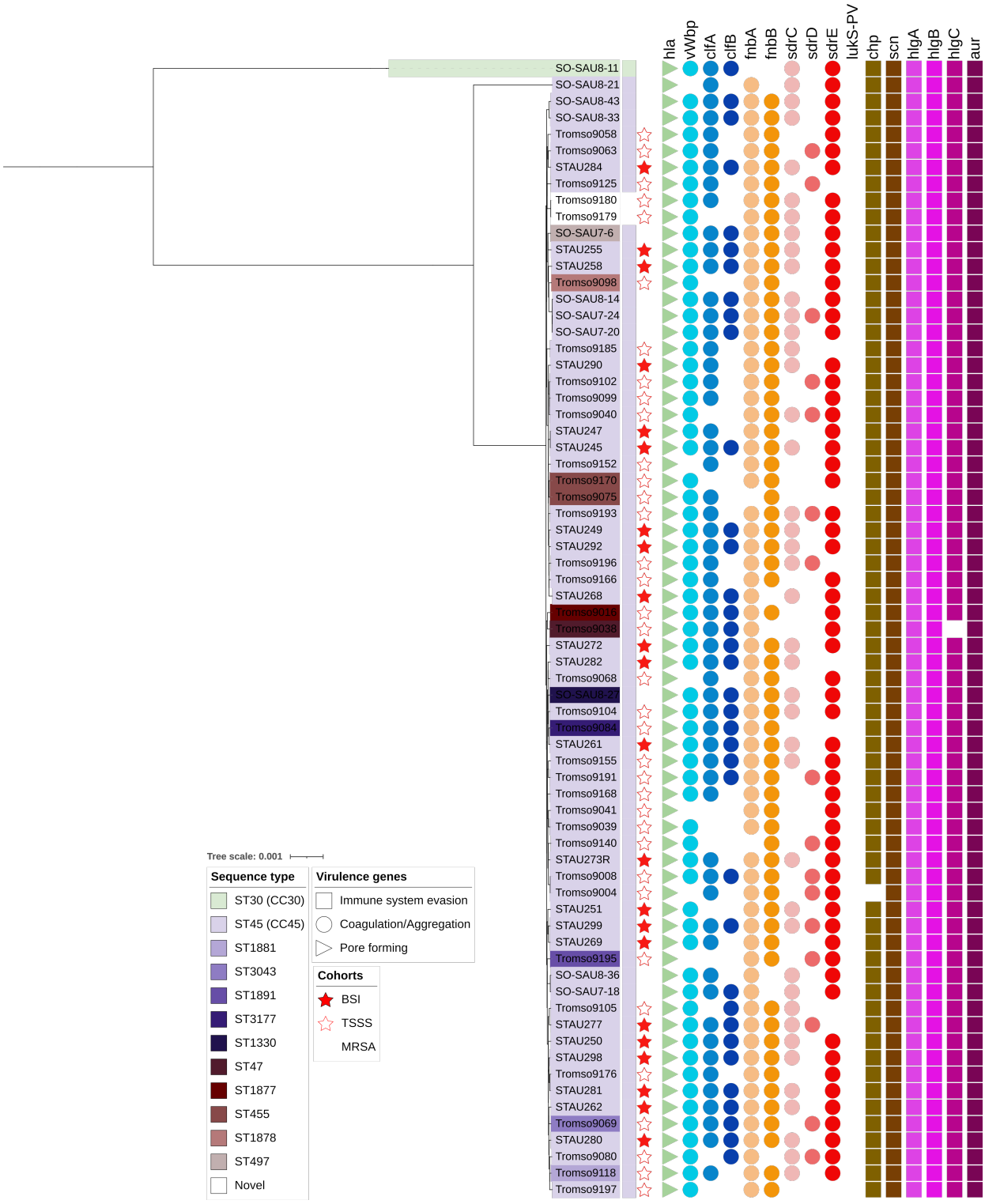


Figure 13.2: Phylogenetic tree showing all strains in CC45 from all three cohorts TSSS, BSI and MRSA, with a ST30 strain as an outgroup. The strains from the BSI cohort are marked with a filled red star, the TSSS strains are marked with a white star and the MRSA strains have no marking. The strains are marked in colours according to their sequence type (ST). The strains with no colour marking did not get a ST during Multi-Locus Sequence Typing. The presence and absence of virulence genes relevant in bloodstream infection-causing *Staphylococcus aureus*, are displayed to the right of the tree. Virulence genes marked with a rectangle plays a role in immune system evasion, the circular plays a role in coagulation and aggregation of blood and the pore forming gene *hla* is marked with a triangle.

14 Microbial GWAS

Microbial GWAS was performed with Scoary, which was run individually for the largest CC groups CC45, CC15, CC1, CC8, CC30, CC5 and CC22. The aim was to see if any genes were significantly associated with BSI-causing strains, so the strains from the BSI cohort was compared to the strains from the TSSS and MRSA cohorts.Three significant hits after correction, with a significance level of 0.05, were obtained from the Scoary run with CC45. The significant hits are shown in table 14.1, presenting the genes that are present in the BSI cohort compared to those in the TSSS and MRSA cohorts. The hits were *clfB* and S-formylglutathione hydrolase (group3115, group3114), where S-formylglutathione hydrolase (group3115) was significant in the TSSS and MRSA cohorts (35/44 strains), while S-formylglutathione hydrolase (group3114) (18/22 strains) and *clfB* (17/22 strains) were present in BSI-causing strains. This is similar to what was observed in the CC45 subtree. There was a noticeable difference between the frequency of *clfB* between the strains from the BSI cohort and the ones from the TSSS cohort. The clumping factor *clfB* is involved in aggregation of blood and *S. aureus*, which is a central part of BSI. When this gene is more present in strains that have caused BSI, it indicates that it could be an important gene in *S. aureus* strains causing BSI. S-formylglutathione hydrolase is present in both the BSI-causing strains and the non-causing strains from the TSSS and MRSA cohorts, but with different assigned groups. This indicates that the BSI strains have a different version of the gene than the strains from the TSSS and MRSA cohorts. None of the other Scoary runs gave significant results.

Table 14.1: The significant hits (p<0.05) for CC45 after running a microbial Genome-Wide Association Study with Scoary, where BSI-causing strains is set as trait. The names and annotation of the genes significantly more or less present in BSI-causing strains opposed to strains in the TSSS and MRSA cohorts are presented. The number of strains with or without the trait and with or without the specific genes are shown in columns 3-6, where P=positive, N=negative, T=trait and G=gene. Sensitivity, specificity and odds ratio is presented for each hit. The naive p-value is adjusted for with Bonferroni correction and the Benjamini-Hochberg procedure.

Gene	Annotation	#PT PG	#NT PG	#PT NG	#NT NG	Sensitivity	Specificity	Odds ratio	Naive p	Bonferroni p	Benjamini H. p
clfB	hypothetical protein	18	9	4	35	81.82	79.55	17.5	2.51e-06	0.0060	0.0060
group3115	S-formylglutathione hydrolase	5	35	17	9	22.73	20.45	0.0756	1.27e-05	0.0306	0.0102
group3114	S-formylglutathione hydrolase	17	9	5	35	77.27	79.55	13.22	1.27e-05	0.0306	0.0102

15 Assembly of the *SdrC* gene

In order to identify virulence genes that are shared and unique within BSI or the carriage strains, strains belonging to the different cohorts with less than 110 SNPs difference in the SNP matrix was selected. The presence of virulence genes associated with BSI was compared between the selected strains from the different cohorts. The *sdrC* gene appeared more frequently in the selected BSI strains than the TSSS strains, as it was detected in 90% of the BSI strains and 38% of the TSSS strains. The gene was analysed further as the difference in frequency of the gene between the BSI-causing strains and the carriage TSSS strains could indicate an association to BSI-causing strains.

The genome of each strain was searched using NCBI-BLAST against a local database with four *sdrC* reference sequences to see if the frequency of the gene would be similar to the result obtained with Nullarbor. *SdrC* was found in 94% (59/63) of the BSI strains, 91% (109/120) of the TSSS strains and 100% (72/72) of the MRSA strains. This is a big difference in frequency for the TSSS strains compared to the results obtained with Nullarbor, where *sdrC* was detected in 38% (46/120) of the TSSS strain. The difference could indicate error during assembly leading to incomplete *sdrC* sequences that were not detected by Nullarbor (Abricate) due to a coverage below the threshold.

The frequency of *sdrC* for each ST with more than five strains, detected in Nullarbor and with sequence-specific analysis within Geneious, is shown in the left and right column of each cohort for BSI, TSSS and MRSA (this cohort has only one column) in table 15.1. For the Nullarbor results, a difference in *sdrC* frequency between the TSSS and BSI strains was detected for ST45, ST30, ST15, ST50 and ST22. *SdrC* was detected in all strains in both cohorts for ST8, ST25 and ST5. Looking at this result isolated could indicate that the presence of the *sdrC* gene is dependent on the ST of the strain. The biggest difference in *sdrC* frequency was observed between the TSSS and BSI strains in ST30 and ST45. All or most of the BSI strains had the gene, while the frequency was 19% and 32% for the TSSS strains respectively. To see if there could be an association between the *sdrC* gene and BSI-causing strains (based on the Nullarbor results), fisher's exact test was performed on BSI and TSSS strains belonging to the three largest ST groups ST45, ST30 and ST15, as seen in figure 15.1. The test gave a p-value of 6.3 e^{-5} for ST45, indicating that there was a significant association between the presence of *sdrC* and BSI-causing strains with ST45. For ST30, the p-value was 0.004, indicating a significant association. For ST15 the p-value was 0.092, which is not a significant result. For the results from the sequence-specific analysis, the *sdrC* gene was detected in all or almost all of the TSSS strains in all STs, except for ST15 where the frequencies remained the same as observed in the Nullarbor results. The difference in frequency between the TSSS and BSI strains in ST15 did not indicate an association between the presence of *sdrC* and BSI-causing strains, as calculated with fisher's exact test.

A phylogenetic tree (appendix G) with all *sdrC* sequences extracted in Geneious was created. The *SdrC* sequences are marked with the ST of the strain they were extracted from. This was only done for STs with more than 5 strains. The tree shows that *sdrC* sequences belonging to the same ST cluster together. This indicates that similarity in the *sdrC* gene is more related to the ST of the strain rather than the cohort. It also strengthens the indication that most of the *sdrC* genes in the TSSS cohort was missing due to incomplete assembly.

Table 15.1: The frequency of *sdrC* positive strains within each cohort for all sequence types (STs) with more than 5 strains. The left columns for the BSI and TSSS cohorts are the *sdrC* frequencies found by Nullarbor through Abricate. The right columns are the frequencies found with NCBI-BLAST in Geneious. The MRSA cohort has only one column, which contains the frequencies from Nullarbor. The STs are arranged in decreasing order from largest to smallest BSI and TSSS (combined) sample size.

	BSI		**TSSS**		**MRSA**
	Nullarbor	Geneious	Nullarbor	Geneious	Nullarbor
ST45	20/22 (91%)	22/22 (100%)	8/25 (32%)	24/25 (96%)	8/8 (100%)
ST30	4/4 (100%)	4/4 (100%)	5/27 (19%)	26/27 (96%)	9/9 (100%)
ST15	7/9 (78%)	7/9 (78%)	4/11 (36%)	4/11 (36%)	-
ST8	5/5 (100%)	5/5 (100%)	4/4 (100%)	4/4 (100%)	10/10 (100%)
ST25	2/2 (100%)	2/2 (100%)	4/4 (100%)	4/4 (100%)	-
ST50	1/2 (50%)	2/2 (100%)	0/3 (0%)	3/3 (100%)	-
ST5	2/2 (100%)	2/2 (100%)	2/2 (100%)	2/2 (100%)	11/11 (100%)
ST22	1/1 (100%)	1/1 (100%)	1/2 (50%)	2/2 (100%)	7/7 (100%)
ST1	3/3 (100%)	3/3 (100%)	-	-	3/3 (100%)

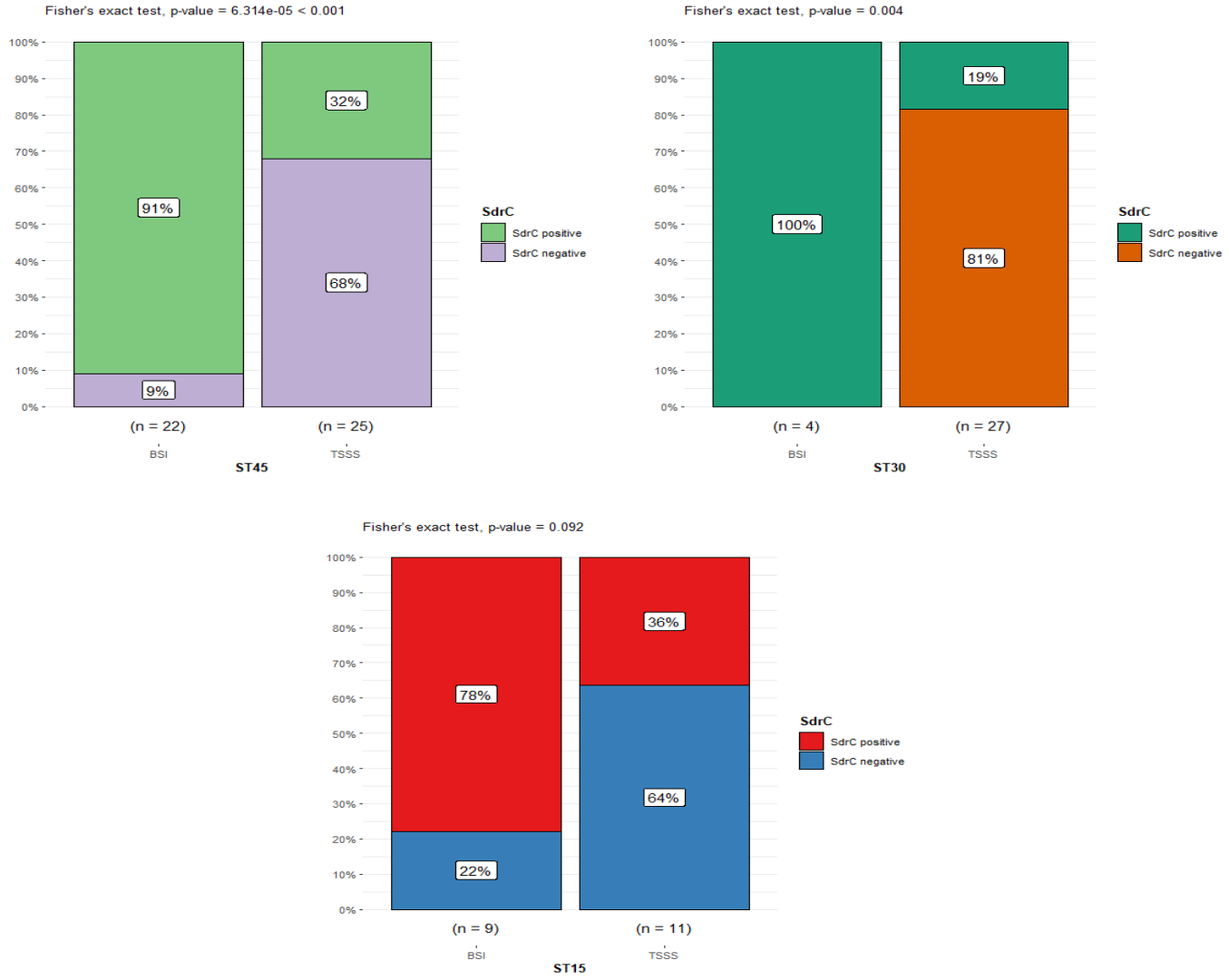


Figure 15.1: Fisher's exact test looking at a possible relationship between bloodstream infection-causing strains and the presence of the virulence gene *SdrC*. The test was performed on strains with ST45, ST30 and ST15 belonging to the BSI and TSSS cohorts.

Part V

Discussion

16 Comparison of MRSA, BSI-causing and carriage TSSS strains

The *S. aureus* strains belonging to the three cohorts MRSA, TSSS and BSI were characterized with regard to genotypes, virulence genes and resistance genes. Comparative genomics was performed using a phylogenetic tree and mGWAS. The aim was to identify important lineages or genes associated with pathogenicity.

16.1 MRSA strains

The factor determining whether a strain is methicillin-resistant, is the presence of the *mecA* gene. This gene was detected in all the strains from the MRSA cohort, confirming that all the strains were MRSA.

The strains from the MRSA cohort were associated with multiple dominating CCs, as shown in previous studies[118]. Enrigth *et al.*[119] found their epidemic MRSA strains, collected from 20 European countries, to be associated with the five CCs CC8, CC30, CC45, CC5 and CC22, while other studies on MRSA strains in Europe have mostly been associated with CC5, CC8, CC1, CC22, CC30 and CC45[120,121,122]. Rolo *et al.*[123] characterized a selection of infection-associated and carriage MRSA strains collected between 2000-2010 in 16 different European countries. The most prevalent CC observed was CC8, which was also one of the largest CCs in the MRSA cohort in this thesis. The study also showed a high frequency of ST80, but only two of our MRSA strains had this ST. Deurenberg *et al.*[120] characterized infection-associated and carriage MRSA strains collected between 1999-2004 in Belgium, Germany and the Netherlands, and found CC8 and CC5 to be most prevalent. Also Aanensen *et al.*[122] found CC5 and CC8, in addition to CC22, to be the largest CCs in a collection of infection-associated MRSA strains collected in 26 European countries in the period 2006-2007. These findings correlates well with the results obtained from our MRSA cohort, as CC5 and CC8 were the largest CCs, along with CC22, CC45 and CC30.

Several of the virulence genes relevant to BSI were detected with high frequency in the MRSA cohort. The genes *hlgA*, *hlgB* and *hlgC* active in immune system evasion were present in all or almost all of the MRSA strains. This fits well with a study conducted by Wang *et al.*[124], where all of the 24 infection-associated MRSA strains, collected in China, contained the *hlg* genes. The less frequent immune system evasion gene *chp* was seen in 68% of the strains in the MRSA cohort in this thesis, which is a bit lower than what Campbell *et al.*[125] reported in their study, where *chp* was seen in 81% of their 52 infection-associated MRSA strains collected in Australia and New-Zealand. Despite the small difference in frequency in this thesis and the Campbell *et al.* study, they both indicate high prevalence. In the study conducted by Wang *et al.*[124] 21% of the 24 strains contained *pvl*, which correlates well with the frequency of *LukS-PV* observed in our MRSA cohort, which was also 21%.

The frequency of the *sdr* genes was studied in 115 carriage MRSA strains collected in 12 European countries by Sabat *et al.*[126]. They detected *sdrC* in all strains, while the frequency of *sdrD* and *sdrE* was 99% and 88% respectively. The high prevalence of the *sdrC* gene is very similar to the results obtained in this thesis, where it was detected by Abricate in 97% of the strains, and after sequence-specific analysis within Geneious in all of the strains in the MRSA cohort. The frequency of the *sdrD* gene was also high, but not as high as in the Sabat *et al.* study, as 75% of MRSA strains contained the gene. The *sdrE* gene was observed in 97% of the strains, which is a bit more than reported in the Sabat *et al.* study, but equally of high prevalence. Two other studies conducted by Liu *et al.*[127] and Campbell *et al.*[125] also detected high frequencies of the *sdr* genes. Liu *et al.* studied 109 infection-associated MRSA strains collected in China, while Campbell *et al.* studied 52 infection-associated MRSA strains detected in Australia and New-Zealand. They both detected a lower frequency of *sdrD* (64% and 73% respectively) than *sdrC* (92% and 73% respectively) and *sdrE* (78% and 90% respectively). This trend was also seen for the MRSA strains in this thesis.

The adhesion factors *fnbpA* and *fnbpB* were detected in 74% and 19% of 43 non-redundant MRSA strains obtained in Tunisia in a study conducted by Haddad *et al.*[128]. The MRSA strains analysed in this thesis also had a higher frequency of *fnbpA* compared to *fnbpB*. However, the frequencies were higher in the MRSA cohort as 89% of the strains contained *fnbpA* and 63% contained *fnbpB*. The biggest difference is seen for the *fnbpB* gene, as it has a low prevalence in the Haddad *et al.* study and a high prevalence in this MRSA cohort. The clumping factors *clfA* (97%) and *clfB* (90%) were present in a high number of MRSA strains. A high frequency of *clfA* was also observed by Li *et al.*[129] who reported that all of their 76 infection-associated MRSA strains collected in China had the gene, while Campbell *et al.*[125] reported that only 62% of their 52 MRSA strains in their study contained *clfB*. The two studies indicate that the frequency of *clfB* is lower than the frequency of *clfA*, which was also observed in thesis. The pore forming *hla* gene was in the three studies conducted by Wang *et al.*[124], Li *et al.* and Campbell *et al.* observed in all or almost all of the strains, which also was the case for the strains in the MRSA cohort. The MRSA cohort had overall a high prevalence of virulence genes, with the exception of *pvl*, as seen in other studies. This means that most MRSA strains contain multiple relevant genes associated with *S. aureus* BSI.

16.2 BSI-causing strains

Similar to the MRSA strains, multiple CCs have previously been associated with BSI causing strains. Pérez-Montarelo *et al.*[130] studied BSI-causing strains collected between 2002-2017 in Spain, and the most prevalent CCs detected from this collection were CC5, CC30, CC45, CC8, CC15 and CC22. The exact same CCs were determined as major CCs in a study conducted by Rasmussen *et al.*[131] on BSI-causing strains in Sweden collected from 1980 to 2010. Similar results have also been detected in Norway and the Netherlands between 2001-2017, as CC45, CC30, CC15 and CC5 were determined major CCs[132,133]. Aanensen *et al.*[122] studied BSI-causing strains collected in 26 European countries in the period 2006-2007, and defined CC45, CC30 and CC15 as major CCs. Even though there was a span in collection years of the strains from the different studies, the same major CCs were observed. All of these CCs were also detected in our BSI cohort, with CC30, CC1 and CC15 being defined as the major groups. CC1 was not considered a major CC in the other European studies. CC97 was additionally detected in the BSI cohort, but was

not considered to be a major group, neither in the BSI cohort nor the European studies.

The analysis of the virulence genes relevant to BSI showed that the genes *aur*, *hlgA*, *hlgB* and *hlgC*, known to be relevant in immune system evasion, were present in all of the strains. This high frequency has also been seen in other studies. Rasmussen *et al.*[134] reported that all 88 analysed BSI-causing strains, collected in Sweden, contained *aur* and *hlgA*, while Wang *et al.*[124] reported 100% presence of the *hlg* genes in the 77 BSI-causing strains studied. Rasmussen *et al.* also reported that 78% of 88 strains contained *chp*, and 98% contained scn. This high frequency is similar to what was observed in the BSI cohort, where 76% of the strains had *chp* and 94% had *scn*. High prevalence of these genes was also reported by Blomfeldt *et al.*[135], that studied 126 BSI-causing strains collected in Norway, with a 75% frequency of *chp* and 96% frequency of *scn*. Blomfeldt *et al.*[135] also found *pvl* to be present in 2% of their 126 strains, while Rasmussen *et al.*[134] found *pvl* in 1% of 88 strains. This low prevalence correlates well with the frequency of *lukS-PV* found in the BSI cohort in this thesis, as 1% of the strains contained this gene.

The studies conducted by Sabat *et al.*[126], which studied 52 strains collected from 12 European countries, Blomfeldt *et al.*[135] and Rasmussen *et al.*[134] show varying prevalence of the *sdr* genes in their collections of BSI-causing strains. The frequency of *SdrC* was reported to be 75%-100% in the different studies, which fits well with the 89% frequency of the gene observed in the BSI cohort, and the 94% detected by sequence-specific analysis. The frequency of the *sdrD* varied the most, as it was found in 44%-92% of the strains in the different studies[126,135,134]. In the BSI cohort in this thesis, the prevalence of the *sdrD* gene was on the lower side of the reported frequencies, as it was detected in 46% of the strains. A high prevalence of *sdrE* was detected in the BSI cohort and the three studies. The frequency of the gene in the studies was 73%-85%, while it was 89% in the BSI cohort.

Li *et al.*[136] (80 strains collected in China) and Rasmussen *et al.*[134] both reported higher frequencies of *fnbpA* (56% and 100%) than *fnbpB* (36% and 81%), which was also seen in the BSI cohort where the frequencies were 92% strains for *fnbpA* and 78% for *fnbpB*. The studies also reported that all their strains contained *cflA*, which is similar to the finding in the BSI cohort, where the frequency was 94%. The *clfB* gene was also present in all the strains in the Rasmussen *et al.*, while the frequency was 89% in the BSI cohort. The pore forming gene *hla* was present in all strains in the BSI cohort, and this high prevalence is also reported by other studies[134,124,136]. The prevalence was high for most of the BSI-associated virulence genes in the BSI cohort. An exception was *sdrD*, which was detected in less than half of the strains, as well as *pvl*, which was detected in only one BSI strain. That not all the BSI-associated virulence genes was detected in all BSI strains, indicates that it is not necessary for all genes to be present for a strain to cause BSI.

16.3 Carriage *S. aureus* strains

Carriage *S. aureus* detected in Europe have been associated with multiple CCs, mostly with CC30, CC45, CC15, CC8 and CC22[137,134,138], which is similar to what was observed in the TSSS cohort. Holtfreter *et al.*[137] did a study on carriage *S. aureus* strains detected in Germany between 2008-2012. CC30, CC45, CC15, CC8 and CC22 were the five biggest CCs detected in that study. In studies of carriage strains collected in 2005-2008 from France, Moldova and Switzerland, the most prevalent CCs were CC30, CC45, CC5, CC8,

CC15 and CC121[139,140]. The major CCs observed in these studies were also detected in the TSSS cohort, where the largest CCs were CC30, CC45 and CC15. CC8, CC5, CC22 and CC121 were also detected in the TSSS cohort, but were not considered major CCs. In addition to the CCs most prevalent in the studies, strains belonging to CC1 and CC97 was observed in the TSSS cohort, but was not considered major groups. Sangvik *et al.*[141] found the ST and CC of 176 strains from The Tromsø Staph and Skin Study, which is the same study the TSSS strains are collected from. CC30, CC45 and CC15 were the major CCs, exactly as observed for our TSSS strains. A total of 16 CCs was detected in the study, which is more than the 9 CCs detected in our TSSS strains. They also detected 49 unique STs, which is similar to the amount of STs detected in our TSSS, as we found 43 unique STs, including novel STs.

In the TSSS cohort, the virulence genes *aur*, *hlgC*, *hlgB* and *hlgA* involved in immune evasion and the pore forming virulence gene *hla* were present in all or almost all of the strains. In studies performed by Peacock *et al.*[142], which analysed 179 carriage strains, and Rasmussen *et al.*[134], which studied 46 carriage strains, the frequency of the *hla* gene was 100%. This high prevalence of *hla* was also observed in the TSSS cohort. Rasmussen *et al.*[134] also found *aur* and *hlgA* in all studied strains. The genes *chp* and *scn*, also involved in immune system evasion, were in the study found in 83% and 89% of the strains respectively. Also in the TSSS cohort, *scn* was detected in more strains than *chp*, with frequencies of 94% and 79% respectively. None of the strains in the Rasmussen *et al.* study contained *pvl*, which was also the case for *lukS-PV* in the TSSS cohort.

For the virulence genes *sdrC*, *sdrD* and *sdrE* involved in coagulation and aggregation, the studies conducted by Peacock *et al.*[142], Sabat *et al.*[126] and Rasmussen *et al.*[134] found them to be present in various degrees in their collections of carriage strains. *SdrC* was detected in all the strains in all three studies, while *sdrD* was detected in 42%-87% of the strains, and *SdrE* was detected in 40%-89% of the strains. The biggest difference was seen for *sdrC*, where all the studies showed a high prevalence of the gene, while it was only detected in 38% of the TSSS strains. However, after sequence-specific analysis of the *sdrC* gene, the frequency of the gene in the TSSS cohort increased to 91%, which is a lot closer to the high prevalence observed in other studies. Possible reasons for this could be incomplete assembly and limitations in annotation of the *sdrC* gene, as discussed below. *SdrE* had a frequency of 76% in the TSSS cohort, which is within the values reported in other studies. The *sdrD* gene had a lower prevalence in the TSSS cohort, with a frequency of 34%, which is also seen in the Sabat *et al.* study.

Nashev *et al.*[143] (32 strains) and Rasmussen *et al.*[134] found that *fnbpA* were present in all strains and *fnbpB* were present in 40%-59% of the strains. Both the studies and the TSSS cohort show a larger frequency of the *fnbpA* gene, but the frequency of *fnbpA* (68%) was lower in the TSSS cohort than in the studies. The frequency of *fnbpB* (55%) was similar to the values reported by the two studies. Rasmussen *et al.*[134] also reported that *clfA* and *clfB* were present in all carriage strains in the study. This was a higher prevalence than in the TSSS cohort from this thesis, where 82% of the strains had *clfA* and 63% had *clfB*. For the TSSS strains, it was observed higher prevalence of virulence genes involved in immune system evasion compared to those involved in coagulation and aggregation. This indicates that the strains is more capable of affecting the immune system than promote the production of fibrin clots.

16.4 Comparison of cohorts

When comparing the major CCs between MRSA, BSI-causing and carriage *S. aureus* strains detected in studies, they are mainly associated with the same CCs. Previous European studies have however shown that MRSA strains have fewer main CCs than BSI-causing and carriage *S. aureus*. In this thesis, the MRSA cohort had more major CCs (>10% of the strains) than the BSI and TSSS cohorts, but more CCs was detected in the TSSS cohort. It is important to note that studies can have different definitions of what is considered a major group, which will impact how many major CCs is reported. That more major CCs were detected in the MRSA cohort, compared to other European studies, could indicate that there was a greater diversity of MRSA strains within Norway compared to Europe. 72 MRSA strains were studied in this thesis, and a larger collection could have impacted the distribution of CCs. CC15 was determined a major CC in previous studies on BSI-causing and carriage *S. aureus* strains, but not for MRSA strains. Few cases of CC15 is generally reported for MRSA strains[144]. This was also the case for the strains from the three cohorts studied in this thesis. None of the MRSA strains were associated with CC15, while in the BSI and TSSS cohorts, CC15 was one of the main CCs. One reason for few or no reported CC15 in MRSA strains could be that methicillin-resistance developed later in CC15 strains than in other CCs, and is therefore less common.

Differences in the presence of virulence genes between the BSI-causing strains and the carriage TSSS strains are especially interesting, as it could indicate that certain virulence genes contributes to or are important for *S. aureus* strains to cause BSI. The genes involved in coagulation of blood and aggregation did all have a higher frequency in the BSI-causing strains than in the TSSS cohort. It could indicate that having these virulence genes increases the possibility of a *S. aureus* strain to cause BSI. However, the differences in frequencies are not that big for most of the virulence genes, and additional analysis would be necessary to determine with greater certainty the activity of these genes. The only indication from the results that the presence of specific virulence genes could be associated with BSI, is the significant difference between the presence of *clfB* in BSI-causing strains and the carriage strains from the TSSS and MRSA cohorts, obtained from the mGWAS. This is only seen in the CC45 strains, which could indicate that *S. aureus* strains belonging to CC45 and has *clfB* is more likely to cause BSI than a CC45 strain without the gene. The gene is active in aggregation where *S. aureus* binds to blood clots, which is an important step in the development of sepsis. It therefore is possible that *clfB* could be a factor increasing the probability of sepsis. The gene has previously been found to improve the adherence to human nasal epithelial cells, and therefore promote colonization[145]. This is however relevant in both BSI-causing and carriage strains, and does not explain why *clfB* could possibly increase chances of BSI. The mGWAS of the CC45 strains also showed a significant association of S-formylglutathione hydrolase to BSI-causing strains for one allele of the gene, and association to the carriage MRSA and TSSS strains for another allele of the gene. However, the function of the gene could be the same for both alleles and not have an impact on the pathogenicity of the strains. One reason why significant associations is detected in only one CC could be that the strain collection is too heterogeneous to compare all strains to each other at once. The other CCs had less strains than CC45, which could have given a poorer foundation for comparison and for giving significant results. A larger collection of more similar strains, for instance belonging to the same CC, would have provided a better foundation for comparison.

The virulence genes *aur*, *hlg* and *hla* were present in all or almost all of the strains in the TSSS, BSI and MRSA cohorts, as well as in other studies on carriage and BSI-causing *S. aureus* strains. This could indicate that these genes are important for the survival of *S. aureus*, or makes it more likely that it survives in a human host. When comparing the MRSA cohort to the others, the frequency of *lukS-PV* stands out as none of the TSSS strains and only 2% of the BSI strains contain the gene, while the frequency is 21% for the MRSA strains.

The presence of specific virulence genes is not necessarily the main reason why some strains cause BSI while some remain carriage strains, as bigger differences between the virulence gene presence was seen between CCs and STs clustering together in the phylogenetic tree rather than between cohorts. *VWbp* was detected in all CC1 strains, but not in any of the CC15 strains. *FnbpB* was not seen in any of the CC30 strains, but seen in almost all of the CC45 strains, regardless of what cohort the strains belonged to. When comparing specific strains from different cohorts that have the same ST and are closely related according to the phylogenetic tree, they mostly have more virulence genes in common than with strains from the same cohort, but with a different ST. For example Tromso9115 (TSSS) and STAU302 (BSI) are both ST22 and according to the tree closest related to each other, and the difference in virulence genes is STAU302 having *clfB*, which Tromso9115 does not have. While the difference between STAU302 from ST22 and STAU275 (BSI) from ST30 is the presence of *sdrE*, *sdrD*, *fnbA* and *fnbB* in STAU302, while they are absent in STAU275. The genes are also present in Tromso9115 from ST22. STAU275 (BSI) is closest related to Tromso9100 (TSSS), and the difference between them is STAU275 missing *sdrD* and *sdrE*. This shows that the strains mostly have more virulence genes in common with strains from the same ST rather than the strains from the same cohort, which indicates that strains from both the TSSS cohort and BSI cohort had the ability to cause BSI, However, as mentioned previously, only a selection of virulence genes have been analysed. There could be other genes or factors determining if a strain is able to cause BSI.

17 Assembly and detection of repeat-rich genes

Most of the virulence genes in the BSI cohort had frequencies similar to those seen in previous studies. Many of the genes in the TSSS did however have a lower frequency than observed in other studies. This was especially seen for the virulence genes associated with coagulation and aggregation. These genes (except *vWpb*) are MSCRAMMs with tandem repeats. The *sdr* and *clf* genes have repeats of serine and aspartate, while the *fnbp* genes have fibronectin-binding repeats. As Tørresen *et al.*[64] have suggested, tandem repeats could lead to errors in assembly, which could mean that the MSCRAMM genes in some strains have not been assembled properly and therefore not been detected. When performing sequence-specific analysis of the *sdrC* gene, many strains were missing the end of the gene sequence, or one part of the gene was present on one contig, while the rest of the gene was present on another contig. This indicates that the repeat region located towards the end on the *sdrC* gene could have caused trouble for the assembly tool. When large parts of a gene is missing due to incomplete assembly, it can be hard to detect this during annotation as the gene sequence will often be too short for the coverage threshold to detect the gene.

Even though incomplete assembly of these genes could be one of the reasons why the annotation of *sdrC* is unpredictable in the TSSS cohort, it is questionable why this is not seen to such a large extent in the BSI cohort as well, as they were both sequenced with a HiSeq sequencer and the same read length. The average number of contigs and N50 values in the two cohorts were also relatively similar. One possible explanation for the difference in detection of the *sdrC* gene in the BSI and TSSS cohort could be that BSI-causing strains and carriage strains differ in the number of repeats in their repeat regions. Long repeat regions could make it harder to assemble the gene[64]. If the carriage strains typically have longer repeat sequences, this could be the reason why they were harder to assemble completely. This is however something that would have to be analysed further.

Limitations in databases could also, in addition to incomplete assembly, be a reason as to why virulence genes that are present in a strain is not detected. If a virulence gene has different alleles, and the database only has a few or one reference sequence of the virulence factor, some alleles could be too different from the reference to be detected. Large differences in repeat sequence lengths could limit the reference databases[64]. If the MSCRAMM genes have large differences in their repeat sequence lengths between isolates, STs or CCs, some could have not been detected.

The virulence factor database (VFDB)[109] was used by Nullarbor to detect virulence genes of each strain, and uses for instance the *clfB* sequence from 14 different reference genomes. If the diversity of the gene is well documented, then *clfB* is correctly identified, but if the reference sequences are too similar, it could lead to missing detection of *clfB*. There were 13 reference sequences for *sdrC* in the VFDB, and little variation in these could be one reason as to why *sdrC* was not detected in many of the strains. As seen in the phylogenetic tree constructed with the *sdrC* genes, the *sdrC* sequences detected in strains from the same STs cluster together. This indicates that there are allelic differences of the *sdrC* gene between different STs. Limited representation of this variation in VFDB could therefore be one reason why this gene was not detected in all strains.

Differences in annotation could also be one reason why *sdrC* was not significantly related to BSI strains in the mGWAS, as one would expect from the frequency of the gene identified from Abricate/VFDB. The mGWAS is based on the presence/absence file from Roary, which is based on the annotation performed with Prokka. Prokka uses other databases than VFDB, which appears to have limited variation of virulence gene reference sequences as the presence/absence file from Roary shows that *sdrC* was annotated in only 9 strains. This would not give a significant result in a mGWAS.

There are some limitations to the study. Which virulence genes that was focused on in this thesis was based on literature describing the pathogenisis of *S. aureus* BSIs. This limits the possibility of finding differences between BSI-causing and carriage strains. Even though the selected genes are relevant players in a BSI, there could still be others that have an impact on whether a strain causes a BSI or remains carriage strains. Another limitation of the study is the use of different sequencing equipment for the different cohorts. The BSI and TSSS strains were sequenced with a HiSeq machine giving 150 bp reads, while the MRSA strains were sequenced with a MiSeq machine giving 300 bp reads. This could give different quality of assembly leading to different foundations during detection of virulence genes.

18 Conclusions

In this thesis, three cohorts of *S. aureus* have been characterized with respect to genotype, virulence and resistance genes in order to determine if there are genetic traits that distinguish BSI-causing strains from carriage strains. A discovery of specific factors that are different in BSI-causing and carriage strains could potentially contribute to a decrease in BSI and sepsis cases caused by *S. aureus*. No apparent difference was found between virulence genes presence in BSI-causing strains and carriage strains, but rather between strains belonging to different STs and CCs. This indicates that there are other factors that contribute to the pathogenicity of *S. aureus* strains. The host-pathogen interaction is an important factor to consider that could be investigated further. This involves components in the host that could increase the adherence to *S. aureus*, make the endothelial barrier more prone to damage or make the immune system more likely to be inhibited. It could explain why, in a selection of genetically similar strains, some strains have caused infection in certain individuals while others remain carriage strains in other individuals. It is also important to acknowledge that possible problems relating to assembly and limitations in databases makes it more challenging to detect true differences between strains. Additional analyses could thus be necessary to confirm or invalidate possible findings.

References

[1] Kevin B. Laupland and Deirdre L. Church. Population-based epidemiology and microbiology of community-onset bloodstream infections. *Clin Microbiol Rev*, 27 (4):647–664, 2014. ISSN 0893-8512. doi: 10.1128/CMR.00002-14.

[2] Claudio Viscoli. Bloodstream infections: The peak of the iceberg. *Virulence*, 7(3): 248–251, 2016. ISSN 2150-5594. doi: 10.1080/21505594.2016.1152440.

[3] James M. M. D. MSc O'Brien, Naeem A. M. D. Ali, Scott K. M. D. M. P. H. Aberegg, and Edward M. D. Abraham. Sepsis. *Am J Med*, 120(12):1012–1022, 2007. ISSN 0002-9343. doi: 10.1016/j.amjmed.2007.01.035.

[4] G. S. Martin. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. *Expert Rev Anti Infect Ther*, 10(6):701–6, 2012. ISSN 1478-7210 (Print) 1478-7210. doi: 10.1586/eri.12.50.

[5] Mark Verway, Kevin A. Brown, Alex Marchand-Austin, Christina Diong, Samantha Lee, Bradley Langford, Kevin L. Schwartz, Derek R. MacFadden, Samir N. Patel, Beate Sander, Jennie Johnstone, Gary Garber, and Nick Daneman. Prevalence and mortality associated with bloodstream organisms: a population-wide retrospective cohort study. *J Clin Microbiol*, 60(4):e0242921–e0242921, 2022. ISSN 0095-1137. doi: 10.1128/jcm.02429-21.

[6] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J. D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J. L. Vincent, and D. C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–10, 2016. ISSN 0098-7484 (Print) 0098-7484. doi: 10.1001/jama.2016.0287.

[7] James A. Russell. Management of sepsis. *New England Journal of Medicine*, 355 (16):1699–1713, 2006. doi: 10.1056/NEJMra043632. URL `https://www.nejm.org/doi/full/10.1056/NEJMra043632`.

[8] Lars Ljungström, Rune Andersson, and Gunnar Jacobsson. Incidences of community onset severe sepsis, sepsis-3 sepsis, and bacteremia in sweden - a prospective population-based study. *PLoS One*, 14(12):e0225700–e0225700, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0225700.

[9] David Yu, David Unger, Christian Unge, Åsa Parke, Jonas Sundén-Cullberg, Kristoffer Strålin, and Volkan Özenci. Correlation of clinical sepsis definitions with microbiological characteristics in patients admitted through a sepsis alert system; a prospective cohort study. *Annals of Clinical Microbiology and Antimicrobials*, 21(1):7, 2022. ISSN 1476-0711. doi: 10.1186/s12941-022-00498-3. URL `https://doi.org/10.1186/s12941-022-00498-3`.

[10] Kristina E. Rudd, Sarah Charlotte Johnson, Kareha M. Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V. Colombara, Kevin S. Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R. Machado, Konrad K. Reinhart, Kathryn Rowan, Christopher W. Seymour, R. Scott Watson, T. Eoin West, Fatima Marinho, Simon I. Hay, Rafael Lozano, Alan D. Lopez, Derek C. Angus, Christopher J. L. Murray, and Mohsen Naghavi. Global, re-

gional, and national sepsis incidence and mortality, 1990x2013;2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020. ISSN 0140-6736. doi: 10.1016/S0140-6736(19)32989-7. URL https://doi.org/10.1016/S0140-6736(19)32989-7.

[11] Luis E Huerta and Todd W Rice. Pathologic difference between sepsis and bloodstream infections. *The Journal of Applied Laboratory Medicine*, 3(4):654–663, 2019. ISSN 2576-9456. doi: 10.1373/jalm.2018.026245. URL https://doi.org/10.1373/jalm.2018.026245.

[12] Michael T. Madigan, Michael T. Madigan, Kelly S. Bender, Daniel H. Buckley, W. Matthew Sattley, David A. Stahl, and Thomas D. Brock. *Brock biology of microorganisms*. Biology of microorganisms. Pearson, NY, NY, fifteenth edition.; global edition. edition, 2019. ISBN 9781292235103,1292235101.

[13] Franklin D. Lowy. Medical progress: Staphylococcus aureus infections. *The New England journal of medicine*, 339(8):520–532, 1998. ISSN 0028-4793. doi: 10.1056/NEJM199808203390806.

[14] G. Licitra. Etymologia: Staphylococcus. *Emerg Infect Dis*, 19(9):1553, 2013. ISSN 1080-6040 (Print). doi: 10.3201/eid1909.ET1909.

[15] Steven Y. C. Tong, Joshua S. Davis, Emily Eichenberger, Thomas L. Holland, and Vance G. Fowler. Staphylococcus aureus infections: Epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev*, 28(3):603–661, 2015. ISSN 0893-8512. doi: 10.1128/CMR.00134-14.

[16] Karsten Becker. *Chapter 2 - Pathogenesis of Staphylococcus aureus*, pages 13–38. Elsevier Inc, 2018. ISBN 9780128096710,9780128097977,0128097973,0128096713. doi: 10.1016/B978-0-12-809671-0.00002-4.

[17] Steven Y. C. Tong, Joshua S. Davis, Emily Eichenberger, Thomas L. Holland, and Vance G. Fowler. Staphylococcus aureus infections: Epidemiology, pathophysiology, clinical manifestations, and management. *Clinical Microbiology Reviews*, 28(3):603–661, 2015. doi: doi:10.1128/CMR.00134-14. URL https://journals.asm.org/doi/abs/10.1128/CMR.00134-14.

[18] K. B. Laupland, O. Lyytikäinen, M. Sgaard, K. J. Kennedy, J. D. Knudsen, C. Ostergaard, J. C. Galbraith, L. Valiquette, G. Jacobsson, P. Collignon, H. C. Schnheyder, and Collaborative for the International Bacteremia Surveillance. The changing epidemiology of staphylococcus aureus bloodstream infection: a multinational population-based surveillance study. *Clin Microbiol Infect*, 19(5):465–471, 2013. ISSN 1198-743X. doi: 10.1111/j.1469-0691.2012.03903.x.

[19] Daniel J. Diekema, Po-Ren Hsueh, Rodrigo E. Mendes, Michael A. Pfaller, Kenneth V. Rolston, Helio S. Sader, and Ronald N. Jones. The microbiology of bloodstream infection: 20-year trends from the sentry antimicrobial surveillance program. *Antimicrob Agents Chemother*, 63(7), 2019. ISSN 0066-4804. doi: 10.1128/AAC.00355-19.

[20] S. J. van Hal, S. O. Jensen, V. L. Vaska, B. A. Espedido, D. L. Paterson, and I. B. Gosbell. Predictors of mortality in staphylococcus aureus bacteremia. *Clin Microbiol Rev*, 25(2):362–86, 2012. ISSN 0893-8512 (Print) 0893-8512. doi: 10.1128/cmr.05022-11.

[21] Heiman F. L. Wertheim, Damian C. Melles, Margreet C. Vos, Willem van Leeuwen, Alex van Belkum, Henri A. Verbrugh, and Jan L. Nouwen. The role of nasal carriage in staphylococcus aureus infections. *Lancet Infect Dis*, 5(12):751–762, 2005. ISSN 1473-3099. doi: 10.1016/S1473-3099(05)70295-4.

[22] Alex van Belkum, Nelianne J. Verkaik, Corné P. de Vogel, Hélène A. Boelens, Jeroen Verveer, Jan L. Nouwen, Henri A. Verbrugh, and Heiman F. L. Wertheim. Reclassification of staphylococcus aureus nasal carriage types. *The Journal of Infectious Diseases*, 199(12):1820–1826, 2009. ISSN 0022-1899. doi: 10.1086/599119. URL `https://doi.org/10.1086/599119`.

[23] N. H. Eriksen, F. Espersen, V. T. Rosdahl, and K. Jensen. Carriage of staphylococcus aureus among 104 healthy persons during a 19-month period. *Epidemiol Infect*, 115(1):51–60, 1995. ISSN 0950-2688 (Print) 0950-2688. doi: 10.1017/s0950268800058118.

[24] Henry F. Chambers and Frank R. DeLeo. Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nature Reviews Microbiology*, 7(9):629–641, 2009. ISSN 1740-1534. doi: 10.1038/nrmicro2200. URL `https://doi.org/10.1038/nrmicro2200`.

[25] Aurelia Kimmig, Stefan Hagel, Sebastian Weis, Christina Bahrs, Bettina Löffler, and Mathias W. Pletz. Management of staphylococcus aureus bloodstream infections. *Frontiers in Medicine*, 7, 2021. ISSN 2296-858X. doi: 10.3389/fmed.2020.616524. URL `https://www.frontiersin.org/articles/10.3389/fmed.2020.616524`.

[26] Lena Thomer, Olaf Schneewind, and Dominique Missiakas. Pathogenesis of staphylococcus aureus bloodstream infections. *Annual Review of Pathology*, 11:343–364, 2016. doi: 10.1146/annurev-pathol-012615-044351.

[27] Andrzej Mynarczyk, Grażyna Mynarczyk, and Janusz Jeljaszewicz. The genome of staphylococcus aureus: A review. *Zentralbl Bakteriol*, 287(4):277–314, 1998. ISSN 0934-8840. doi: 10.1016/S0934-8840(98)80165-5.

[28] A. Muto and S. Osawa. The guanine and cytosine content of genomic dna and bacterial evolution. *Proc Natl Acad Sci U S A*, 84(1):166–169, 1987. ISSN 0027-8424. doi: 10.1073/pnas.84.1.166.

[29] Sang-Cheol Park, Kihyun Lee, Yeong Ouk Kim, Sungho Won, and Jongsik Chun. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Frontiers in Microbiology*, 10, 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.00834. URL `https://www.frontiersin.org/articles/10.3389/fmicb.2019.00834`.

[30] Yan Chen, Lu Sun, Dandan Wu, Haiping Wang, Shujuan Ji, and Yunsong Yu. Using core-genome multilocus sequence typing to monitor the changing epidemiology of methicillin-resistant staphylococcus aureus in a teaching hospital. *Clinical Infectious Diseases*, 67(suppl2):S241–S248, 2018. ISSN 1058-4838. doi: 10.1093/cid/ciy644. URL `https://doi.org/10.1093/cid/ciy644`.

[31] C. Montelongo, C. R. Mores, C. Putonti, A. J. Wolfe, and A. Abouelfetouh. Whole-genome sequencing of staphylococcus aureus and staphylococcus haemolyticus clinical isolates from egypt. *Microbiol Spectr*, 10(4):e0241321, 2022. ISSN 2165-0497. doi: 10.1128/spectrum.02413-21.

[32] E. Bosi, J. M. Monk, R. K. Aziz, M. Fondi, V. Nizet, and BØ Palsson. Comparative genome-scale modelling of staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc Natl Acad Sci U S A*, 113(26): E3801–9, 2016. ISSN 0027-8424 (Print) 0027-8424. doi: 10.1073/pnas.1523199113.

[33] S. J. Ho Sui, A. Fedynak, W. W. Hsiao, M. G. Langille, and F. S. Brinkman. The association of virulence factors with genomic islands. *PLoS One*, 4(12):e8094, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0008094.

[34] Laura S. Frost, Raphael Leplae, Anne O. Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005. ISSN 1740-1534. doi: 10.1038/nrmicro1235. URL `https://doi.org/10.1038/nrmicro1235`.

[35] Ying Jian and Min Li. A narrative review of single-nucleotide polymorphism detection methods and their application in studies of staphylococcus aureus. *Journal of Bio-X Research*, 4(1):1–9, 2021. ISSN 2096-5672. doi: 10.1097/jbr.0000000000000071. URL `https://journals.lww.com/jbioxresearch/Fulltext/2021/03000/A_narrative_review_of_single_nucleotide.1.aspx`.

[36] David L. Nelson, Albert L. Lehninger, and Michael M. Cox. *Lehninger principles of biochemistry*. Principles of biochemistry. W.H. Freeman, New York, 7th int. ed. edition, 2017. ISBN 9781319108243,1-319-10824-5.

[37] J. W. Peterson. *Bacterial Pathogenesis*, book section Chapter 7. Galveston (TX): University of Texas Medical Branch at Galveston, 4th edition edition, 1996. URL `https://www.ncbi.nlm.nih.gov/books/NBK8526/`.

[38] Michael E. Powers and Juliane Bubeck Wardenburg. Igniting the fire: Staphylococcus aureus virulence factors in the pathogenesis of sepsis. *PLoS Pathog*, 10(2): e1003871–e1003871, 2014. ISSN 1553-7366,1553-7374. doi: 10.1371/journal.ppat.1003871.

[39] Jakub M. Kwiecinski and Alexander R. Horswill. Staphylococcus aureus bloodstream infections: pathogenesis and regulatory mechanisms. *Curr Opin Microbiol*, 53:51–60, 2020. ISSN 1369-5274. doi: 10.1016/j.mib.2020.02.005.

[40] Timothy J. Foster. The mscramm family of cell-wall-anchored surface proteins of gram-positive cocci. *Trends in Microbiology*, 27(11):927–941, 2019. ISSN 0966-842X. doi: https://doi.org/10.1016/j.tim.2019.06.007. URL `https://www.sciencedirect.com/science/article/pii/S0966842X19301623`.

[41] Timothy J. Foster, Joan A. Geoghegan, Vannakambadi K. Ganesh, and Magnus Höök. Adhesion, invasion and evasion: The many functions of the surface proteins of staphylococcus aureus. *Nat Rev Microbiol*, 12(1):49–62, 2014. ISSN 1740-1526. doi: 10.1038/nrmicro3161.

[42] Clement Ajayi, Espen Åberg, Fatemeh Askarian, Johanna U. E. Sollid, Mona Johannessen, and Anne-Merethe Hanssen. Genetic variability in the sdrd gene in staphylococcus aureus from healthy nasal carriers. *BMC Microbiol*, 18(1):34–34, 2018. ISSN 1471-2180. doi: 10.1186/s12866-018-1179-7.

[43] L. J. Shallcross, E. Fragaszy, A. M. Johnson, and A. C. Hayward. The role of the panton-valentine leucocidin toxin in staphylococcal disease: a systematic review

and meta-analysis. *Lancet Infect Dis*, 13(1):43–54, 2013. ISSN 1473-3099 (Print) 1473-3099. doi: 10.1016/s1473-3099(12)70238-4.

[44] James A. Russell. Management of sepsis. *New England Journal of Medicine*, 355 (16):1699–1713, 2006. doi: 10.1056/NEJMra043632. URL https://www.nejm.org/doi/full/10.1056/NEJMra043632.

[45] Yunlei Guo, Guanghui Song, Meiling Sun, Juan Wang, and Yi Wang. Prevalence and therapies of antibiotic-resistance in staphylococcus aureus. *Frontiers in Cellular and Infection Microbiology*, 10, 2020. ISSN 2235-2988. doi: 10.3389/fcimb.2020.00107. URL https://www.frontiersin.org/articles/10.3389/fcimb.2020.00107.

[46] Andie S. Lee, Hermínia de Lencastre, Javier Garau, Jan Kluytmans, Surbhi Malhotra-Kumar, Andreas Peschel, and Stephan Harbarth. Methicillin-resistant staphylococcus aureus. *Nature Reviews Disease Primers*, 4(1):18033, 2018. ISSN 2056-676X. doi: 10.1038/nrdp.2018.33. URL https://doi.org/10.1038/nrdp.2018.33.

[47] Nicholas A. Turner, Batu K. Sharma-Kuinkel, Stacey A. Maskarinec, Emily M. Eichenberger, Pratik P. Shah, Manuela Carugati, Thomas L. Holland, and Vance G. Fowler. Methicillin-resistant staphylococcus aureus: an overview of basic and clinical research. *Nature Reviews Microbiology*, 17(4):203–218, 2019. ISSN 1740-1534. doi: 10.1038/s41579-018-0147-4. URL https://doi.org/10.1038/s41579-018-0147-4.

[48] Y. Cong, S. Yang, and X. Rao. Vancomycin resistant staphylococcus aureus infections: A review of case updating and clinical features. *J Adv Res*, 21:169–176, 2020. ISSN 2090-1232 (Print) 2090-1224. doi: 10.1016/j.jare.2019.10.005.

[49] David Chinemerem Nwobodo, Malachy Chigozie Ugwu, Clement Oliseloke Anie, Mushtak T. S. Al-Ouqaili, Joseph Chinedu Ikem, Uchenna Victor Chigozie, and Morteza Saki. Antibiotic resistance: The challenges and some emerging strategies for tackling a global menace. *Journal of clinical laboratory analysis*, 36(9):e24655–n/a, 2022. ISSN 0887-8013. doi: 10.1002/jcla.24655.

[50] Martin C. J. Maiden, Jane A. Bygraves, Edward Feil, Giovanna Morelli, Joanne E. Russell, Rachel Urwin, Qing Zhang, Jiaji Zhou, Kerstin Zurth, Dominique A. Caugant, Ian M. Feavers, Mark Achtman, and Brian G. Spratt. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*, 95(6):3140–3145, 1998. ISSN 0027-8424. doi: 10.1073/pnas.95.6.3140.

[51] Keith A. Jolley, James E. Bray, and Martin C. J. Maiden. Open-access bacterial population genomics: Bigsdb software, the pubmlst.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Res*, 3:124–124, 2018. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.14826.1.

[52] Mark C. Enright, Nicholas P. J. Day, Catrin E. Davies, Sharon J. Peacock, and Brian G. Spratt. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of staphylococcus aureus. *J Clin Microbiol*, 38(3):1008–1015, 2000. ISSN 0095-1137. doi: 10.1128/jcm.38.3.1008-1015.2000.

[53] Edward J. Feil, Bao C. Li, David M. Aanensen, William P. Hanage, and Brian G. Spratt. eburst: Inferring patterns of evolutionary descent among clusters of related

bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, 186(5):1518–1530, 2004. doi: doi:10.1128/JB.186.5.1518-1530.2004. URL https://journals.asm.org/doi/abs/10.1128/JB.186.5.1518-1530.2004.

[54] H. M. E. Frénay, A. E. Bunschoten, L. M. Schouls, W. J. Van Leeuwen, C. M. J. E. Vandenbroucke-Grauls, J. Verhoef, and F. R. Mooi. Molecular typing of methicillin-resistant staphylococcus aureus on the basis of protein a gene polymorphism. *Eur J Clin Microbiol Infect Dis*, 15(1):60–64, 1996. ISSN 0934-9723. doi: 10.1007/BF01586186.

[55] Larry Koreen, Srinivas V. Ramaswamy, Edward A. Graviss, Steven Naidich, James M. Musser, and Barry N. Kreiswirth. spa typing method for discriminating among staphylococcus aureus isolates: Implications for use of a single marker to detect genetic micro- and macrovariation. *Journal of Clinical Microbiology*, 42(2):792–799, 2004. doi: doi:10.1128/JCM.42.2.792-799.2004. URL https://journals.asm.org/doi/abs/10.1128/JCM.42.2.792-799.2004.

[56] Eugene V. Koonin, Kira S. Makarova, and Yuri I. Wolf. Evolution of microbial genomics: Conceptual shifts over a quarter century. *Trends in Microbiology*, 29(7):582–592, 2021. ISSN 0966-842X. doi: https://doi.org/10.1016/j.tim.2021.01.005. URL https://www.sciencedirect.com/science/article/pii/S0966842X2100007X.

[57] Charles Y. Chiu and Steven A. Miller. Clinical metagenomics. *Nature Reviews Genetics*, 20(6):341–355, 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0113-7. URL https://doi.org/10.1038/s41576-019-0113-7.

[58] S. Behjati and P. S. Tarpey. What is next generation sequencing? *Arch Dis Child Educ Pract Ed*, 98(6):236–8, 2013. ISSN 1743-0585 (Print) 1743-0585. doi: 10.1136/archdischild-2013-304340.

[59] Stuart M. Brown. *Next-generation DNA sequencing informatics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y, 2013. ISBN 9781936113873.

[60] An introduction to next-generation sequencing technology. 2017. URL https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

[61] Xingyu Liao, Min Li, You Zou, Fang-Xiang Wu, Pan Yi, and Jianxin Wang. Current challenges and solutions of de novo assembly. *Quantitative Biology*, 7(2):90–109, 2019. ISSN 2095-4697. doi: 10.1007/s40484-019-0166-9. URL https://doi.org/10.1007/s40484-019-0166-9.

[62] P. E. Compeau, P. A. Pevzner, and G. Tesler. How to apply de bruijn graphs to genome assembly. *Nat Biotechnol*, 29(11):987–91, 2011. ISSN 1087-0156 (Print) 1087-0156. doi: 10.1038/nbt.2023.

[63] A. van Belkum, S. Scherer, L. van Alphen, and H. Verbrugh. Short-sequence dna repeats in prokaryotic genomes. *Microbiol Mol Biol Rev*, 62(2):275–93, 1998. ISSN 1092-2172 (Print) 1092-2172. doi: 10.1128/mmbr.62.2.275-293.1998.

[64] O. K. Tørresen, B. Star, P. Mier, M. A. Andrade-Navarro, A. Bateman, P. Jarnot, A. Gruca, M. Grynberg, A. V. Kajava, V. J. Promponas, M. Anisimova, K. S. Jakobsen, and D. Linke. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res*,

47(21):10994–11006, 2019. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkz841.

[65] Fassler J. and Cooper P. Blast® help [internet]. *Bethesda (MD): National Center for Biotechnology Information (US); 2008-*, 2011. URL https://www.ncbi.nlm.nih.gov/books/NBK62051/.

[66] Choudhuri Supratim. *Chapter 2 Fundamentals of Molecular Evolution*, pages 27–53. Elsevier Inc, 2014. ISBN 9780124104716,0124104711,0124105106,9780124105102. doi: 10.1016/B978-0-12-410471-6.00002-5.

[67] Jonathan B. Losos. Seeing the forest for the trees: The limitations of phylogenies in comparative biology. *Am Nat*, 177(6):709–727, 2011. ISSN 0003-0147. doi: 10.1086/660020.

[68] L. Gabora. *Convergent Evolution*, pages 178–180. Academic Press, San Diego, 2013. ISBN 978-0-08-096156-9. doi: https://doi.org/10.1016/B978-0-12-374984-0.00336-3. URL https://www.sciencedirect.com/science/article/pii/B9780123749840003363.

[69] Mikkel H Schierup and Jotun Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–891, 2000. ISSN 1943-2631. doi: 10.1093/genetics/156.2.879. URL https://doi.org/10.1093/genetics/156.2.879.

[70] Claudia Augusta de Moraes Russo and Alexandre Pedro Selvatti. Bootstrap and rogue identification tests for phylogenetic analyses. *Molecular Biology and Evolution*, 35(9):2327–2333, 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy118. URL https://doi.org/10.1093/molbev/msy118.

[71] Kevin de Queiroz. Nodes, branches, and phylogenetic definitions. *Syst Biol*, 62(4):625–632, 2013. ISSN 1063-5157. doi: 10.1093/sysbio/syt027.

[72] Sandra L. Baldauf. Phylogeny for the faint of heart: a tutorial. *Trends Genet*, 19(6):345–351, 2003. ISSN 0168-9525. doi: 10.1016/S0168-9525(03)00112-4.

[73] Yu-Chen Hu, Shailesh Tiwari, Krishn K. Mishra, and Munesh C. Trivedi. *Phylogenetics Algorithms and Applications*, volume 904 of *Advances in Intelligent Systems and Computing*, pages 187–194. Singapore: Springer Singapore Pte. Limited, Singapore, 2019. ISBN 2194-5357 9811359334,9789811359330. doi: 10.1007/978-981-13-5934-7_17.

[74] Miguel Arenas. Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 6, 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00319. URL https://www.frontiersin.org/articles/10.3389/fgene.2015.00319.

[75] Peter K. Dunn and Gordon K. Smyth. *Beyond Linear Regression: The Method of Maximum Likelihood*. United States: Springer New York, United States, 2018. ISBN 1441901175,9781441901170.

[76] Guohua Jin, Luay Nakhleh, Sagi Snir, and Tamir Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 23(8):1046–1047, 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm075.

[77] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and

Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9. URL https://doi.org/10.1038/s43586-021-00056-9.

[78] Shiro Ikegawa. A short history of the genome-wide association study: Where we were and where we are going. *Genomics informatics*, 10(4):220–225, 2012. ISSN 2234-0742,1598-866X.

[79] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. Microbial genome-wide association studies: lessons from human gwas. *Nat Rev Genet*, 18(1):41–50, 2017. ISSN 1471-0056. doi: 10.1038/nrg.2016.132.

[80] James Emmanuel San, Shakuntala Baichoo, Aquillah Kanzi, Yumna Moosa, Richard Lessells, Vagner Fonseca, John Mogaka, Robert Power, and Tulio de Oliveira. Current affairs of microbial genome-wide association studies: Approaches, bottlenecks and analytical pitfalls. *Front Microbiol*, 10:3119–3119, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2019.03119.

[81] Alita R. Burmeister. Horizontal gene transfer. *Evol Med Public Health*, 2015(1): 193–194, 2015. ISSN 2050-6201. doi: 10.1093/emph/eov018.

[82] Ola Brynildsrud, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. Rapid scoring of genes in microbial pan-genome-wide association studies with scoary. *Genome Biol*, 17(1):238–238, 2016. ISSN 1474-7596,1474-760X. doi: 10.1186/s13059-016-1108-8.

[83] Michael H. Herzog, Gregory Francis, and Aaron Clarke. *The Multiple Testing Problem*, pages 63–66. Springer International Publishing, Cham, 2019. ISBN 978-3-030-03499-3. doi: 10.1007/978-3-030-03499-3_5. URL https://doi.org/10.1007/978-3-030-03499-3_5.

[84] Winston Haynes. *Benjamini–Hochberg Method*, pages 78–78. Springer New York, New York, NY, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_1215. URL https://doi.org/10.1007/978-1-4419-9863-7_1215.

[85] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*, 56 (1):45–50, 2008. ISSN 0301-4738 (Print) 0301-4738. doi: 10.4103/0301-4738.37595.

[86] Geir Sverre Braut and Sigmund Grønmo. *odds ratio*. Store norske leksikon. snl.no, 2020. URL https://snl.no/odds_ratio.

[87] M. Szumilas. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, 19(3): 227–9, 2010. ISSN 1719-8429 (Print) 1719-8429.

[88] Bjarne K. Jacobsen, Anne Elise Eggen, Ellisiv B. Mathiesen, Tom Wilsgaard, and Inger Njølstad. Cohort profile: The tromsø study. *Int J Epidemiol*, 41(4):961–967, 2012. ISSN 0300-5771. doi: 10.1093/ije/dyr049.

[89] Anne Elise Eggen, Ellisiv B. Mathiesen, Tom Wilsgaard, Bjarne K. Jacobsen, and Inger Njølstad. The sixth survey of the tromso study (tromse 6) in 2007-08: Collaborative research in the interface between clinical medicine and epidemiology: Study objectives, design, data collection procedures, and attendance in a multipurpose population-based health survey. *Scand J Public Health*, 41(1):65–80, 2013. ISSN 1403-4948. doi: 10.1177/1403494812469851.

[90] Registre i helse nord-trøndelag. *hnt.no*, 2020. URL https://hnt.no/helsefaglig/forskning/forsknings-og-kvalitetsregistre#hva-brukes-registeret-til.

[91] Nasjonalt referanselaboratorium mrsa. *https://stolavs.no*, 2022. URL https://stolav.no/fag-og-forskning/lab/nasjonalt-referanselaboratorium-mrsa.

[92] Illumina dna prep reference guide. *Illumina.com*, 2022. URL https://support.illumina.com/sequencing/sequencing_kits/illumina-dna-prep/documentation.html.

[93] Miseq system denature and dilute libraries guide. *Illumina.com*, 2019. URL https://support.illumina.com/downloads/prepare_libraries_for_sequencing_miseq_15039740.html.

[94] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res*, 8(3):186–94, 1998. ISSN 1088-9051 (Print) 1088-9051.

[95] S. Andrews. Fastqc. a quality control tool for high throughput sequence data. 2010. URL https://www.bibsonomy.org/bibtex/2b6052877491828ab53d3449be9b293b3/ozborn.

[96] Chen Shifu, Zhou Yanqing, Chen Yaru, and Gu Jia. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018. doi: https://doi.org/10.1093/bioinformatics/bty560.

[97] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. ISSN 1367-4803,1367-4811. doi: 10.1093/bioinformatics/btw354.

[98] T Seemann, A Goncalves da Silva, DM Bulach, MB Schultz, JC Kwong, and BP Howden. Nullarbor. *Github*. URL https://github.com/tseemann/nullarbor.

[99] Staphylococcus aureus subsp. aureus strain atcc 25923, complete genome. *European Nucleotide Archive*. URL https://www.ebi.ac.uk/ena/browser/view/CP009361?dataType=&show=publications.

[100] Alexandre Souvorov, Richa Agarwala, and David J. Lipman. Skesa: Strategic k-mer extension for scrupulous assemblies. *Genome Biol*, 19(1):153–153, 2018. ISSN 1474-7596,1474-760X. doi: 10.1186/s13059-018-1540-z.

[101] Torsten Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu153.

[102] Derrick E. Wood and Steven L. Salzberg. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*, 15(3):R46–R46, 2014. ISSN 1474-760X,1465-6906. doi: 10.1186/gb-2014-15-3-r46.

[103] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biology*, 20(1):1–257, 2019. ISSN 1474-7596,1474-760X. doi: 10.1186/s13059-019-1891-0.

[104] Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res*, 26(12):1721–1729, 2016. ISSN 1088-9051. doi: 10.1101/gr.210641.116.

[105] T Seemann. mlst. *Github*. URL https://github.com/tseemann/mlst.

[106] PubMLST. URL https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=downloadProfiles&scheme_id=1.

[107] T Seeman. Abricate. *Github*. URL https://github.com/tseemann/abricate.

[108] Alfred Ferrer Florensa, Rolf Sommer Kaas, Philip Thomas Lanken Conradsen Clausen, Derya Aytan-Aktug, and Frank M. Aarestrup. Resfinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom*, 8 (1), 2022. ISSN 2057-5858. doi: 10.1099/mgen.0.000748.

[109] Bo Liu, Dandan Zheng, Siyu Zhou, Lihong Chen, and Jian Yang. Vfdb 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Res*, 50 (1):D912–D917, 2022. ISSN 0305-1048,1362-4962. doi: 10.1093/nar/gkab1107.

[110] M Sullivan and JF Sanchez-Herrero. spatyper: Staphylococcal protein a (spa) characterization pipeline. URL https://github.com/HCGB-IGTP/spaTyper/tree/v0.3.1.

[111] Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian. Parkhill. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 2015. doi: http://dx.doi.org/10.1093/bioinformatics/btv421.

[112] Morgan N. Price, P.S. Dehal, and A.P. Arkin. Fast tree: Computing large minimum-evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26:1641–1650, 2009. doi: 10.1093/molbev/msp077.

[113] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490–e9490, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009490.

[114] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*, 49(W1):W293–W296, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab301.

[115] Koichiro Tamura, Glen Stecher, and Sudhir Kumar. Mega11: Molecular evolutionary genetics analysis version 11. *Mol Biol Evol*, 38(7):3022–3027, 2021. ISSN 0737-4038,1537-1719. doi: 10.1093/molbev/msab120.

[116] Geneious prime 2022.2.2. URL https://www.geneious.com.

[117] Posit team. *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA, 2023. URL http://www.posit.co/.

[118] Nicholas A. Turner, Batu K. Sharma-Kuinkel, Stacey A. Maskarinec, Emily M. Eichenberger, Pratik P. Shah, Manuela Carugati, Thomas L. Holland, and Vance G. Fowler. Methicillin-resistant staphylococcus aureus: an overview of basic and clinical research. *Nature Reviews Microbiology*, 17(4):203–218, 2019. ISSN 1740-1534. doi: 10.1038/s41579-018-0147-4. URL https://doi.org/10.1038/s41579-018-0147-4.

[119] Mark C. Enright, D. Ashley Robinson, Gaynor Randle, Edward J. Feil, Hajo Grundmann, and Brian G. Spratt. The evolutionary history of methicillin-resistant *Staphy-*

lococcus aureus (mrsa). *Proceedings of the National Academy of Sciences*, 99(11): 7687–7692, 2002. doi: doi:10.1073/pnas.122108599. URL https://www.pnas.org/doi/abs/10.1073/pnas.122108599.

[120] R. H. Deurenberg, C. Vink, G. J. Oudhuis, J. E. Mooij, C. Driessen, G. Coppens, J. Craeghs, E. De Brauwer, S. Lemmen, H. Wagenvoort, A. W. Friedrich, J. Scheres, and E. E. Stobberingh. Different clonal complexes of methicillin-resistant staphylococcus aureus are disseminated in the euregio meuse-rhine region. *Antimicrob Agents Chemother*, 49(10):4263–71, 2005. ISSN 0066-4804 (Print) 0066-4804. doi: 10.1128/aac.49.10.4263-4271.2005.

[121] Serena Manara, Edoardo Pasolli, Daniela Dolce, Novella Ravenni, Silvia Campana, Federica Armanini, Francesco Asnicar, Alessio Mengoni, Luisa Galli, Carlotta Montagnani, Elisabetta Venturini, Omar Rota-Stabelli, Guido Grandi, Giovanni Taccetti, and Nicola Segata. Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of staphylococcus aureus strains in a paediatric hospital. *Genome Medicine*, 10(1):82, 2018. ISSN 1756-994X. doi: 10.1186/s13073-018-0593-7. URL https://doi.org/10.1186/s13073-018-0593-7.

[122] David M. Aanensen, Edward J. Feil, Matthew T. G. Holden, Janina Dordel, Corin A. Yeats, Artemij Fedosejev, Richard Goater, Santiago Castillo-Ramírez, Jukka Corander, Caroline Colijn, Monika A. Chlebowicz, Leo Schouls, Max Heck, Gerlinde Pluister, Raymond Ruimy, Gunnar Kahlmeter, Jenny Åhman, Erika Matuschek, Alexander W. Friedrich, Julian Parkhill, Stephen D. Bentley, Brian G. Spratt, and Hajo Grundmann. Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive staphylococcus aureus in europe. *mBio*, 7(3):e00444–16, 2016. doi: doi:10.1128/mBio.00444-16. URL https://journals.asm.org/doi/abs/10.1128/mBio.00444-16.

[123] Joana Rolo, Maria Miragaia, Agata Turlej-Rogacka, Joanna Empel, Bouchami Ons, Nuno A. Faria, Ana Tavares, Waleria Hryniewicz, Ad C. Fluit, Lencastre Hermínia de, and Concord Working Group the. High genetic diversity among community-associated staphylococcus aureus in europe: Results from a multicenter study. *PLoS One*, 7(4), 2012. doi: https://doi.org/10.1371/journal.pone.0034768.

[124] X. Wang, D. Lin, Z. Huang, J. Zhang, W. Xie, P. Liu, H. Jing, and J. Wang. Clonality, virulence genes, and antibiotic resistance of staphylococcus aureus isolated from blood in shandong, china. *BMC Microbiol*, 21(1):281, 2021. ISSN 1471-2180. doi: 10.1186/s12866-021-02344-6.

[125] Anita J. Campbell, Shakeel Mowlaboccus, Geoffrey W. Coombs, Denise A. Daley, Laila S. Al Yazidi, Linny K. Phuong, Clare Leung, Emma J. Best, Rachel H. Webb, Lesley Voss, Eugene Athan, Philip N. Britton, Penelope A. Bryant, Coen T. Butters, Jonathan R. Carapetis, Natasha S. Ching, Joshua Francis, Te-Yu Hung, Clare Nourse, Samar Ojaimi, Alex Tai, Nan Vasilunas, Brendan McMullan, Asha C. Bowen, and Christopher C. Blyth. Whole genome sequencing and molecular epidemiology of paediatric staphylococcus aureus bacteraemia. *Journal of Global Antimicrobial Resistance*, 29:197–206, 2022. ISSN 2213-7165. doi: https://doi.org/10.1016/j.jgar.2022.03.012. URL https://www.sciencedirect.com/science/article/pii/S2213716522000649.

[126] A. Sabat, D. C. Melles, G. Martirosian, H. Grundmann, A. van Belkum, and

W. Hryniewicz. Distribution of the serine-aspartate repeat protein-encoding sdr genes among nasal-carriage and invasive staphylococcus aureus strains. *J Clin Microbiol*, 44(3):1135–8, 2006. ISSN 0095-1137 (Print) 0095-1137. doi: 10.1128/jcm.44.3.1135-1138.2006.

[127] Huanle Liu, Jingnan Lv, Xiuqin Qi, Yu Ding, Dan Li, Longhua Hu, Liangxing Wang, and Fangyou Yu. The carriage of the serine-aspartate repeat protein-encoding sdr genes among staphylococcus aureus lineages. *The Brazilian Journal of Infectious Diseases*, 19(5):498–502, 2015. ISSN 1413-8670. doi: https://doi.org/10.1016/j.bjid.2015.07.003. URL https://www.sciencedirect.com/science/article/pii/S1413867015001373.

[128] Ons Haddad, Abderrahmen Merghni, Aida Elargoubi, Hajer Rhim, Yosr Kadri, and Maha Mastouri. Comparative study of virulence factors among methicillin resistant staphylococcus aureus clinical isolates. *BMC Infectious Diseases*, 18(1):560, 2018. ISSN 1471-2334. doi: 10.1186/s12879-018-3457-2. URL https://doi.org/10.1186/s12879-018-3457-2.

[129] Xuehan Li, Tao Huang, Kai Xu, Chenglin Li, and Yirong Li. Molecular characteristics and virulence gene profiles of staphylococcus aureus isolates in hainan, china. *BMC Infectious Diseases*, 19(1):873, 2019. ISSN 1471-2334. doi: 10.1186/s12879-019-4547-5. URL https://doi.org/10.1186/s12879-019-4547-5.

[130] Dafne Pérez-Montarelo, Esther Viedma, Nieves Larrosa, Carmen Gómez-González, Enrique Ruiz de Gopegui, Irene Muñoz-Gallego, Rafael San Juan, Nuria Fernández-Hidalgo, Benito Almirante, and Fernando Chaves. Molecular epidemiology of staphylococcus aureus bacteremia: Association of molecular factors with the source of infection. *Frontiers in Microbiology*, 9, 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.02210. URL https://www.frontiersin.org/articles/10.3389/fmicb.2018.02210.

[131] Gunlög Rasmussen, Stefan Monecke, Ole Brus, Ralf Ehricht, and Bo Söderquist. Long term molecular epidemiology of methicillin-susceptible staphylococcus aureus bacteremia isolates in sweden. *PLOS ONE*, 9(12):e114276, 2014. doi: 10.1371/journal.pone.0114276. URL https://doi.org/10.1371/journal.pone.0114276.

[132] A. Blomfeldt, A. N. Eskesen, H. V. Aamot, T. M. Leegaard, and J. V. Bjørnholt. Population-based epidemiology of staphylococcus aureus bloodstream infection: clonal complex 30 genotype is associated with mortality. *European Journal of Clinical Microbiology Infectious Diseases*, 35(5):803–813, 2016. ISSN 1435-4373. doi: 10.1007/s10096-016-2601-4. URL https://doi.org/10.1007/s10096-016-2601-4.

[133] Bibi C. G. C. Slingerland, Margreet C. Vos, Willeke Bras, René F. Kornelisse, Dieter De Coninck, Alex van Belkum, Irwin K. M. Reiss, Wil H. F. Goessens, Corné H. W. Klaassen, and Nelianne J. Verkaik. Whole-genome sequencing to explore nosocomial transmission and virulence in neonatal methicillin-susceptible staphylococcus aureus bacteremia. *Antimicrobial Resistance Infection Control*, 9(1):39, 2020. ISSN 2047-2994. doi: 10.1186/s13756-020-0699-8. URL https://doi.org/10.1186/s13756-020-0699-8.

[134] G. Rasmussen, S. Monecke, R. Ehricht, and B. Söderquist. Prevalence of clonal complexes and virulence genes among commensal and invasive staphylococcus au-

reus isolates in sweden. *PLoS One*, 8(10):e77477, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0077477.

[135] A. Blomfeldt, H. V. Aamot, A. N. Eskesen, F. Müller, and S. Monecke. Molecular characterization of methicillin-sensitive staphylococcus aureus isolates from bacteremic patients in a norwegian university hospital. *J Clin Microbiol*, 51(1):345–7, 2013. ISSN 0095-1137 (Print) 0095-1137. doi: 10.1128/jcm.02571-12.

[136] Xuehan Li, Fang Fang, Jin Zhao, Ning Lou, Chenglin Li, Tao Huang, and Yirong Li. Molecular characteristics and virulence gene profiles of staphylococcus aureus causing bloodstream infection. *The Brazilian Journal of Infectious Diseases*, 22(6):487–494, 2018. ISSN 1413-8670. doi: https://doi.org/10.1016/j.bjid.2018.12.001. URL `https://www.sciencedirect.com/science/article/pii/S141386701830970X`.

[137] S. Holtfreter, D. Grumann, V. Balau, A. Barwich, J. Kolata, A. Goehler, S. Weiss, B. Holtfreter, S. S. Bauerfeind, P. Döring, E. Friebe, N. Haasler, K. Henselin, K. Kühn, S. Nowotny, D. Radke, K. Schulz, S. R. Schulz, P. Trübe, C. H. Vu, B. Walther, S. Westphal, C. Cuny, W. Witte, H. Völzke, H. J. Grabe, T. Kocher, I. Steinmetz, and B. M. Bröker. Molecular epidemiology of staphylococcus aureus in the general population in northeast germany: Results of the study of health in pomerania (ship-trend-0). *J Clin Microbiol*, 54(11):2774–2785, 2016. ISSN 0095-1137 (Print) 0095-1137. doi: 10.1128/jcm.00312-16.

[138] Jaishri Mehraj, Wolfgang Witte, Manas K. Akmatov, Franziska Layer, Guido Werner, and Gérard Krause. *Epidemiology of Staphylococcus aureus Nasal Carriage Patterns in the Community*, pages 55–87. Springer International Publishing, Cham, 2016. ISBN 978-3-319-49284-1. doi: 10.1007/82_2016_497. URL `https://doi.org/10.1007/82_2016_497`.

[139] Raymond Ruimy, Laurence Armand-Lefevre, Francois Barbier, Etienne Ruppé, Radu Cocojaru, Yasmine Mesli, Aminata Maiga, Mokhtar Benkalfat, Samia Benchouk, Hafida Hassaine, Jean-Baptiste Dufourcq, Chhor Nareth, Jean-Louis Sarthou, Antoine Andremont, and Edward J. Feil. Comparisons between geographically diverse samples of carried staphylococcus aureus. *Journal of Bacteriology*, 191(18):5577–5583, 2009. doi: doi:10.1128/JB.00493-09. URL `https://journals.asm.org/doi/abs/10.1128/JB.00493-09`.

[140] O. Sakwinska, G. Kuhn, C. Balmelli, P. Francioli, M. Giddey, V. Perreten, A. Riesen, F. Zysset, D. S. Blanc, and P. Moreillon. Genetic diversity and ecological success of staphylococcus aureus strains colonizing humans. *Appl Environ Microbiol*, 75(1): 175–83, 2009. ISSN 0099-2240 (Print) 0099-2240. doi: 10.1128/aem.01860-08.

[141] Maria Sangvik, Renate Slind Olsen, Karina Olsen, Gunnar Skov Simonsen, Anne-Sofie Furberg, and Johanna U. Ericson Sollid. Age- and gender-associated staphylococcus aureus spa types found among nasal carriers in a general population: the tromsxf8; staph and skin study. *Journal of Clinical Microbiology*, 49(12):4213–4218, 2011. doi: doi:10.1128/JCM.05290-11. URL `https://journals.asm.org/doi/abs/10.1128/JCM.05290-11`.

[142] S. J. Peacock, C. E. Moore, A. Justice, M. Kantzanou, L. Story, K. Mackie, G. O'Neill, and N. P. Day. Virulent combinations of adhesin and toxin genes in natural populations of staphylococcus aureus. *Infect Immun*, 70(9):4987–96, 2002. ISSN 0019-9567 (Print) 0019-9567. doi: 10.1128/iai.70.9.4987-4996.2002.

[143] Dimitar Nashev, Katia Toshkova, S.Isrina O. Salasia, Abdulwahed A. Hassan, Christoph Lämmler, and Michael Zschöck. Distribution of virulence genes of staphylococcus aureus isolated from stable nasal carriers. *FEMS Microbiology Letters*, 233(1):45–52, 2004. ISSN 0378-1097. doi: 10.1016/j.femsle.2004.01.032. URL https://doi.org/10.1016/j.femsle.2004.01.032.

[144] A. C. Senok, A. M. Somily, P. Slickers, M. A. Raji, G. Garaween, A. Shibl, S. Monecke, and R. Ehricht. Investigating a rare methicillin-resistant staphylococcus aureus strain: first description of genome sequencing and molecular characterization of cc15-mrsa. *Infect Drug Resist*, 10:307–315, 2017. ISSN 1178-6973 (Print) 1178-6973. doi: 10.2147/idr.S145394.

[145] L. M. O'Brien, E. J. Walsh, R. C. Massey, S. J. Peacock, and T. J. Foster. Staphylococcus aureus clumping factor b (clfb) promotes adherence to human type i cytokeratin 10: implications for nasal colonization. *Cell Microbiol*, 4(11):759–70, 2002. ISSN 1462-5814 (Print) 1462-5814. doi: 10.1046/j.1462-5822.2002.00231.x.

A Commands

This appendix lists all commands run during the bioinformatic analysis.

FastQC

```
>fastqc $input_fastq  -o $output
```

MultiQC

```
>multiqc $input -o $output --interactive
```

FastP

```
>fastp -i $input_forward_reads -I $input_reverse_reads
-o $output_paired_forward -O $output_paired_reverse
--unpaired1 $output_unpaired_forward --unpaired2
$output_unpaired_reverse --failed_out $failed_output
-j $json_output -h $html_out
```

Nullarbor

```
>nullarbor.pl --name $project_name --mlst $input_organism
--ref $input_reference_strain  --input $input_strains
--outdir $output_directory --run
```

SpaTyper

```
>spaTyper -f $input
```

Roary

```
>roary -e -n -f $output $input
```

FastTree

```
>FastTree -gtr -nt $input > $output
```

Roary plots

```
 >roary_plots.py $input_tree $input_gene_presence_absence
```

Scoary

```
>scoary -g $input_genes_presence_absence -t $input_trait
-n $input_tree -r $input_strains
```

Snippy

```
 >snippy --cpus 16 --outdir $output --ref $input_reference
 --R1 $input_forward_reads --R2 input_reverse_reads
```

Snippy core

```
>snippy-core --ref $input_reference -ver-prefix core
$input_strains
```

SNP tree

```
>snippy-clean_full_aln $input_snippy_core > $output_modified
>run_gubbins.py -p gubbins $input_modified
>snp-sites -c gubbins.filtered_polymorphic_sites.fasta
    > $output_clean_core
>FastTree -gtr -nt $input_clean_core > $output_tree
```

B Quality control data

This appendix shows data from the quality control including the number of reverse- and forward reads and base pairs before and after trimming, number of reads passing the quality filter and number of reads failing to pass the quality filter for all three cohorts BSI (B.1) MRSA (B.2) and TSSS (B.3).

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
STAU303	1570878	1570878	237202578	237202578	1497776	1497776	214945879	214946331	2995552	143958
STAU302	4269664	4269664	644719264	644719264	4132839	4132839	615336858	615336858	8265678	270292
STAU301	4773022	4773022	720726322	720726322	4697115	4697115	692497449	692497449	9394230	146906
STAU300	2314563	2314563	349499013	349499013	2262953	2262953	329323633	329323633	4525906	101380
STAU299	5086642	5086642	768082942	768082942	5012595	5012595	732884668	732884668	10025190	143872
STAU298	2556507	2556507	386032557	386032557	2516123	2516123	370443553	370443553	5032246	78682
STAU297	5061993	5061993	764360943	764360943	4972398	4972398	735967819	735967819	9944796	174946
STAU296	5524908	5524908	834261108	834261108	5356768	5356768	797379625	797379625	10713536	331814
STAU295	2241174	2241174	338417274	338417274	2200475	2200475	323616306	323616306	4400950	79660
STAU294	4193372	4193372	633199172	633199172	4108771	4108771	606861670	606861670	8217542	165742
STAU293	4654678	4654678	702856378	702856378	4579529	4579529	673230825	673230825	9159058	146384
STAU292	2851860	2851860	430630860	430630860	2785140	2785140	412028765	412028765	5570280	131168
STAU291	1712527	1712527	258591577	258591577	1636803	1636803	235712077	235711967	3273606	149212
STAU290	2119789	2119789	320088139	320088139	2107032	2107032	296854852	296854852	4214064	23786
STAU289	3923912	3923912	592510712	592510712	3867408	3867408	557711163	557711163	7734816	109308
STAU288	4514489	4514489	681687839	681687839	4441138	4441138	656474965	656474965	8882276	142740
STAU287	2560611	2560611	386652261	386652261	2531878	2531878	365888424	365888424	5063756	55284
STAU286	4196332	4196332	633646132	633646132	4102286	4102286	607764983	607764983	8204572	184834
STAU285	4094580	4094580	618281580	618281580	4023794	4023794	581900068	581900068	8047588	137588
STAU284	2806122	2806122	423724422	423724422	2761300	2761300	405554439	405554439	5522600	87310
STAU283	2238714	2238714	338045814	338045814	2144274	2144274	309550611	309550431	4288548	186138
STAU282	5105162	5105162	770879462	770879462	5027816	5027816	722796098	722796098	10055632	150316
STAU281	3143892	3143892	474727692	474727692	3052299	3052299	444102100	444102100	6104598	180632
STAU280	3874712	3874712	585081512	585081512	3810218	3810218	547383127	547383127	7620436	125436
STAU279	3894378	3894378	588051078	588051078	3832481	3832481	566630395	566630395	7664962	120446
STAU278	2698154	2698154	407421254	407421254	2660382	2660382	390858019	390858019	5320764	73210
STAU277	3196416	3196416	482658816	482658816	3101416	3101416	451691435	451691435	6202832	187366
STAU276	3212644	3212644	485109244	485109244	3051990	3051990	453529334	453529334	6103980	319026
STAU275	3727121	3727121	562795271	562795271	3650533	3650533	532157709	532157709	7301066	149524
STAU274	2435981	2435981	367833131	367833131	2403119	2403119	351903461	351903461	4806238	63740
STAU273R	3147279	3147279	475239129	475239129	3113592	3113592	429511370	429511370	6227184	65508
STAU272	2586699	2586699	390591549	390591549	2552340	2552340	373870298	373870298	5104680	66526
STAU271	5152917	5152917	778090467	778090467	5021151	5021151	739051853	739051853	10042302	259218
STAU270	2900123	2900123	437918573	437918573	2846833	2846833	420957781	420957781	5693666	104160
STAU269	4721862	4721862	713001162	713001162	4550045	4550045	634037364	634038004	9100090	340206
STAU268	4293014	4293014	648245114	648245114	4223647	4223647	621029884	621029884	8447294	135264
STAU267	2816933	2816933	425356883	425356883	2777961	2777961	404004701	404004701	5555922	75520
STAU266	5285267	5285267	798075317	798075317	5219422	5219422	764835393	764835393	10438844	126438
STAU265	4103945	4103945	619695695	619695695	4043190	4043190	594233060	594233060	8086380	117370
STAU264	5892838	5892838	889818538	889818538	5600935	5600935	838964122	838964122	11201870	579106
STAU262	4428922	4428922	668767222	668767222	4358221	4358221	636525992	636525992	8716442	137406
STAU261	4205531	4205531	635035181	635035181	4157470	4157470	605041798	605041798	8314940	91864
STAU260	5130997	5130997	774780547	774780547	5021596	5021596	745844430	745844430	10043192	214704
STAU259	2869409	2869409	433280759	433280759	2824295	2824295	416668821	416668821	5648590	87852
STAU258	3891621	3891621	587634771	587634771	3848733	3848733	559076044	559076044	7697466	82376
STAU257	6139158	6139158	927012858	927012858	6067504	6067504	884325624	884325624	12135008	138150

Table B.1: Quality control data for the BSI cohort.

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
STAU256	2741790	2741790	414010290	414010290	2687095	2687095	391949653	391949653	5374190	107076
STAU255	3735378	3735378	564042078	564042078	3599525	3599525	530567410	530567410	7199050	269086
STAU254	5189593	5189593	783628543	783628543	5072827	5072827	744442941	744442941	10145654	229198
STAU253	3713432	3713432	560728232	560728232	3664345	3664345	528331359	528331359	7328690	94700
STAU252	1753425	1753425	264767175	264767175	1679920	1679920	241746195	241746325	3359840	145084
STAU251	1948816	1948816	294271216	294271216	1855060	1855060	269256169	269256289	3710120	185356
STAU250	5663349	5663349	855165699	855165699	5536863	5536863	807623714	807623714	11073726	247402
STAU249	3569279	3569279	538961129	538961129	3490460	3490460	516609513	516609513	6980920	154564
STAU248	4815499	4815499	727140349	727140349	4699250	4699250	697979858	697979858	9398500	228266
STAU247	2494891	2494891	376728541	376728541	2392780	2392780	345330058	345330474	4785560	200916
STAU246	4852823	4852823	732776273	732776273	4735035	4735035	697742278	697742278	9470070	231704
STAU245	1770581	1770581	267357731	267357731	1740252	1740252	256147661	256147661	3480504	59102
STAU244	2569747	2569747	388031797	388031797	2471445	2471445	353347773	353348368	4942890	193582
STAU243	4992491	4992491	753866141	753866141	4875465	4875465	723044970	723044970	9750930	229810
STAU242	6198947	6198947	936040997	936040997	6075149	6075149	897022030	897022030	12150298	242300
STAU241	2904456	2904456	438572856	438572856	2861290	2861290	420923733	420923733	5722580	83916
STAU240	4636507	4636507	700112557	700112557	4557130	4557130	663208781	663208781	9114260	153882

Table B.1: Quality control data for the BSI cohort (cont.).

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
SO-SAU8-9	855871	855871	?57617171	?57617171	848172	848172	?30601470	?30601470	1696344	15396
SO-SAU8-8	596851	596851	?79652151	?79652151	590239	590239	?60673452	?60673452	1180478	13224
SO-SAU8-7	731939	731939	?20313639	?20313639	722145	722145	?03174807	?03174807	1444290	19588
SO-SAU8-6	542435	542435	?63272935	?63272935	537352	537352	?50321726	?50321726	1074704	10166
SO-SAU8-5	513971	513971	?54705271	?54705271	503093	503093	?41969350	?41969350	1006186	21756
SO-SAU8-43	698504	698504	?10249704	?10249704	690807	690807	?83523263	?83523263	1381614	15394
SO-SAU8-42	624355	624355	?87930855	?87930855	617744	617744	?63473461	?63473461	1235488	13222
SO-SAU8-41	677555	677555	?03944055	?03944055	671123	671123	?78401302	?78401302	1342246	12864
SO-SAU8-40	550241	550241	?65622541	?65622541	544247	544247	?46025408	?46025408	1088494	11988
SO-SAU8-4	686985	686985	?06782485	?06782485	679821	679821	?87375791	?87375791	1359642	14328
SO-SAU8-39	737953	737953	?22123853	?22123853	730780	730780	?96585115	?96585115	1461560	14346
SO-SAU8-38	780917	780917	?35056017	?35056017	773023	773023	?08163603	?08163603	1546046	15784
SO-SAU8-37	556397	556397	?67475497	?67475497	552283	552283	?47455112	?47455112	1104566	8228
SO-SAU8-36	767271	767271	?30948571	?30948571	761388	761388	?03642759	?03642759	1522776	11766
SO-SAU8-35	753964	753964	?26943164	?26943164	747327	747327	?99174021	?99174021	1494654	13274
SO-SAU8-34	850784	850784	?56085984	?56085984	840772	840772	?31226902	?31226902	1681544	20024
SO-SAU8-33	642370	642370	?93353370	?93353370	634956	634956	?67338418	?67338418	1269912	14828
SO-SAU8-32	525777	525777	?58258877	?58258877	523316	523316	?41250188	?41250188	1046632	4922
SO-SAU8-31	862656	862656	?59659456	?59659456	856649	856649	?27560678	?27560678	1713298	12012
SO-SAU8-30	605407	605407	?82227507	?82227507	601898	601898	?61408699	?61408699	1203796	7018
SO-SAU8-3	578853	578853	?74234753	?74234753	570942	570942	?59436044	?59436044	1141884	15822
SO-SAU8-29	668237	668237	?01139337	?01139337	665529	665529	?76002686	?76002686	1331058	5416
SO-SAU8-28	713398	713398	?14732798	?14732798	707449	707449	?89632373	?89632373	1414898	11896
SO-SAU8-27	770452	770452	?31906052	?31906052	762728	762728	?01131193	?01131193	1525456	15446
SO-SAU8-26	699965	699965	?10689465	?10689465	693345	693345	?83593143	?83593143	1386690	13240
SO-SAU8-25	628603	628603	?89209503	?89209503	625421	625421	?66887771	?66887771	1250842	6364
SO-SAU8-24	775981	775981	?33570281	?33570281	769359	769359	?03367844	?03367844	1538718	13244
SO-SAU8-23	685128	685128	?06223528	?06223528	679657	679657	?80994913	?80994913	1359314	10942
SO-SAU8-22	672227	672227	?02340327	?02340327	665082	665082	?80499007	?80499007	1330164	14290
SO-SAU8-21	647980	647980	?95041980	?95041980	642371	642371	?71826700	?71826700	1284742	11216
SO-SAU8-20	698423	698423	?10225323	?10225323	691755	691755	?84774810	?84774810	1383510	13334
SO-SAU8-2	486208	486208	?46348608	?46348608	481791	481791	?34287999	?34287999	963582	8834
SO-SAU8-19	644072	644072	?93865672	?93865672	636464	636464	?73953268	?73953268	1272928	15216
SO-SAU8-18	616086	616086	?85441886	?85441886	609680	609680	?66018261	?66018261	1219360	12812
SO-SAU8-17	724110	724110	?17957110	?17957110	718275	718275	?95539886	?95539886	1436550	11670
SO-SAU8-16	625787	625787	?88361887	?88361887	618067	618067	?67448516	?67448516	1236134	15440
SO-SAU8-15	561250	561250	?68936250	?68936250	554805	554805	?49040946	?49040946	1109610	12888
SO-SAU8-14	624041	624041	?87836341	?87836341	617773	617773	?66935068	?66935068	1235546	12536
SO-SAU8-13	481548	481548	?44945948	?44945948	476749	476749	?29698723	?29698723	953498	9598
SO-SAU8-12	572033	572033	?72181933	?72181933	569460	569460	?46522665	?46522665	1138920	5146
SO-SAU8-11	592146	592146	?78235946	?78235946	582024	582024	?58285811	?58285811	1164048	20244
SO-SAU8-10	720373	720373	?16832273	?16832273	712184	712184	?93884779	?93884779	1424368	16376
SO-SAU8-1	608697	608697	?83217797	?83217797	600443	600443	?64569082	?64569082	1200886	16508
SO-SAU7-9	605205	605205	?82166705	?82166705	601312	601312	?65574354	?65574354	1202624	7786
SO-SAU7-8	543925	543925	?63721425	?63721425	534726	534726	?52791826	?52791826	1069452	18398
SO-SAU7-7	917829	917829	?76266529	?76266529	897853	897853	?45631913	?45631913	1795706	39952

Table B.2: Quality control data for the MRSA cohort.

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
SO-SAU7-6	719261	719261	216497561	216497561	710511	710511	193493478	193493478	1421022	17500
SO-SAU7-5	736339	736339	221638039	221638039	726608	726608	199583824	199583824	1453216	19462
SO-SAU7-4	779705	779705	234691205	234691205	770630	770630	208535268	208535268	1541260	18150
SO-SAU7-3	639358	639358	192446758	192446758	632240	632240	165712650	165712650	1264480	14236
SO-SAU7-29	563887	563887	169729987	169729987	559218	559218	157101115	157101115	1118436	9338
SO-SAU7-28	731788	731788	220268188	220268188	723020	723020	190444117	190444117	1446040	17536
SO-SAU7-27	497745	497745	149821245	149821245	486593	486593	135865596	135865596	973186	22304
SO-SAU7-26	621925	621925	187199425	187199425	616285	616285	170172520	170172520	1232570	11280
SO-SAU7-25	659887	659887	198625987	198625987	655446	655446	178031530	178031530	1310892	8882
SO-SAU7-24	662318	662318	199357718	199357718	656426	656426	184501890	184501890	1312852	11784
SO-SAU7-23	669998	669998	201669398	201669398	659562	659562	179866405	179866405	1319124	20872
SO-SAU7-22	681915	681915	205256415	205256415	673465	673465	187924515	187924515	1346930	16900
SO-SAU7-21	595963	595963	179384863	179384863	574673	574673	160274661	160274661	1149346	42580
SO-SAU7-20	837007	837007	251939107	251939107	829760	829760	219152348	219152348	1659520	14494
SO-SAU7-2	508677	508677	153111777	153111777	497549	497549	136470307	136470307	995098	22256
SO-SAU7-19	865635	865635	260556135	260556135	856635	856635	226311747	226311747	1713270	18000
SO-SAU7-18	753953	753953	226939853	226939853	745307	745307	195182310	195182310	1490614	17292
SO-SAU7-17	663843	663843	199816743	199816743	654360	654360	181144492	181144492	1308720	18966
SO-SAU7-16	841690	841690	253348690	253348690	833036	833036	223248981	223248981	1666072	17308
SO-SAU7-15	606114	606114	182440314	182440314	599323	599323	168169323	168169323	1198646	13582
SO-SAU7-14	886935	886935	266967435	266967435	879067	879067	225414870	225414870	1758134	15736
SO-SAU7-13	729162	729162	219477762	219477762	721011	721011	192038042	192038042	1442022	16302
SO-SAU7-12	774739	774739	233196439	233196439	765796	765796	212353817	212353817	1531592	17886
SO-SAU7-11	749641	749641	225641941	225641941	742730	742730	202562238	202562238	1485460	13822
SO-SAU7-10	723521	723521	217779821	217779821	717686	717686	191907992	191907992	1435372	11670
SO-SAU7-1	625847	625847	188379947	188379947	617403	617403	169231624	169231624	1234806	16888

Table B.2: Quality control data for the MRSA cohort (cont.).

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
Tromso9200	2303261	2303261	347792411	347792411	2251490	2251490	310221376	310224362	4502980	102376
Tromso9199	999752	999752	150962552	150962552	980886	980886	141014803	141014516	1961772	37196
Tromso9198	2233335	2233335	337233585	337233585	2187000	2187000	306745176	306746785	4374000	91326
Tromso9197	1671243	1671243	252357693	252357693	1636663	1636663	245479693	245479720	3273326	68036
Tromso9196	2340415	2340415	353402665	353402665	2287038	2287038	320991034	320992650	4574076	105428
Tromso9195	2409600	2409600	363849600	363849600	2350279	2350279	352913869	352914312	4700558	117368
Tromso9194	2429363	2429363	366833813	366833813	2381568	2381568	324733217	324736053	4763136	94360
Tromso9193	2132689	2132689	322036039	322036039	2082004	2082004	312276613	312275610	4164008	100128
Tromso9192	2539795	2539795	383509045	383509045	2490993	2490993	337445347	337448521	4981986	96426
Tromso9191	1625971	1625971	245521621	245521621	1588985	1588985	238383176	238383453	3177970	73000
Tromso9190	962542	962542	145343842	145343842	943278	943278	141406895	141406897	1886556	38054
Tromso9189	2562671	2562671	386963321	386963321	2507553	2507553	339706211	339710221	5015106	109040
Tromso9188	1203727	1203727	181762777	181762777	1179672	1179672	161569852	161571273	2359344	47348
Tromso9187	382329	382329	57731679	57731679	373937	373937	56106592	56106628	747874	16532
Tromso9185	2122273	2122273	320463223	320463223	2075085	2075085	288503082	288504719	4150170	93378
Tromso9183	2313368	2313368	349318568	349318568	2263136	2263136	339708518	339707331	4526272	99250
Tromso9182	2342766	2342766	353757666	353757666	2290241	2290241	315502329	315505236	4580482	103972
Tromso9181	2073260	2073260	313062260	313062260	2031336	2031336	304956578	304956352	4062672	82610
Tromso9180	1945522	1945522	293773822	293773822	1905180	1905180	286347314	286347542	3810360	79772
Tromso9179	1878911	1878911	283715561	283715561	1840271	1840271	260179697	260181161	3680542	76326
Tromso9178	2233361	2233361	337237511	337237511	2187808	2187808	297379832	297381880	4375616	89814
Tromso9177	1908196	1908196	288137596	288137596	1870381	1870381	279500316	279500437	3740762	74606
Tromso9176	1737144	1737144	262308744	262308744	1695027	1695027	254209526	254210490	3390054	83306
Tromso9174	1760045	1760045	265766795	265766795	1722363	1722363	258322427	258322366	3444726	74384
Tromso9173	2093484	2093484	316116084	316116084	2042238	2042238	306634215	306633698	4084476	101480
Tromso9172	1491947	1491947	225283997	225283997	1457455	1457455	219146426	219146878	2914910	68186
Tromso9170	2151223	2151223	324834673	324834673	2113989	2113989	290999215	291000491	4227978	73366
Tromso9169	3045859	3045859	459924709	459924709	2981837	2981837	448399336	448398667	5963674	126504
Tromso9168	1370618	1370618	206963318	206963318	1344568	1344568	199404875	199404665	2689136	51338
Tromso9167	1661447	1661447	250878497	250878497	1622103	1622103	243221108	243221153	3244206	77808
Tromso9166	909194	909194	137288294	137288294	891111	891111	133625975	133625744	1782222	35582
Tromso9165	1937577	1937577	292574127	292574127	1897394	1897394	285151929	285152258	3794788	79490
Tromso9164	1424535	1424535	215104785	215104785	1392309	1392309	208630935	208630357	2784618	63666
Tromso9163	1967985	1967985	297165735	297165735	1928141	1928141	289485870	289485682	3856282	78706
Tromso9162	2130283	2130283	321672733	321672733	2081976	2081976	312988665	312988235	4163952	95486
Tromso9161	3841347	3841347	580043397	580043397	3746840	3746840	559182434	559183413	7493680	187130
Tromso9159	2191018	2191018	330843718	330843718	2137375	2137375	320768972	320769617	4274750	106112
Tromso9157	1756854	1756854	265284954	265284954	1719811	1719811	257086206	257086681	3439622	73278
Tromso9156	3165492	3165492	477989292	477989292	3091936	3091936	463990488	463989314	6183872	145428
Tromso9155	1616245	1616245	244052995	244052995	1581412	1581412	217831829	217833488	3162824	68794
Tromso9154	1075318	1075318	162373018	162373018	1054226	1054226	143965096	143965620	2108452	41706
Tromso9152	1964181	1964181	296591331	296591331	1911815	1911815	286295029	286294089	3823630	103770
Tromso9151	2201808	2201808	332473008	332473008	2144551	2144551	321918684	321918354	4289102	113202
Tromso9150	2066948	2066948	312109148	312109148	2020361	2020361	302735388	302735376	4040722	92124
Tromso9149	2060248	2060248	311097448	311097448	2010311	2010311	300947893	300947825	4020622	98828
Tromso9148	1719013	1719013	259570963	259570963	1682061	1682061	251773950	251773710	3364122	73066

Table B.3: Quality control data for the TSSS cohort.

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
Tromso9147	2692223	2692223	406525673	406525673	2639298	2639298	367165800	367167189	5278596	104324
Tromso9146	2209724	2209724	333668324	333668324	2157650	2157650	323524373	323523809	4315300	102852
Tromso9145	2684589	2684589	405372939	405372939	2626939	2626939	351481885	351485549	5253878	113976
Tromso9144	2393263	2393263	361382713	361382713	2334574	2334574	350643449	350643988	4669148	116118
Tromso9143	2170586	2170586	327758486	327758486	2111427	2111427	316660341	316660231	4222854	117020
Tromso9141	4488105	4488105	677703855	677703855	4387108	4387108	657564399	657565198	8774216	199596
Tromso9140	2114425	2114425	319278175	319278175	2072852	2072852	283764381	283766952	4145704	82072
Tromso9139	712372	712372	107568172	107568172	697415	697415	104314263	104314055	1394830	29588
Tromso9138	1558663	1558663	235358113	235358113	1524977	1524977	228422556	228422563	3049954	66650
Tromso9137	2339011	2339011	353190661	353190661	2286021	2286021	342237747	342237554	4572042	104784
Tromso9135	1732889	1732889	261666239	261666239	1692750	1692750	254246712	254247061	3385500	79354
Tromso9134	1339007	1339007	202190057	202190057	1311086	1311086	196191724	196190985	2622172	55130
Tromso9133	2586786	2586786	390604686	390604686	2529180	2529180	346198304	346201369	5058360	113866
Tromso9132	2084088	2084088	314697288	314697288	2043260	2043260	306669016	306669270	4086520	80638
Tromso9131	2533966	2533966	382628866	382628866	2472088	2472088	371175231	371175469	4944176	122600
Tromso9130	2620156	2620156	395643556	395643556	2545893	2545893	382900430	382900104	5091786	147236
Tromso9128	2759518	2759518	416687218	416687218	2697693	2697693	375996935	375999204	5395386	122248
Tromso9126	2381749	2381749	359644099	359644099	2338984	2338984	351513190	351512187	4677968	84230
Tromso9125	1834307	1834307	276980357	276980357	1793578	1793578	268927444	268927253	3587156	80480
Tromso9124	1642359	1642359	247996209	247996209	1609205	1609205	241099955	241100167	3218410	65530
Tromso9123	2009241	2009241	303395391	303395391	1966207	1966207	294911500	294911303	3932414	85134
Tromso9121	1876298	1876298	283320998	283320998	1834637	1834637	275221338	275221277	3669274	82398
Tromso9120	1291319	1291319	194989169	194989169	1265587	1265587	188761788	188761640	2531174	50794
Tromso9118	2125792	2125792	320994592	320994592	2080173	2080173	311779284	311778979	4160346	90120
Tromso9117	2306958	2306958	348350658	348350658	2258098	2258098	322245491	322246843	4516196	96442
Tromso9115	1940631	1940631	293035281	293035281	1894393	1894393	283752396	283751716	3788786	91362
Tromso9114	2214941	2214941	334456091	334456091	2169644	2169644	323506971	323506717	4339288	89456
Tromso9113	2512535	2512535	379392785	379392785	2457906	2457906	368682940	368682790	4915812	107820
Tromso9112	2468299	2468299	372713149	372713149	2411586	2411586	360935910	360935854	4823172	112112
Tromso9111	2103702	2103702	317659002	317659002	2046051	2046051	307181679	307181751	4092102	114350
Tromso9110	2033293	2033293	307027243	307027243	1978267	1978267	296730966	296730031	3956534	108890
Tromso9109	1661253	1661253	250849203	250849203	1619885	1619885	242257840	242259300	3239770	81772
Tromso9107	2138907	2138907	322974957	322974957	2088994	2088994	313772582	313773069	4177988	98606
Tromso9106	2458878	2458878	371290578	371290578	2407587	2407587	324794881	324797090	4815174	101200
Tromso9105	2414124	2414124	364532724	364532724	2358922	2358922	304969716	304974121	4717844	109066
Tromso9104	3249964	3249964	490744564	490744564	3170881	3170881	474903600	474904244	6341762	156458
Tromso9103	2693118	2693118	406660818	406660818	2630281	2630281	366077871	366081532	5260562	123956
Tromso9102	2757063	2757063	416316513	416316513	2696391	2696391	366009554	366015061	5392782	119640
Tromso9101	973953	973953	147066903	147066903	950290	950290	141941447	141941478	1900580	46892
Tromso9100	2616747	2616747	395128797	395128797	2557581	2557581	367342847	367343602	5115162	116552
Tromso9099	2741776	2741776	414008176	414008176	2680461	2680461	401178112	401178440	5360922	121382
Tromso9098	2712329	2712329	409561679	409561679	2645486	2645486	397139424	397138910	5290972	132228
Tromso9097	1753272	1753272	264744072	264744072	1709885	1709885	255501526	255502136	3419770	85942
Tromso9096	1385179	1385179	209162029	209162029	1351870	1351870	202480305	202480626	2703740	65966
Tromso9095	1446024	1446024	218349624	218349624	1408237	1408237	210669119	210669612	2816474	74664
Tromso9094	2357567	2357567	355992617	355992617	2311806	2311806	309786596	309791679	4623612	90448

Table B.3: Quality control data for the TSSS cohort (cont.).

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
Tromso9093	2106520	2106520	318084520	318084520	2054436	2054436	308611082	308611563	4108872	102920
Tromso9092	2628973	2628973	396974923	396974923	2570642	2570642	386108997	386108472	5141284	115256
Tromso9091	6331461	6331461	956050611	956050611	6184891	6184891	873917620	873921726	12369782	289048
Tromso9090	1478562	1478562	223262862	223262862	1442242	1442242	215337427	215337021	2884484	71808
Tromso9089	1958694	1958694	295762794	295762794	1913467	1913467	285537355	285537371	3826934	89538
Tromso9088	2736397	2736397	413195947	413195947	2673269	2673269	401302969	401302947	5346538	124598
Tromso9087	2339226	2339226	353223126	353223126	2280593	2280593	342815052	342814791	4561186	116106
Tromso9086	2241542	2241542	338472842	338472842	2182665	2182665	327525173	327525102	4365330	116514
Tromso9085	2265004	2265004	342015604	342015604	2213329	2213329	331892563	331892505	4426658	102180
Tromso9084	1063182	1063182	160540482	160540482	1039575	1039575	155662327	155662580	2079150	46652
Tromso9082	2577842	2577842	389254142	389254142	2524021	2524021	354408365	354410302	5048042	106242
Tromso9081	2387544	2387544	360519144	360519144	2335390	2335390	350386840	350386385	4670780	103152
Tromso9080	2707617	2707617	408850167	408850167	2632081	2632081	394800424	394800385	5264162	149818
Tromso9079	2648661	2648661	399947811	399947811	2595004	2595004	354866889	354870829	5190008	105778
Tromso9078	2236486	2236486	337709386	337709386	2186806	2186806	327570437	327570312	4373612	98182
Tromso9075	2068694	2068694	312372794	312372794	2027025	2027025	303544662	303544126	4054050	82200
Tromso9074	2583268	2583268	390073468	390073468	2516889	2516889	377983680	377983478	5033778	131492
Tromso9072	1951352	1951352	294654152	294654152	1902770	1902770	285204857	285205736	3805540	96170
Tromso9071	2487406	2487406	375598306	375598306	2436947	2436947	331576254	331579615	4873894	99538
Tromso9069	2423079	2423079	365884929	365884929	2362083	2362083	355025309	355024406	4724166	120820
Tromso9068	1500426	1500426	226564326	226564326	1464972	1464972	219636827	219636605	2929944	69920
Tromso9067	2412894	2412894	364346994	364346994	2362749	2362749	318014265	318016689	4725498	98862
Tromso9063	2840764	2840764	428955364	428955364	2754909	2754909	414519100	414519173	5509818	169978
Tromso9062	2124821	2124821	320847971	320847971	2073139	2073139	310673731	310674071	4146278	102158
Tromso9060	1467857	1467857	221646407	221646407	1433400	1433400	214998569	214998945	2866800	68176
Tromso9058	1814825	1814825	274038575	274038575	1778069	1778069	262633062	262633148	3556138	72454
Tromso9057	2042932	2042932	308482732	308482732	1988896	1988896	299052878	299053613	3977792	106758
Tromso9056	1901355	1901355	287104605	287104605	1847755	1847755	277363628	277363114	3695510	106264
Tromso9055	2452615	2452615	370344865	370344865	2398519	2398519	360218785	360218526	4797038	106816
Tromso9054	2107801	2107801	318277951	318277951	2049919	2049919	307522054	307521743	4099838	114568
Tromso9053	2118405	2118405	319879155	319879155	2063109	2063109	309306440	309306148	4126218	109634
Tromso9050	1833309	1833309	276829659	276829659	1790912	1790912	268225970	268225623	3581824	83702
Tromso9049	2440939	2440939	368581789	368581789	2382468	2382468	359027549	359027074	4764936	115880
Tromso9048	2070905	2070905	312706655	312706655	2016249	2016249	302961343	302960172	4032498	108170
Tromso9047	2062615	2062615	311454865	311454865	2005335	2005335	300452288	300451829	4010670	113360
Tromso9046	483363	483363	72987813	72987813	471178	471178	68897354	68896993	942356	24086
Tromso9045	2194850	2194850	331422350	331422350	2137191	2137191	320988265	320987733	4274382	114092
Tromso9043	2088424	2088424	315352024	315352024	2037043	2037043	306660570	306660478	4074086	101454
Tromso9042	1272740	1272740	192183740	192183740	1241962	1241962	187170641	187170574	2483924	61042
Tromso9041	2328512	2328512	351605312	351605312	2277326	2277326	341622064	341622234	4554652	101222
Tromso9040	2388762	2388762	360703062	360703062	2336283	2336283	333617218	333617332	4672566	103594
Tromso9039	1797389	1797389	271405739	271405739	1760432	1760432	263107060	263107146	3520864	73046
Tromso9038	2875706	2875706	434231606	434231606	2793431	2793431	420734775	420734194	5586862	162852
Tromso9037	2208172	2208172	341525156	341525156	2154597	2154597	331675718	331675504	4309194	105854
Tromso9036	2261756	2261756	341525156	341525156	2210151	2210151	331675718	331675504	4420302	102120
Tromso9035	2183151	2183151	329655801	329655801	2116107	2116107	318537500	318536751	4232214	132738

Table B.3: Quality control data for the TSSS cohort (cont.).

	Before filtering				After filtering					
Strain	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Forward reads	Reverse reads	Base pairs forward reads	Base pairs reverse reads	Reads passed filter	Low quality reads
Tromso9030	2706846	2706846	408733746	408733746	2654910	2654910	357925498	357928698	5309820	102332
Tromso9029	2104170	2104170	317729670	317729670	2055910	2055910	308530372	308530086	4111820	95354
Tromso9028	2613273	2613273	394604223	394604223	2556828	2556828	382726492	382725554	5113656	111612
Tromso9027	5036923	5036923	760575373	760575373	4927914	4927914	696013555	696014243	9855828	215096
Tromso9026	2278683	2278683	344081133	344081133	2226801	2226801	333747611	333747372	4453602	102530
Tromso9023	2123353	2123353	320626303	320626303	2075233	2075233	311629361	311629044	4150466	95182
Tromso9022	2398435	2398435	362163685	362163685	2341817	2341817	351472986	351472135	4683634	112064
Tromso9021	2023738	2023738	305584438	305584438	1973119	1973119	295950229	295950053	3946238	100044
Tromso9020	1793960	1793960	270887960	270887960	1752466	1752466	262990326	262990321	3504932	82172
Tromso9019	2712670	2712670	409613170	409613170	2642754	2642754	396935039	396934069	5285508	138462
Tromso9018	2053504	2053504	310079104	310079104	1998231	1998231	300407788	300407163	3996462	109446
Tromso9016	2662607	2662607	402053657	402053657	2601272	2601272	389991710	389991748	5202544	121470
Tromso9015	1998792	1998792	301817592	301817592	1946422	1946422	292281699	292281417	3892844	103708
Tromso9014	1897137	1897137	286467687	286467687	1852079	1852079	277408157	277408007	3704158	88878
Tromso9013	2178146	2178146	328900046	328900046	2126092	2126092	319519547	319519084	4252184	102904
Tromso9012	1707174	1707174	257783274	257783274	1674167	1674167	249851474	249851207	3348334	65120
Tromso9011	2454156	2454156	370577556	370577556	2405009	2405009	361565108	361564843	4810018	97116
Tromso9010	2675693	2675693	404029643	404029643	2618797	2618797	392203182	392203517	5237594	112474
Tromso9008	2552813	2552813	385474763	385474763	2490270	2490270	374651915	374651236	4980540	123602
Tromso9007	2666921	2666921	402705071	402705071	2608999	2608999	391642903	391642117	5217998	114452
Tromso9006	2254842	2254842	340481142	340481142	2199735	2199735	329948754	329949275	4399470	109094
Tromso9004	2410228	2410228	363944428	363944428	2346768	2346768	353088381	353087549	4693536	125470
Tromso9003	2285045	2285045	345041795	345041795	2225629	2225629	334381500	334381260	4451258	117524
Tromso9001	765691	765691	115619341	115619341	749645	749645	107423812	107423899	1499290	31546

Table B.3: Quality control data for the TSSS cohort (cont.).

C Mean quality score

This appendix shows the mean quality score for each base in the forward and reverse reads for all strains in each cohort TSSS (C.1), BSI (C.2) and MRSA (C.3) for raw and trimmed data.

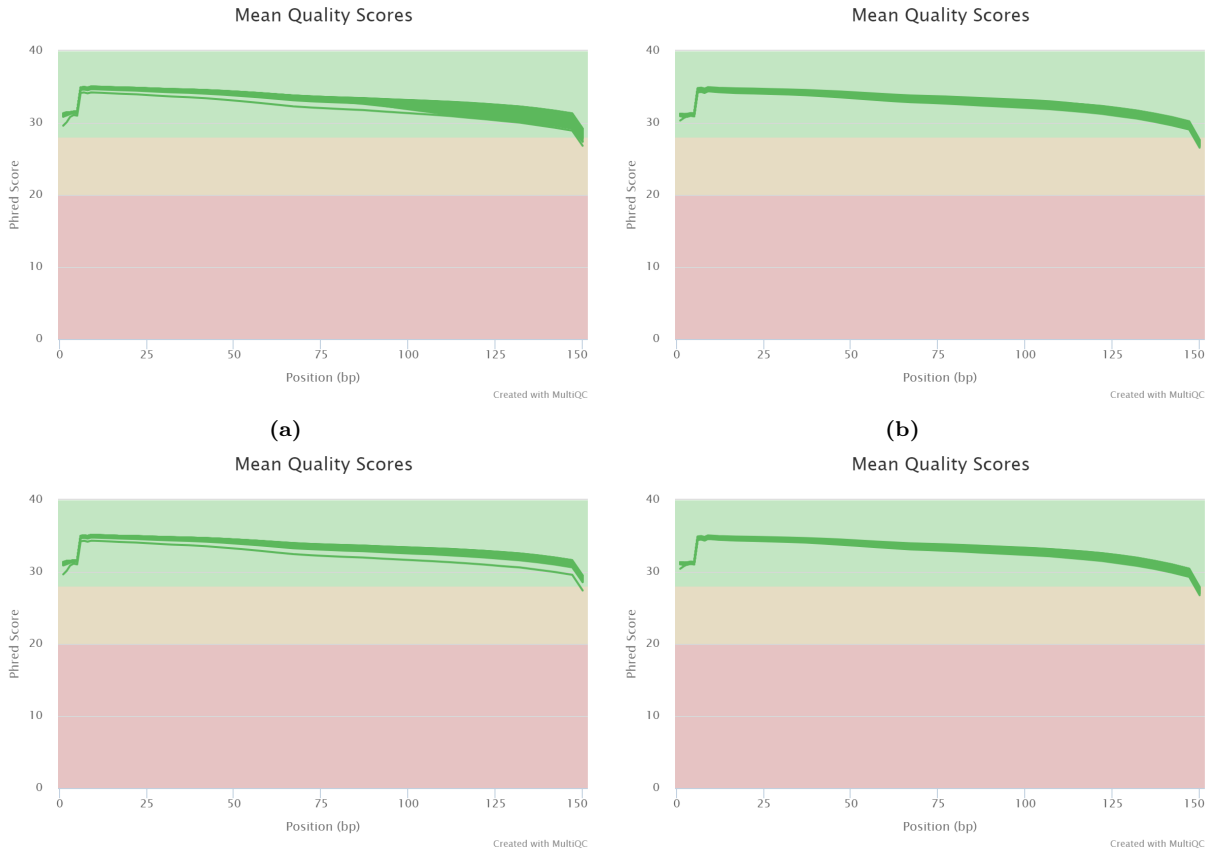


Figure C.1: The mean quality score for each base in the forward and reverse reads for all strains in the TSSS cohort for raw and trimmed data. The subfigures show the mean quality score for: a) Forward reads, raw data b) reverse reads, raw data c) forward reads, trimmed data d) reverse reads, trimmed data.

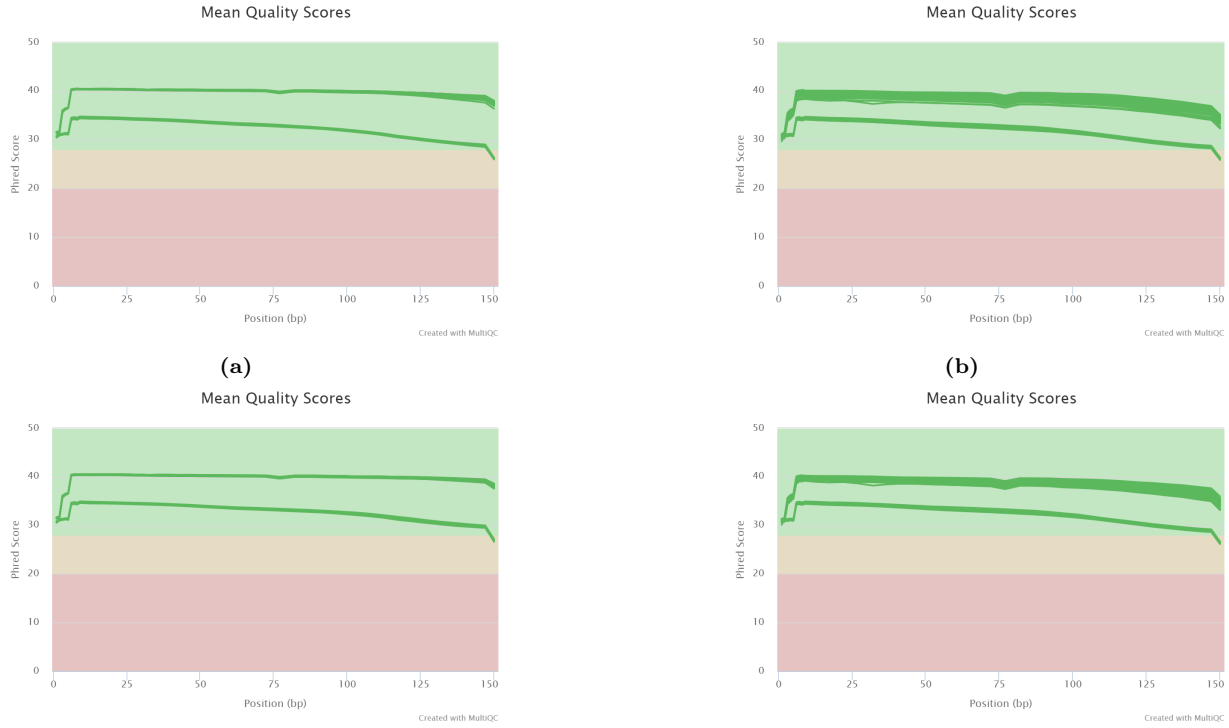


Figure C.2: The mean quality score for each base in the forward and reverse reads for all strains in the BSI cohort for raw and trimmed data. The subfigures show the mean quality score for: a) Forward reads, raw data b) reverse reads, raw data c) forward reads, trimmed data d) reverse reads, trimmed data.

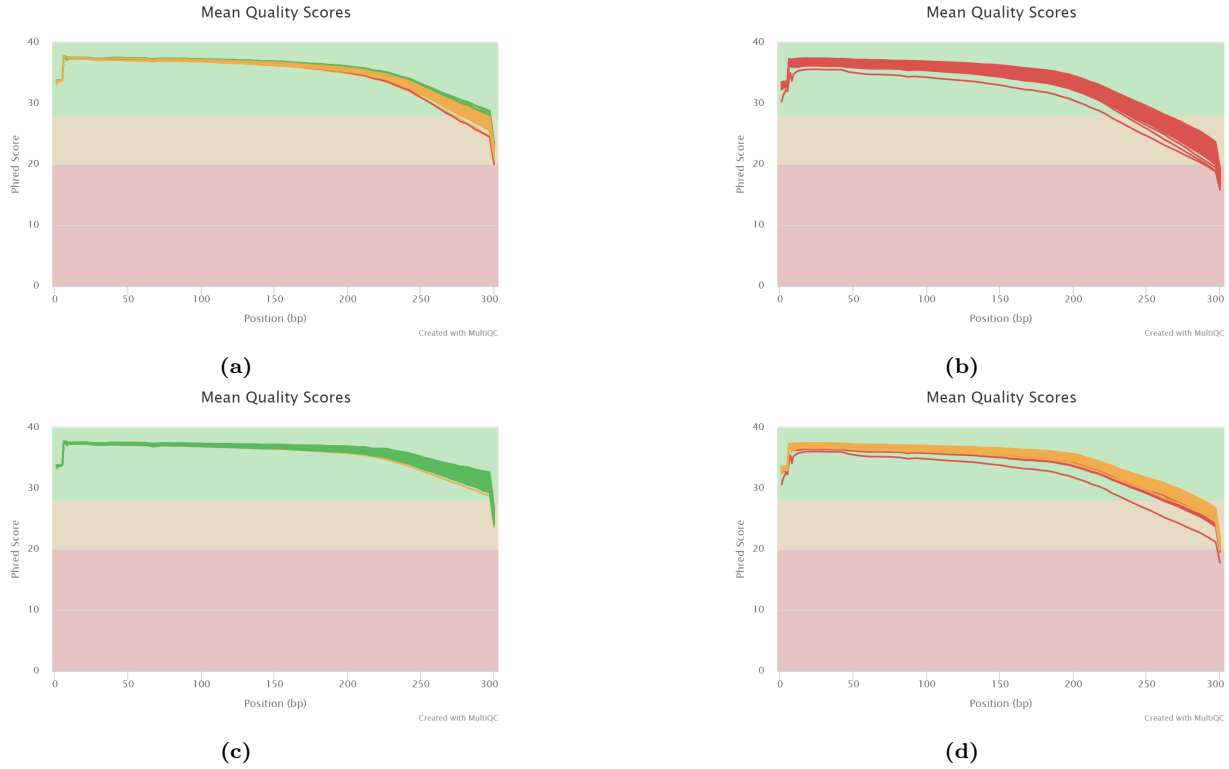


Figure C.3: The mean quality score for each base in the forward and reverse reads for all strains in the MRSA cohort for raw and trimmed data. The subfigures show the mean quality score for: a) Forward reads, raw data b) reverse reads, raw data c) forward reads, trimmed data d) reverse reads, trimmed data.

D Adapter content

This appendix show the adapter content of reads where the adapter contamination $\geq$ 0.1% for all three cohorts TSSS (D.1), BSI (D.2) and MRSA (D.3). For trimmed reverse reads in the TSSS cohort and trimmed forward and reverse reads in the BSI cohort, no reads had adapter contamination $\geq$ 0.1%.

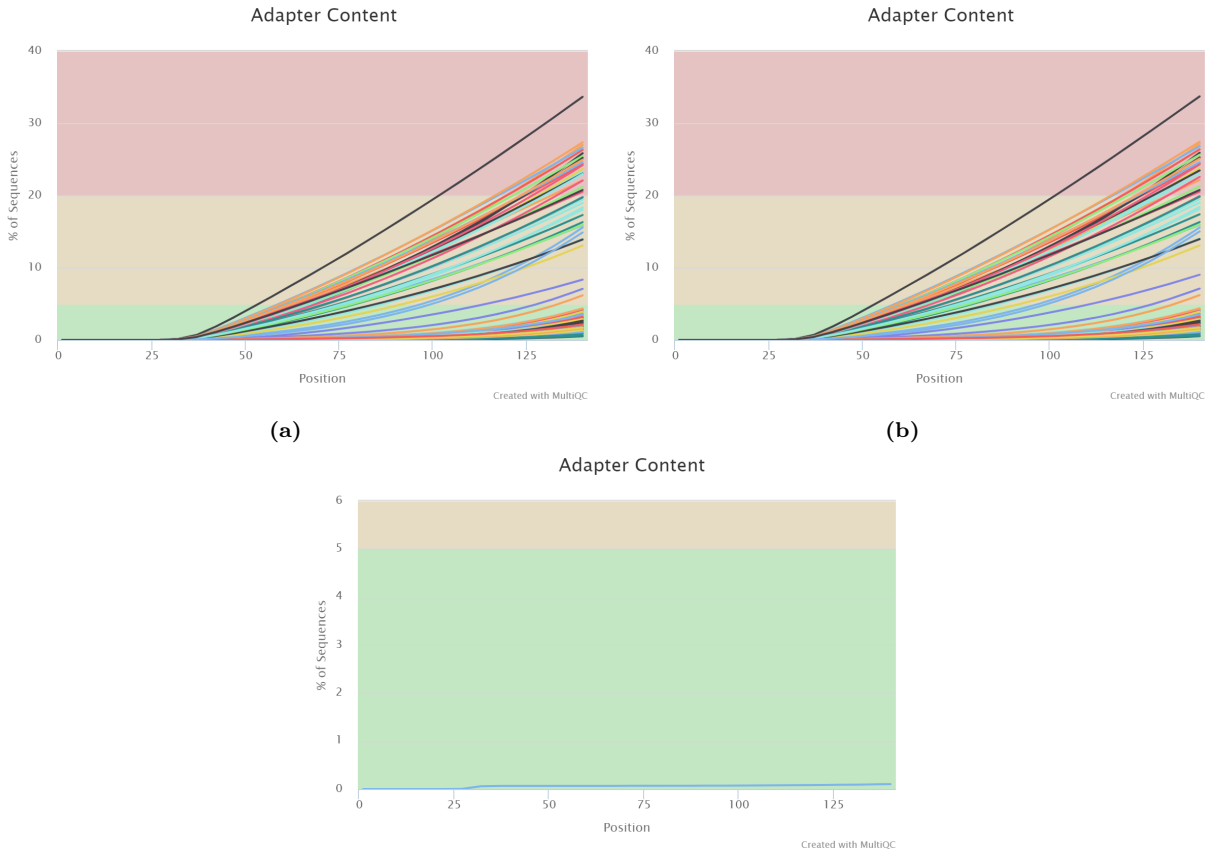
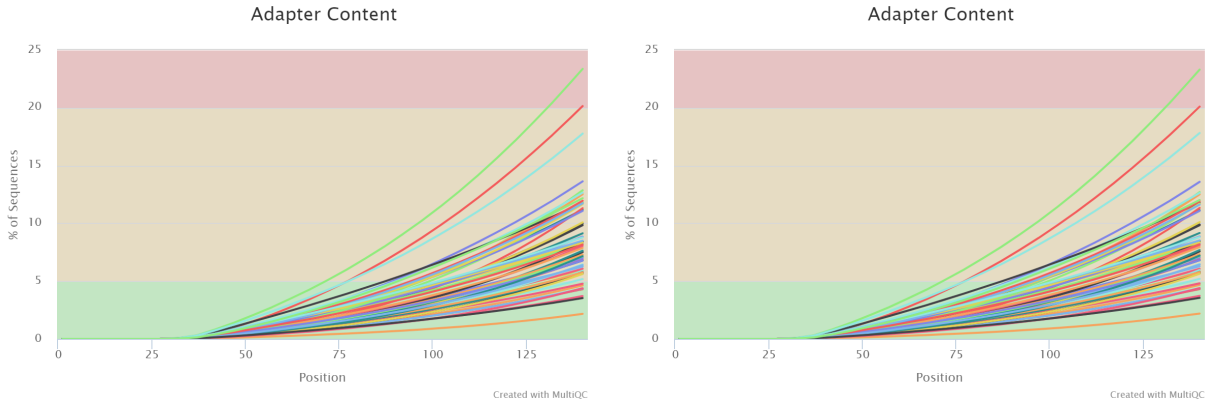
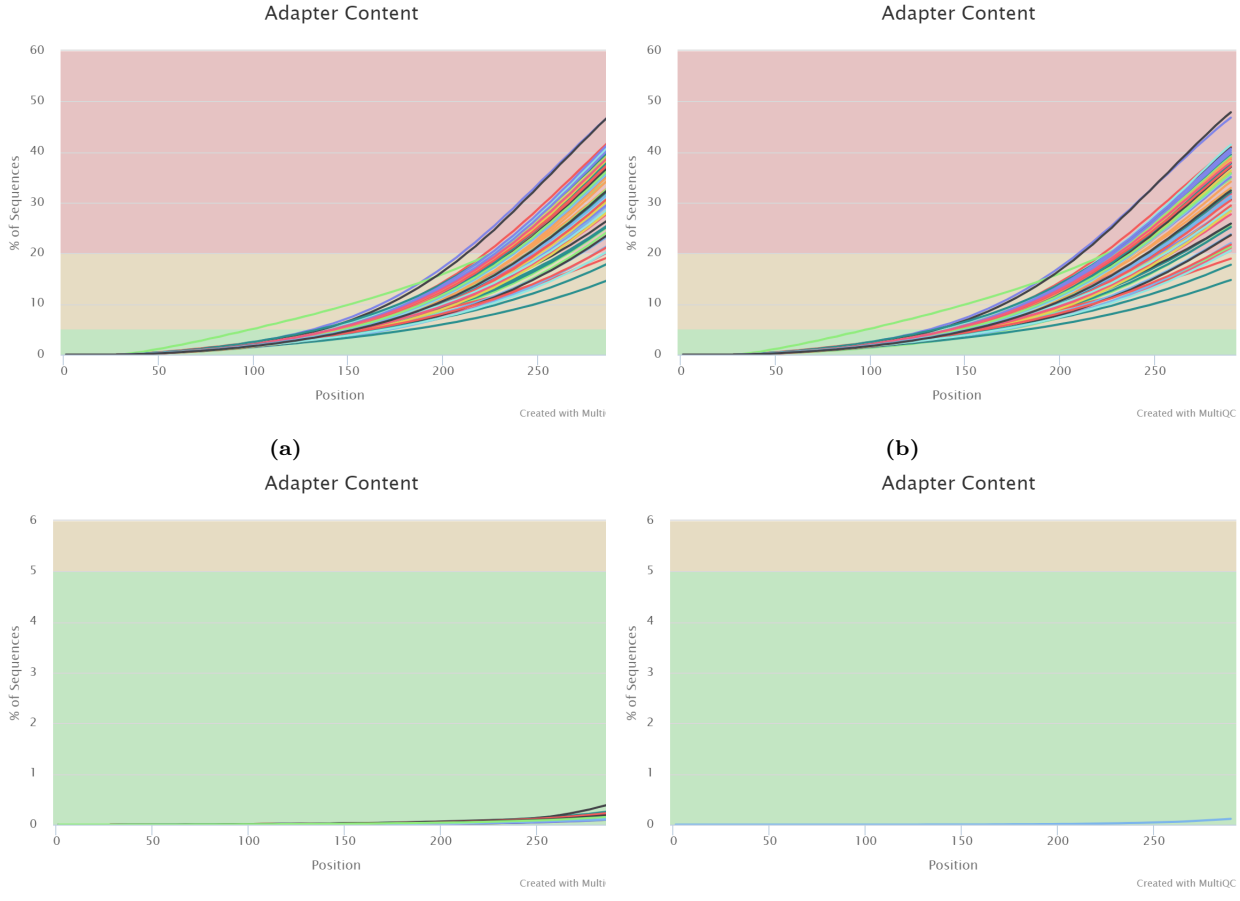


Figure D.1: The percentage of the sequences at each position of the reads from the TSSS cohort with adapter contamination $\geq$ 0.1%. The plots are created with MultiQC, and show the adapter content of: a) raw forward reads b) raw reverse reads c) trimmed forward reads.



(a) **(b)**

Figure D.2: The percentage of the sequences at each position of the reads from the BSI cohort with adapter contamination $\geq 0.1\%$. The plots are created with MultiQC, and show the adapter content of: a) raw forward reads b) raw reverse reads.



(a) **(b)**

(c) **(d)**

Figure D.3: The percentage of the sequences at each position of the reads from the MRSA cohort with adapter contamination $\geq 0.1\%$. The plots are created with MultiQC, and show the adapter content of: a) raw forward reads b) raw reverse reads c) trimmed forward reads d) trimmed reverse reads.

E Pangenome frequency

The pangenome frequency in figure E.1 shows the number of genes detected in the genomes. The full number of genes is given for the first genome in the plot. The subsequently strains show the number of additional genes present in their genomes, that were not present in the previous genomes. The first genome contains almost 2000 genes, which contains the core genes as well as additional genes present in that particular strain. The first 75-80 subsequent genomes have noticeable additional genes, which are genes not present in the core genome, but additional genes not present in all strains in the three cohorts. After that, few new genes are added to the pangenome before the approximately 10 last genomes. The last genome adds almost 1500 new genes to the pangenome, meaning it has a lot of genes not present in any of the other strains.

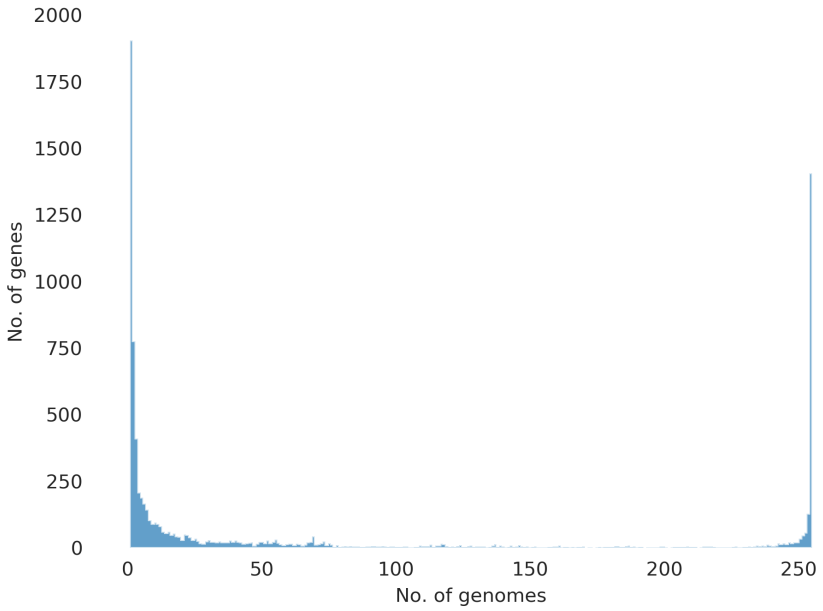


Figure E.1: Pangenome frequency first showing the total number of genes present in one of the genomes. For the subsequently genomes, the number of additional genes not present in previous genomes is shown.

F Phylogenetic tree

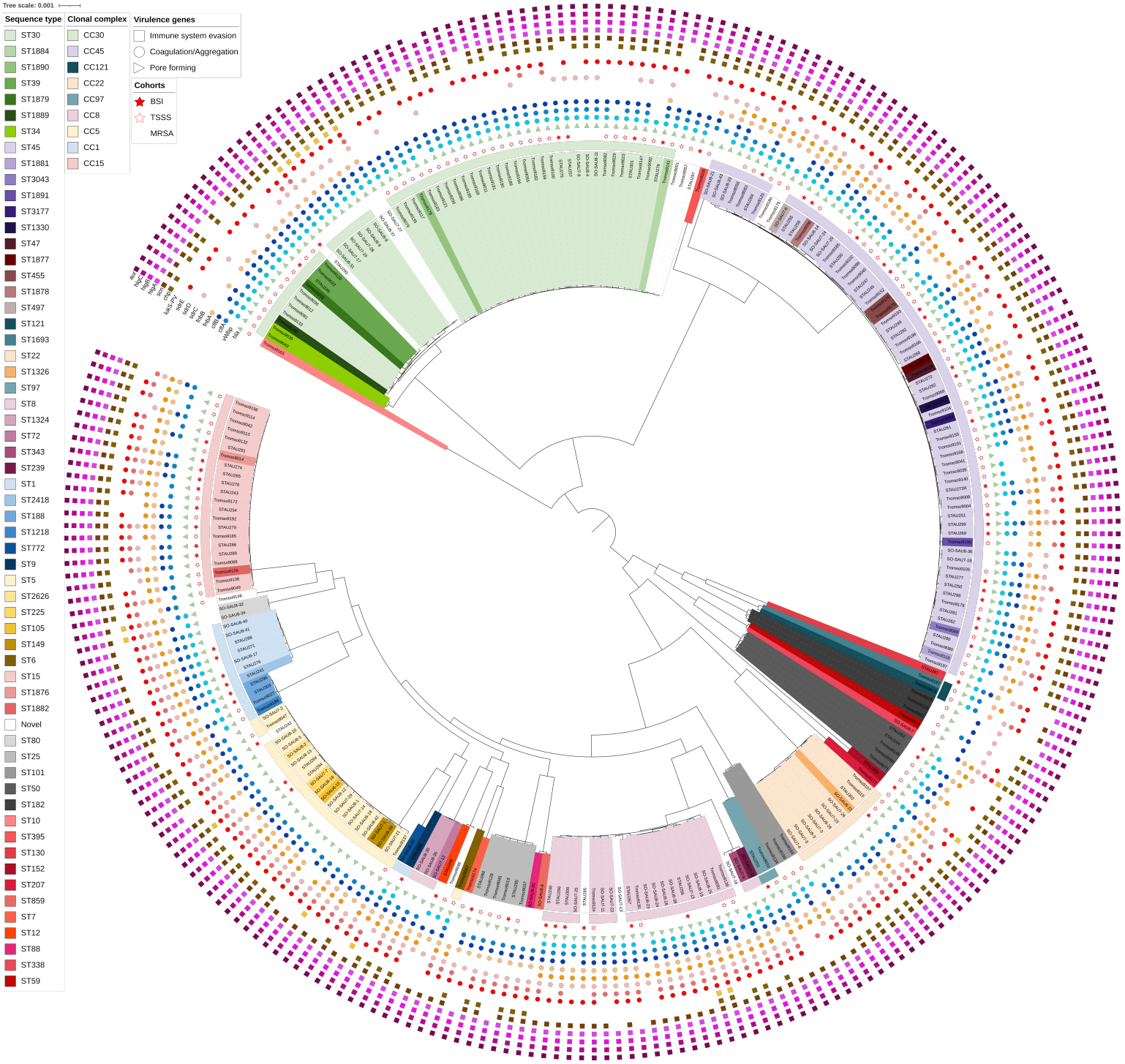


Figure F.1: Phylogenetic tree showing all strains from all three cohorts TSSS, BSI and MRSA. The strains from the BSI cohort are marked with a filled red star, the TSSS strains are marked with a white star and the MRSA strains have no marking. The strains are marked in colours according to their sequence type (ST). The strains with no colour marking did not get a ST during multi-locus sequence typing. The presence and absence of virulence genes relevant in bloodstream infection-causing *Staphylococcus aureus*, are displayed next to the tree. Virulence factors marked with a rectangle plays a role in immune system evasion, the circular plays a role in coagulation and aggregation of blood and the pore forming gene *hla* is marked with a triangle.

G *SdrC* phylogenetic tree

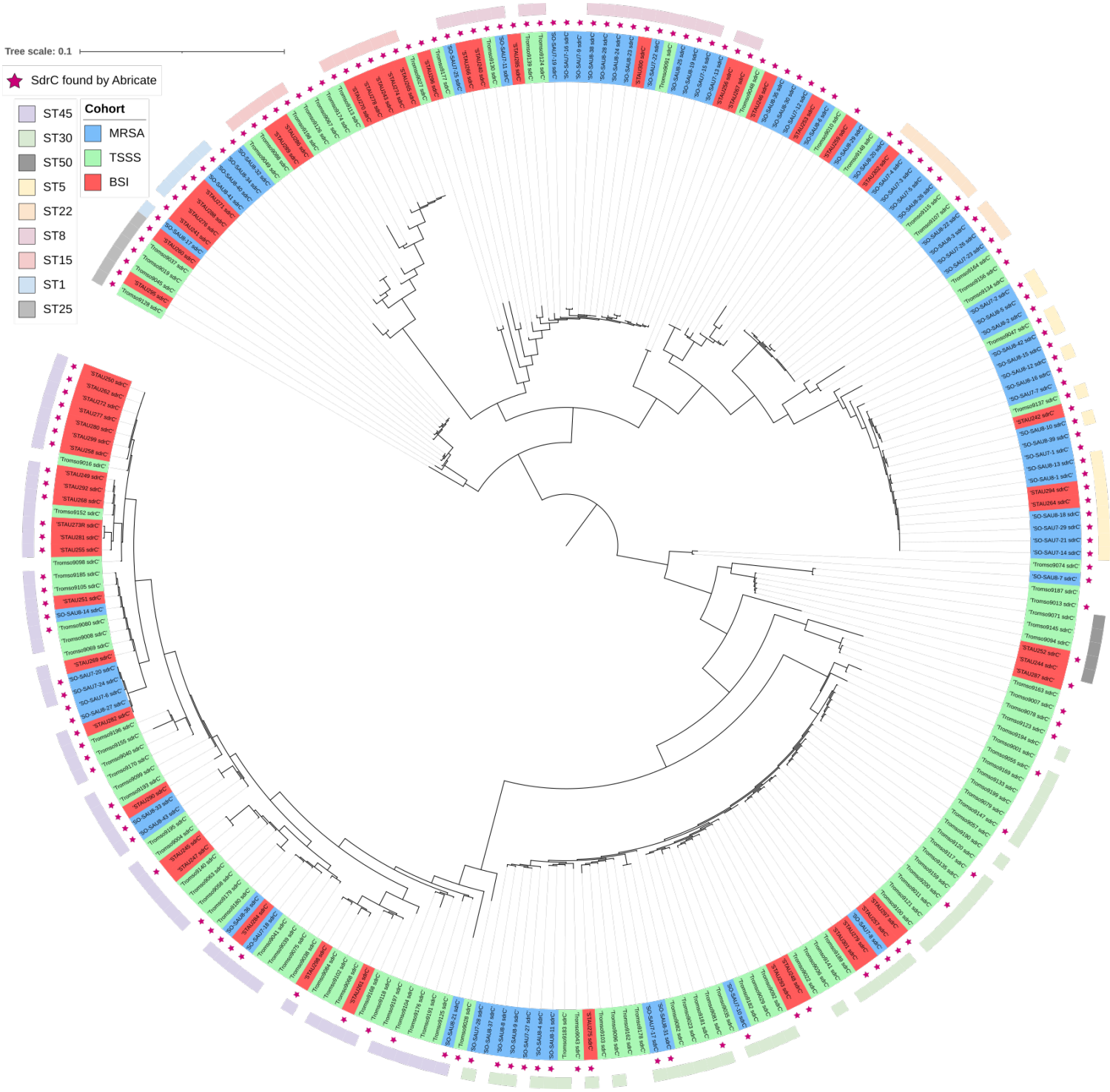
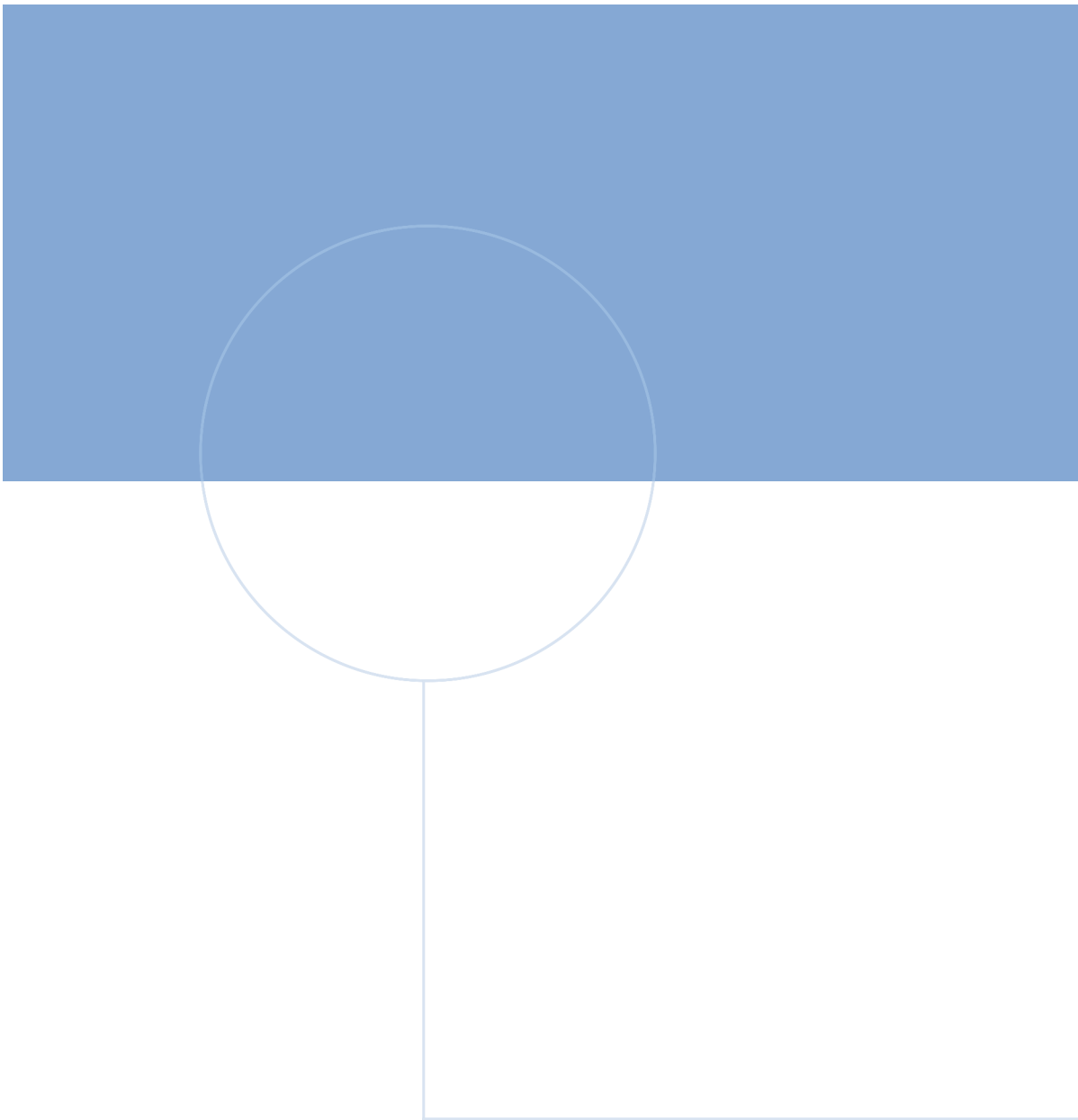


Figure G.1: A phylogenetic tree showing the possible relationship between the strain's *sdrC* genes. The *sdrC* genes was extracted in Geneious and visualized in iTOL. *SdrC* genes from the TSSS cohort is marked in green, the ones from the MRSA cohort is marked in blue and the ones from the BSI cohort is marked in red. What sequence type (for sequence types with 5 or more strains) the strain the gene is extracted from belongs to is shown, and whether it was found by Abricate when running Nullarbor.



NTNU

Norwegian University of
Science and Technology