COMPARATIVE GENOMIC OF METHICILLIN-RESISTANT *Staphylococcus aureus* PR01 (MRSA PR01) AND METHICILLIN-SENSITIVE *Staphylococcus aureus* SA D22901 (MSSA SA D22901) AND METHICILLIN RESISTANT DERIVATIVES OF THE LATTER

by

LEE LIAN SHIEN

Thesis submitted in fulfilment of the requirements for the degree of Master of Science

November 2014

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank a couple of people that have helped made this possible for me:

To my supervisors, Professor Mohd. Zaki Salleh and Professor Zainul F. Zainuddin - Thank you for providing this opportunity to learn and grow. I have learnt a lot about research and science under your tutelage and with your guidance.

To the members of the PROMISE Centre, UiTM Puncak Alam, where most of the work for this study was done: Thank you for providing the facilities and means by which the sequencing and data analysis could be done. To Professor Teh Lay Kek, who provided so much guidance and help with the research project and thesis writing.

Last but not least, to my family and friends. My husband, Alan Lee Hann Wu was instrumental in supporting me emotionally throughout. To my parents, Lee Sing Lee and Lam Swee Wan, and siblings Lian Ni, Wei Yen and Wei Ran, who were always there for me. Last but certainly not least; this work of study is dedicated to the little one keeping me company while this thesis was in preparation, my daughter Sophie Lee Yue Ning.

This project was supported by a grant from the Ministry of Higher Education Malaysia (Grant no. 600-RMI/ST/FRGS 5/3/Fst (58/2010))

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

<	Less than
>	More than
×	-fold
°	Degree
°C	Degree Celsius
µg	Microgram
µl	Microlitre
Bp	Base pair
CA-MRSA	Community-acquired methicillin-resistant *Staphylococcus aureus*
CC	Clonal Complex
CLSI	Clinical and Laboratory Standards Institute
dH_2O	Distilled water
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
EMBL	European Molecular Biology Laboratory
EST(s)	Expressed sequence tag(s)
FU	Fluorescence Unit
gDNA	Genomic DNA
HA-MRSA	Healthcare-acquired methicillin-resistant *Staphylococcus aureus*
HGT	Horizontal Gene Transfer
InDel	Insertion/Deletion
Kb	Kilobase

Mb	Megabase
MGE(s)	Mobile Genetic Element(s)
MIC	Minimum Inhibitory Concentration
Min	Minute
ml	Millilitre
MLST	Multi-locus sequence typing
mM	Millimolar
MRSA	Methicillin-resistant *Staphylococcus aureus*
MSSA	Methicillin-susceptible *Staphylococcus aureus*
NCBI	National Centre for Biotechnology Information
OD	Optical Density
ORF(s)	Open Reading Frame(s)
PCR	Polymerase Chain Reaction
PE	Paired-end
pM	Picomolar
qPCR	Quantitative PCR
QRDR	Quinolone-resistance Determining Region
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
RT-PCR	Real time PCR
SaPI(s)	*Staphylococcus aureus* Pathogenicity Island(s)
SBS	Sequencing by synthesis

SCC	Staphylococcal Chromosomal Cassette
Sec	Second
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
Tn	Transposon
UV	Ultraviolet
VISA	Vancomycin-resistant *Staphylococcus aureus*
A	Alpha
B	Beta
N	Nu
Φ	Phi
Ψ	Psi

PERBANDINGAN GENOMIK METHICILLIN-RESISTANT *Staphylococcus aureus* PR01 (MRSA PR01) DAN METHICILLIN-SENSITIVE *Staphylococcus aureus* SA D22901 (MSSA SA D22901) SERTA DUA STRAIN DERIVATIF DARIPADA MSSA SA D22901

ABSTRAK

Staphylococcus aureus (S. aureus) adalah penyumbang utama jangkitan nosokomial manusia seluruh dunia. Kajian ini berjaya menghuraikan jujukan genomik beberapa spesis *S. aureus*, iaitu *Methicillin-resistant Staphylococcus aureus (MRSA)*, yang dipencilkan daripada pesakit septisemia di Kuala Lumpur (MRSA PR01), pencilan klinikal *Methicillin susceptible S. aureus* (*MSSA*; SA D22901), serta dua strain *ín vitro* CR dan COR yang merupakan strain derivatif yang mempunyai rintangan terhadap antibiotik. CR dan COR telah dijana melalui dedahan SA D22901 kepada antibiotik ciprofloxacin, dan ciprofloxacin dan oxacillin, masing-masing. Jujukan genom *S. aureus* telah diperolehi melalui platform penjujukan genom generasi kedua. Data jujukan ini kemudiannya diproses melalui kombinasi pemasangan *de novo* dan perbandingan (*de novo and comparative assemblies*), penjelasan (*annotation*) dan pengemasan (*finishing*) genom untuk menghasilkan jujukan genom keseluruhan yang lengkap atau deraf. Genom MRSA PR01 telah dihuraikan ke taraf "non-contiguous finished", manakala genom SA D22901 telah dihuraikan ke taraf "improved high quality draft", dan strain-strain derivatif CR dan COR telah melalui pemasangan (*assembly*) secara "reference mapping" dengan menggunakan genom SA D22901 sebagai rujukan. MRSA PR01 didapati mempunyai corak rintangan antibiotik terhadap kebanyakan kelas antibiotik. Analisis pengetipan urutan multilokus (*multi-locus sequence typing*) mengesahkan ia adalah strain ST239 yang merupakan keturunan utama dalam kesihatan di seluruh dunia. SA D22901 mengandungi profil MLST yang novel dan berlainan secara filogenetik dibanding

dengan *S. aureus* yang terdapat di Malaysia. Selepas didedahkan kepada antibiotik, strain rintangan terhadap ciprofloxacin (dinamakan CR), serta ciprofloxacin dan oxacillin (dinamakan COR) berjaya dihasilkan. Analisis genom MRSA PR01 menunjukkan bukti adaptasi untuk terus hidup selepas pendedahan terhadap antibiotik dan gen rintangan antiseptik dikodkan pada elemen bimbit genetik, sisipan kromosom besar dan mutasi titik dalam gen "*housekeeping"*. Analisis perbandingan MRSA klinikal dan MSSA menunjukkan peranan gen yang terlibat dalam pembentukan dinding sel dalam pengambilalihan rintangan methicillin. Analisis varian daripada kajian *in vitro* menunjukkan strain yang dijana memperolehi mutasi yang mungkin mencetuskan mekanisme evolusi rintangan terhadap ciprofloxacin dan oxacillin oleh *S. aureus*. Rintangan ciprofloxacin telah dicapai melalui mutasi dalam dua gen "topoisomerase", manakala mutasi dalam gen yang terlibat dalam pembentukan dinding sel dan pelekatan sel ditemui dalam strain COR yang rintang terhadap oxacillin. Kesimpulannya, kajian ini berjaya menghasilkan jujukan genom keseluruhan beberapa strain *S. aureus* klinikal dan makmal dan mengenalpasti beberapa ciri utama yang terlibat dalam fenomena rintangan terhadap antibiotik.

COMPARATIVE GENOMIC OF METHICILLIN-RESISTANT *Staphylococcus aureus* PR01 (MRSA PR01) AND METHICILLIN-SENSITIVE *Staphylococcus aureus* SA D22901 (MSSA SA D22901) AND METHICILLIN RESISTANT DERIVATIVES OF THE LATTER

ABSTRACT

Staphylococcus aureus is a major contributor of nosocomial infections of humans in the world. This study elucidates the genomic sequences of an MRSA (MRSA PR01) isolate from a patient with septicaemia in Kuala Lumpur, a clinical methicillin-susceptible *S. aureus* (MSSA) isolate (SA D22901), and two *in vitro* generated strains CR and COR which are derivatives of SA D22901 that developed resistance to antibiotics. CR and COR were isolated by exposing SA D22901 to become resistant to the antibiotics ciprofloxacin, and ciprofloxacin and oxacillin, respectively. These *S. aureus* strains were sequenced using a second generation sequencing platform. A combination of *de novo* and comparative assemblies, annotation and genome finishing was then applied to sequenced data to obtain the finished or draft genomes. The MRSA PR01 genome was sequenced to "non-contiguous-finished" status, whereas MSSA SA D22901 was sequenced to "improved high-quality draft" status and the derivatives CR and COR were assembled using reference mapping to MSSA D22901. MRSA PR01 has an extended antibiotic resistance pattern, and is resistant to many classes of antibiotics. Multi locus sequence typing analysis determined it to be of type ST239, a prevalent healthcare lineage across the world. SA D22901 contains a novel MLST profile, and is phylogenetically diverse from other *S. aureus* strains found in Malaysia. Genomic analysis of MRSA PR01 provides evidence of its adaptation to survive in a health care setting through acquisition of drug and antiseptic resistance genes encoded on mobile genetic elements, large chromosomal insertions and point mutations in

housekeeping genes. Comparative analysis of the clinical MRSA and MSSA strains showed the importance of genes involved in cell wall synthesis in the acquisition of methicillin resistance. Variant analysis of the *in vitro* generated strains has uncovered mutations and mechanisms important in the evolution of ciprofloxacin and oxacillin resistance in *S. aureus*. Ciprofloxacin resistance was attained via mutations in two topoisomerase genes, while mutations in genes involved in cell wall synthesis and cell adhesion were found in the oxacillin-resistant COR strain. In conclusion, this study provided the whole genome sequences of several clinical and laboratory generated *S. aureus* strains and identified several key features that are involved in conferring antibiotic resistance.

CHAPTER ONE

INTRODUCTION

1.1 Overview

Genomics is the scientific study of the structure and function of genomes, i.e., the complete set of DNA within a cell of an organism (National Institutes of Health, 2014). Structural genomics focuses on the construction of high resolution genetic, physical and transcript maps for each species(King, Stansfield, and Mulligan, 2007). Functional genomics expands the scope of research to the simultaneous study of large numbers of structural genes that respond to a suitable stimulus (King, Stansfield, and Mulligan, 2007). Evolutionary genomics contrasts the genomes of different species to follow evolutionary changes in genome organization. A common exercise in genomics is the *in silico* investigation of orthologs (King, Stansfield, and Mulligan, 2007).

The pathogenic, Gram-positive bacterium *S. aureus* is a common cause of hospital infections. Antibiotic resistance of *S. aureus* is fast becoming a major concern, as an increasing number of infections are caused by methicillin-resistant *S. aureus* strains that are resistant to most antibiotics in hospitals. The mechanism by which a strain acquires antibiotic resistance has been widely attributed to allelic diversity among strains. Therefore, studying the genome of the bacteria would enhance our knowledge about the mechanisms through which antibiotic resistance is attained.

For this study, genomic DNA from an MRSA strain (source: clinical isolate) was sequenced using massively-parallel sequencing methods to produce millions of short reads that would be able to provide deep coverage of the genome. The reads were then assembled into contigs, annotated and then compared with other MRSA strains

as well as non-pathogenic *S. aureus* strains. The draft genome data can be further completed by employing primer-walking to close gaps between contigs. From there, various functional genetic studies can be undertaken to validate the findings of this study.

Comparative genomic analysis of this MRSA strain to other MRSA strains would provide a clearer picture of the mechanism underlying multi-drug resistance. Comparative genomic analysis will also give a wealth of information on different aspects of the strain, such as epidemiological differences from MRSA strains found in other regions. From this study, it is also possible to develop different methods for detection of MRSA strains in hospitals.

Complementing this would be an *in vitro* study of the process by which antibiotic resistance was attained in *S. aureus.* This is achieved by the generation of multiple-antibiotic resistant *S. aureus* strains from a parental, susceptible strain through antibiotic exposure. These strains were then sequenced using the same method as with the clinical MRSA strain. Comparative analysis of the strains was then performed to detect SNPs or other structural changes in the genome that had occurred after exposure to the antibiotics.

Overall, this study attempts to investigate the occurrence of antibiotic resistance in *S. aureus* through the examination of the genomic sequences of strains that are resistant to various antibiotics. This study had two major components: firstly, the genome sequence of an MRSA strain was determined; secondly, antibiotic-resistant strains were generated in an *in vitro* manner and studied. The resultant laboratory-created MRSA genomes were compared with the genome of the clinical MRSA strain in

order to identify genetic elements of importance to antibiotic resistance. This is elaborated in further detail in Section 1.9.

1.2 Pathogenicity of *Staphylococcus aureus* and the problem of increasing resistance to antibiotics

The pathogenic, Gram-positive bacterium *Staphylococcus aureus* is a common cause of hospital acquired infections. This bacterium has been characterized as an opportunistic pathogen, and is extremely adaptive to its environment and host. About a third of humans are estimated to be asymptomatically colonized by *S. aureus*, hence it is considered normal flora (Gorwitz et al., 2008). Clinically important, it causes a wide variety of diseases in humans, e.g., nosocomial and community-acquired infections, toxic shock syndrome, septicaemia, endocarditis and pneumonia (Lowy, 1998).

Pathogenicity of the organism is compounded by the resistance mechanism of *S. aureus* to antibiotic treatment. Antibiotic resistance of *S. aureus* is a major concern, as an increasing number of infections are caused by methicillin-resistant *S. aureus* (MRSA) strains that are resistant to most antibiotics in hospitals. It was estimated that 90% of human *S. aureus* strains are resistant to penicillin (Klein, 2009). Strains of this class may also be resistant to other types of antibiotics.

Historically, there have been four waves of antibiotic resistance in *S. aureus* each corresponding to the introduction of novel antibiotics in hospitals and the community. The first wave began in the 1940s with the incidence of penicillin resistant *S. aureus* (Barber and Rozwadowska-Dowzenko, 1948; Kirby, 1944) that

evolved as a response to the introduction of penicillin. Introduction of methicillin marked the onset of the second wave of resistance. This wave featured strains that were resistant to broad spectrum antibiotics which include penicillins, cephalosporins and carbapenems. This was different from the wave before, where strains had narrow resistance spectrum (Chambers and DeLeo, 2009). From this wave, a number of notable clones emerged, most prominently the "archaic" clone, one of the most successful of all MRSA lineages. The archetypal strain, COL is a member of this group and is the most studied MRSA isolate. This archaic clone of MRSA disseminated across hospitals in Europe until the 1970s, but was largely contained within the region and never gained a foothold across the rest of the world or into the community. The third wave of resistance occurred in the late 1970s and marked the rapid spread of successful MRSA lineages across the world. This wave established the successful lineages seen today, such as those shown in Table 1.1. The latest and ongoing fourth wave of resistance features MRSA invasion of the community, producing community-associated MRSA (CA-MRSA) clones (Henry F. Chambers and DeLeo 2009).

Table 1.1. Lineages of common nosocomial MRSA.

Clonal Complex	**Multilocus sequence type**	**Common names for specific MRSA clone**	**Comment**
CC5	ST5	USA100 and New York/Japan clone	Most common US healthcare-associated MRSA, SCC*mec*II
	ST5	EMRSA-3	*SCCmec*I
	ST5	USA800/Paediatric clone	Prevalent in Argentina, Columbia, United States, SCC*mec*IV
	ST5	HDE288/Paediatric clone (Portugal)	SCC*mec*VI
CC8	ST250	Archaic	First MRSA clone identified, COL strain as an example; SCC*mec*I
	ST247	Iberian clone and EMRSA-5	Descendent of COL-type strains, SCC*mec*I
	ST239	Brazilian/Hungarian clone	SCC*mec*III
	ST239	EMRSA-1	Eastern Australian epidemic clone of 1980s, SCC*mec*III
	ST8	AUS-2 and AUS-3	SCC*mec*II
	ST8	Irish-1	Common nosocomial isolate in the 1990s in Europe and the US
	ST8	USA500 and EMRSA-2,6	SCC*mec*IV
CC22	ST22	EMRSA-15	International clone, prominent in Europe and Australia, *SCCmec*IV
CC30	ST36	USA200 and EMRSA-16	Single most abundant cause of MRSA infections in United Kingdom; second most common cause of MRSA infections in US hospitals in 2003, SCCmecII
CC45	ST45	USA600 and Berlin	SCC*mec*II

Note: Adapted from Chambers and DeLeo, 2009.
Legend: SCC*mec*[number] refers to the type of SCC*mec* (Staphylococcal chromosomal cassette *mec*) found. This is elaborated more in Section 1.3.1.

1.3 The case for a genetic basis as the cause of adaptability of the organism to its environment

Staphylococcus aureus is extremely dangerous as a pathogen because of its ability to adapt quickly to external pressures. A major way in which this ability manifests itself is through its ability to attain resistance to antibiotics quickly. The main mechanism by which this occurs is through Horizontal Gene Transfer (HGT) (Figure 1.1).

Horizontal gene transfer is the process by which transfer of mobile genetic elements (MGEs) occurs between cells (Malachowa and DeLeo 2010). MGEs are short DNA sequences that frequently encode proteins that mediate the movement of DNA between bacterial cells or within genomes. Many resistance genes are carried on MGEs that will then act as vectors that transfer these genes to other bacteria of the same species, or of different genus or species. MGEs may consist of insertion sequences, transposons, phages, pathogenicity islands, and chromosome cassettes. In *S. aureus*, MGEs account for approximately 15-20% of the genome (Lindsay and Holden, 2006). This is contrasted with the typical method by which genetic material is passed on from parent to progeny via sexual or asexual reproduction, termed vertical gene transfer. HGT in prokaryotes occurs mainly through three mechanisms: Conjugation, transformation and transduction.

Conjugation is the method by which transfer of genetic bacteria occurs between independently replicating bacterial cells, either by direct cell-to-cell contact or by a bridge-like connection between the two cells (Lederberg and Tatum, 1946). Conjugation is commonly associated with the transfer of plasmids between cells. Transformation is the natural uptake of exogenous genetic material from the cells'

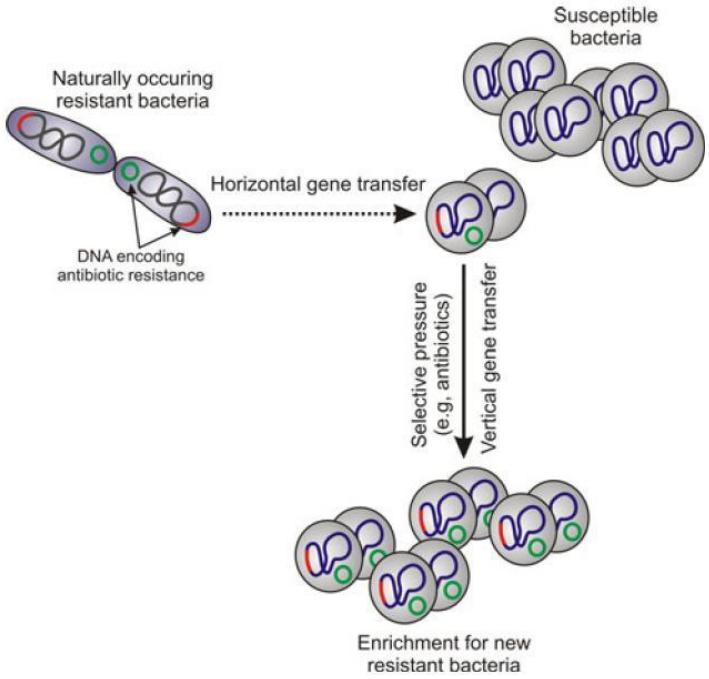

Figure 1.1. Horizontal Gene Transfer (HGT) in bacteria. The transfer of DNA encoding antibiotic resistance from bacteria that is naturally resistant to the antibiotic can be passed onto a susceptible bacterial cell via horizontal gene transfer. Due to the presence of selective pressures, the newly integrated DNA is then inherited by progeny, resulting in enrichment and proliferation of newly resistant bacteria. Adapted from (Malachowa and DeLeo, 2010). Appears as Figure 1 in original text.

surroundings through the cell membrane (reviewed by Chen and Dubnau, 2004). In fact, it was the observation of transformation of virulence determinants from *pneumococci* into infected mice that provided the first evidence that HGT could occur (Griffith, 1966; Thomas and Nielsen, 2005). There are only a number of prokaryotes capable of natural transformation, *S. aureus* among them. The last mechanism, transduction is the process by which DNA is transferred from one bacterium to another by a virus, i.e., a bacteriophage (Zinder and Lederberg 1952). This method does not need physical contact between the donor and acceptor cell. Taken together, these three mechanisms of horizontal gene transfer explain the genetic basis by which antibiotic drugs become ineffective due to the transfer of antibiotic-resistance genes between bacteria (Frost et al., 2005).

The bacterial genome consists of two different components: a core genome, containing all genes necessary for cell survival and an accessory genome, containing genes that encode for proteins required for adaptation of bacteria to their ecological niche. Genes that make up the core genome are highly conserved and are present in all strains of that species, whereas genes in the accessory genome tend to vary between strains and tend to be unique to that particular strain. Accessory genome differs from the core genome typically in GC content because they are usually obtained from other species through horizontal gene transfer mechanisms (Hacker and Kaper, 2000). Most MGEs are contained within the accessory genome.

Staphylococcal MGEs consists of plasmids, transposons (Tn), insertion sequences (IS), bacteriophages, pathogenicity islands, and staphylococcal cassette chromosomes (Malachowa and DeLeo, 2010). Most genes encoded by MGEs are under the control of global regulators located within the core genome (Malachowa and DeLeo, 2010).

1.3.1 Genome features of *Staphylococcus aureus*

The analyses of *Staphylococcus aureus* genomes have revealed several key features that characterize the organism. Central to methicillin resistance is the *mecA* gene, located on a genomic region known as the Staphylococcal chromosomal cassette (SCC*mec*) (Hiramatsu et al., 2001). Most molecular tests for MRSA are designed to detect the presence of the *mecA* gene in *S. aureus* (Otte, Jenner, and Wulffen, 2005).

SCC*mec* is a mobile genetic element characterized by the presence of terminal inverted and direct repeats, a set of site-specific recombinase genes (*ccrA* and *ccrB*) and the *mecA* gene complex (Hiramatsu et al., 2001; Ito et al., 2003). This element is integrated into the *orfX* gene (Berger-Bächi and Rohrer, 2002). To date, the SCC*mec* elements can be characterized into at least eight distinct types, with numerous subtypes and presumably many more to be discovered with the sequencing of more *S. aureus* strains (IWG-SCC, 2009). The *mecA* gene is responsible for the conference of resistance to β-lactam antibiotics, and is a major key feature present in all methicillin-resistant *Staphylococcus aureus* strains (Ubukata et al., 1989; Chambers, Hartman, and Tomasz 1985). The *mecA* gene encodes a low affinity penicillin-binding protein (PBP), PBP2a that is expressed in addition to the complement of native PBPs and is the sole PBP in the presence of β-lactam antibiotics that is capable of maintaining cell wall integrity (Berger-Bächi and Rohrer, 2002).

In addition to the above, specific to the *Staphylococcus* genome are regions within the *Staphylococcus aureus* genome that were thought to have arisen via horizontal gene transfer (Kuroda et al., 2001). These genomic regions have been termed νSaα, encoding multiple exotoxins (*set*) (Williams et al., 2000; Jarraud et al., 2001) and lipoproteins (*lpl*), and νSaβ, encoding multiple *set* and serine protease homologues

(*spl*) (Reed et al., 2001). Although stable (they are found in all *S. aureus* strains that have been sequenced so far), they vary in genetic composition and often carry genes involved in pathogenicity (Lindsay and Holden, 2006).

Another class of *Staphylococcus* genomic features are super antigen encoding pathogenicity islands (SaPIs.), approximately 15 kb genomic regions which notably encode a number of virulence genes, including the superantigens toxic shock syndrome toxin-1 (*tst*) and enterotoxins B and C (*seb, sec*) (Yarwood et al., 2002). Superantigens are able to elicit a non-specific T-cell response, which then leads to massive cell expansion and shock. SaPIs themselves have been shown to be able to transfer horizontally in the presence of 'helper' phage (Lindsay et al., 1998).

As mentioned in Section 1.3, MGEs play a vital role in the adaptability of the *S. aureus* genome. In *S. aureus,* a variety of MGEs are responsible for the antibiotic resistance and pathogenicity of the organism apart from those mentioned earlier. These include plasmids, transposons and bacteriophages (Malachowa and DeLeo, 2010).

Plasmids in *S. aureus* are divided into three types, class I-III based on their size and ability to conjugate (Lindsay and Holden, 2006). Class I plasmids are the smallest in size (<5kb) and have the highest copy number. They have been known to encode several resistance genes, e.g., tetracycline resistance is encoded by pT181 in COL (Khan and Novick, 1983; Gill et al., 2005). Class II plasmids are larger (up to 40 kb) and usually contain resistance elements that were originally on transposons that have integrated into the plasmid. Class I and II plasmids are usually transferred by transduction. Class III plasmids are similar to Class II plasmids, except that they often also contain transfer (*tra*) genes that allow conjugative transfer of the plasmid

between bacterial isolates at low frequency on solid surfaces (Thomas and Archer, 1989). This occurs because of the large size (up to 60 kb) of class III plasmids which may be too large to be transferred by transduction.

In *S. aureus*, transposons are another MGE component that often encode for resistance genes. Transposons encode a transposase gene, the product of which catalyses excision and/or replication of the element, as well as integration. Horizontal transfer of transposons to other *S. aureus* cells is presumably by "piggybacking" onto another MGE, most likely a plasmid, which in turn can then be transferred by transduction or conjugation. An example of a transposon commonly found in *S. aureus* genomes is Tn*554*, which contains *erm* genes that encode for resistance to erythromycin (Phillips and Novick, 1979). Tn*554* is also capable of integrating in multiple sites around the chromosome. Other examples include Tn*552* which contains the *blaZ* gene that is associated with β-lactamase resistance (Rowland and Dyke, 1989); and Tn5801, a large transposon integrated into the Mu50 genome and encodes tetracycline resistance (*tetM*) (Kuroda et al., 2001). When the transposable elements lack additional genes, they are known as insertion sequences.

The last major component of MGEs that make up a big part of the *S. aureus* accessory genome is bacteriophages, which are viruses that infect and replicate within bacteria. Bacteriophage genomic areas mainly contain virulence factors, such a *sak* (staphylokinase A thrombolytic enzyme) (Collen, 1998), *chp* (chemotaxis inhibitory protein and inhibitor of leucocyte migration) (Haas et al., 2004), *sea* (enterotoxin A, a common cause of food poisoning) (Betley and Mekalanos, 1985), *eta* (exfoliative toxin A, a cause of scalded skin syndrome) (Yamaguchi et al., 2000) and *lukSF-PV* (Panton–Valentine leucocidin implicated in haemolytic pneumonia and severe skin and soft tissue infections) (Kaneko et al., 1998; Narita et al., 2001).

Phages that have integrated into the bacterial chromosome are called prophages. These lysogenic phages are "dormant" and are replicated as part of the entire bacterial chromosome and passed to daughter cells during cell division (vertical transfer). Distribution of prophages in *S. aureus* strain populations has been shown to correlate with the clonal types of *S. aureus* (Goerke et al., 2009).This was shown in a study that classified *Staphylococcal* phages according to polymorphism of the integrase gene, which found that by linking the phage content to dominant *S. aureus* clonal complexes, the distribution of bacteriophages varied remarkably between lineages, indicating restriction-based barriers (Goerke et al., 2009). A comparison of colonizing and invasive *S. aureus* strains revealed that *hlb*-converting phages were significantly more frequent in colonizing strains (Goerke et al., 2009).

Horizontal transfer of virulence factors by bacteriophage occurs in two ways: first, by transfer and integration of a bacteriophage that encodes a virulence gene on its genome (phage conversion); secondly, by generalized transduction. Generalized transduction (i.e., the transfer of bacterial DNA and not the phage DNA via transduction) in *S. aureus* is thought to be widespread and is the only feasible mechanism for the horizontal transfer of many non-phage MGEs (Lindsay and Holden, 2006).

Table 1.2. Mobile genetic elements (MGEs) in *S. aureus*.

Type of MGE	Mode of transfer	Examples	Associated resistance/virulence elements	References
Plasmid	Conjugation (Class III only), Generalized transduction	pT181, pSAS	*tet, blaZ, cadD*	Massidda et al., 2006; Hou et al., 2007; Bismuth et al., 1990
Transposon	"Piggybacking" onto another MGE, e.g., plasmid	Tn*554*, Tn*552*, Tn*5801*	*ermA*, *spc*, *blaZ*, *tetM*	Westh et al., 1995; Lelièvre et al., 1999; Rowland and Dyke, 1990, 552; Bismuth et al., 1990
Bacteriophage	Transduction	φ11, φETA	*sak*, *chp*, *sea*, *eta*, *lukSF-PV*	Collen, 1998; Haas et al., 2004; Betley and Mekalanos, 1985; Yamaguchi et al., 2000; Kaneko et al., 1998; Narita et al., 2001
Genomic island	Unknown	νSaα, νSaβ	*set*, *lpl*, *spl*, *lukDE*	Williams et al., 2000; Jarraud et al., 2001; Reed et al., 2001
Pathogenicity island	Transduction	SaPI1, SaPImw2	*ear*, *tst* ,*seb* ,*sec*	Novick 2003; Yarwood et al., 2002
Staphylococcal cassette chromosomes	Unknown	SCC*mec*	*mecA*, *far*, *cadA*	Ubukata et al., 1989; Chambers, Hartman, and Tomasz, 1985; Dietrich, 1992; M. Holden et al., 2004)

1.4 Biological mechanisms of resistance

In general, there are various ways in which a bacterial cell is able to attain resistance to an antibiotic. In *Staphylococcus aureus*, the main mechanism to which resistance is attained is dependent on the mode of action taken by the antibiotic in question. For instance, flouroquinones such as ciprofloxacin which inhibit bacterial growth by inhibiting the activity of DNA topoisomerase IV or DNA gyrase, can be neutralized by the presence of mutations that occur in the quinolone-resistance determining region (QRDR) of topoisomerase IV (encoded by the genes *grlA* and *grlB*) and DNA gyrase encoded by the genes *gyrA* and *gyrB* (Kaatz and Seo, 1997). However, fluoroquinolone resistance can also be mediated by drug efflux, a mechanism that is less well characterized (Tanaka et al., 2000).

β-lactam resistance in *Staphylococcus aureus* is conferred by a few genetic elements. By far the most notable and studied is *mecA* operon on the staphylococcal cassette chromosome mec (SCC*mec*). The *mecA* gene encodes an altered penicillin binding protein (PBP2a or PBP2') that has a lower affinity for binding β-lactams (penicillins, cephalosporins, and carbapenems) (Ubukata et al., 1989). This allows for resistance to all β-lactam antibiotics, and obviates their clinical use during MRSA infections.

Aminoglycoside antibiotics, such as kanamycin, gentamicin and streptomycin were once effective against staphylococcal infections until strains evolved mechanisms to inhibit the action of aminoglycosides, which occurs via protonated amine and/or hydroxyl interactions with the ribosomal RNA of the bacterial 30S ribosomal subunit. There are three main mechanisms of aminoglycoside resistance which are currently and widely accepted: aminoglycoside modifying enzymes, ribosomal mutations, and active efflux of the drug out of the bacteria.

Resistance to macrolides such as erythromycin arises due to two distinct mechanisms. Erythromycin-resistant methylase is encoded by *erm* genes. Resultant structural changes to rRNA prevent macrolide binding and allow synthesis of bacterial proteins to continue. The presence of erm gene results in high-level resistance. Modification of the mechanism whereby antibiotics are eliminated from the bacteria also brings about resistance. Bacteria carrying the gene encoding macrolide efflux (*mefE*) display relatively low-level resistance.

1.5 *In vitro* selection and characterization of antibiotic-resistant *Staphylococcus aureus*

In order to explore further the mechanisms by which *Staphylococcus aureus* attains antibiotic resistance, it is possible to generate strains that attain resistance to different antibiotics *in vitro*, through the selection of resistant mutants (from an originally-susceptible strain) by passage in broth culture containing subinhibitory concentrations of the antibiotic of choice. With the creation of these strains, it is then possible to analyse the genomes of these strains prior to and after exposure to that particular antibiotic using the same approaches whereby MRSA PR01 was characterized.

1.6 Sequencing is an important tool for biological research

DNA sequencing technology has advanced in leaps and bounds since the announcement of the completion of the Human Genome Project in 2000. This historic discovery ushered in an era of huge interest in genomics and molecular

biology for the purposes of biological and clinical research. Currently, a new wave of next generation sequencers provides a slew of different massively parallel sequencing strategies and platforms. One common platform is sequencing generated by the Illumina platform which utilises a 'sequencing by synthesis' approach to sequence DNA (Bentley et al., 2008). This approach is a reversible terminator-based method that enables detection of single bases as they are incorporated into extending DNA strands. A fluorescently-labelled terminator is imaged as each dNTP is added and then cleaved to allow incorporation of the next base. All four reversible terminator-bound dNTPs are present during each sequencing cycle, minimizing incorporation bias.

There are many applications in which DNA sequencing plays a role, most notably in the arena of genetics and genomics. For example, whole genome sequencing can be used as a tool to understand the evolution of antibiotic resistance. Mwangi et al. (2007) used this approach to follow the genetic changes that occurred at the start of vancomycin therapy and then sequenced the genome of a non-susceptible isolate recovered 3 months later. This study identified 35 point mutations that conferred a VISA phenotype and defined loci critical for multidrug resistance. These studies provide an illustration of how whole genome sequencing can be used to elucidate the molecular genetic basis of emerging antibiotic resistance.

Another approach that has utilised whole genome sequencing to great effect is in the tracking of transmission in epidemiological outbreaks. Currently, genotyping is performed on *S. aureus* strains in order to type and characterize between different strains. One such method is multilocus sequence typing (MLST), a sequence based genotyping method that characterizes *S. aureus* strains based on the sequence of 450 bp fragments of seven housekeeping genes. Isolates are grouped into clones based on

the allelic profile of the sequenced genes, and assigned a unique sequence type (ST). Sequence types that share at least five alleles in common with each other can be further grouped into clonal complexes (CCs) (Feil et al., 2001). Another similar method of MRSA typing is by the single locus DNA-sequencing of repeat regions of the *Staphylococcus* protein A gene (*spa*), termed *spa* typing (Koreen et al., 2004). These methods serve to provide identification and classification to different *S. aureus* strains.

The aforementioned current molecular methods such as MLST and *spa* typing used to track and differentiate MRSA strains do so with limited discriminatory power. This is problematic, in light of the fact that single genotypic variants commonly predominate in the pathogen population. It has been shown, for example, that most human infecting MRSA strains are derived from one of 10 independent clonal complexes (Feil et al., 2003; Lindsay et al., 2006). Several epidemiological studies have used the combination of whole genome sequence data of MRSA with epidemiological information to trace the transmission history of an MRSA outbreak (Nübel et al., 2013; Eyre et al., 2012; Köser et al., 2012). These studies have shown the ability of bacterial whole genome sequencing to distinguish between strains of MRSA collected within a short timescale in a confined geographic location (e.g. a hospital unit), hence supporting its potential as a routine tool in clinical bacteriology.

1.7 Bioinformatics overview

The term "bioinformatics" was first coined by the theoretical biologist Paulien Hogeweg in 1970, who at the time defined it as "the study of informatics processes in biotic systems" and used the term to describe the research themes that he and his

compatriots were working on, i.e., to understand biological systems as information processing systems (Hogeweg, 2011). With the advent of massive sequencing projects in the 1990s and 2000s, and the shift in molecular biology to become a heavily "data-driven" science, the definition of bioinformatics changed and developed into the field that it is known currently, as that of an "interdisciplinary toolset for applying computer science, mathematics, and statistics to the classification and analysis of biological information. At the highest level, bioinformatics is used to analyse large datasets to help answer biological questions, both in fundamental biology and in the biology that underlies disease"(Rothberg, Merriman, and Higgs, 2012).

Bioinformatics is also defined as conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying informatics techniques (derived from disciplines such as applied maths, computer science and statistics) to understand and organise the information associated with these molecules, on a large scale (Luscombe, Greenbaum, and Gerstein, 2001). In short, bioinformatics is a management information system for molecular biology and has many practical applications (Luscombe, Greenbaum, and Gerstein, 2001). Current work in genomics, proteomics, and many other "omics" research has produced a myriad of available software for many types of analyses.

In the genomics era, biologists must attain, in addition to technical ability at the lab bench, considerable expertise in handling large datasets in arcane command line interfaces, and a considerable knowledge or computer programming or coding is required. The typical workflow for the elucidation of a bacterial genome is depicted in Figure 1.2.

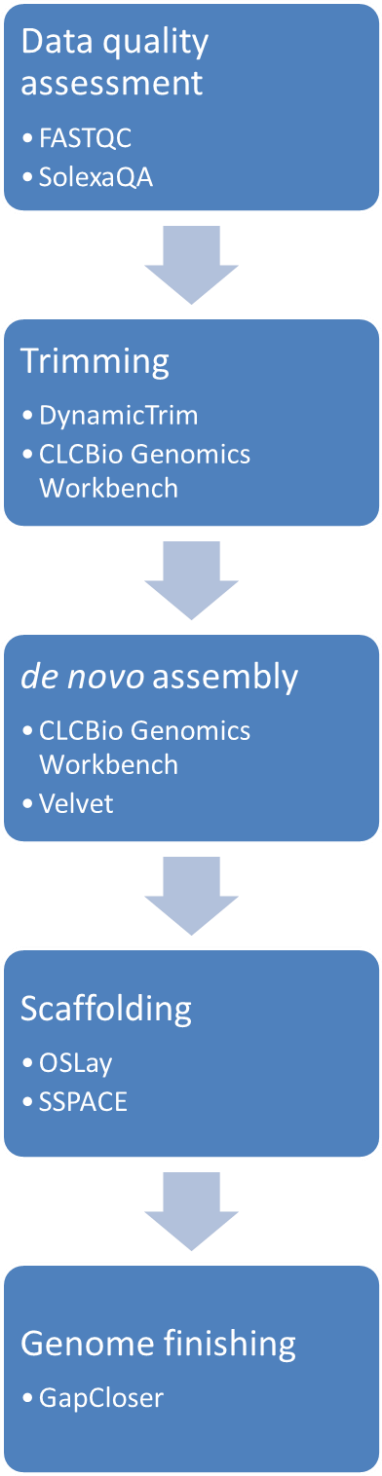

Figure 1.2. An example of a workflow of the data analyses needed to elucidate the whole genome sequence of bacteria.

With the advent of next-generation sequencers there has been an exponential increase in the deposit of draft genomes into databases such as NCBI. However, these genome data cannot be held to the same standard as some of them may have been produced from low-quality data or through mis-assembly. Chain et al., (2009) proposed a new set of standards that fit the community-defined categories of standards that reflect the quality of the genome sequence based on the understanding of the available technologies, assemblers and efforts to improve upon drafted genomes. In this paper, they published a set of categories upon which genomes could be distinguished from each other based on combinations of technology, chemistry, assembler or improvement and/or finishing process. A summary of the categories is found in Table 1.3.

To attain these standards, several criteria must be considered including quality, coverage and assembly. Data quality is important in ensuring the validity of results, as is discussed further in Section 1.7.1. Coverage, or 'depth' of sequencing, refers to the number of times each nucleotide within the genome is expected to be sequenced taking into account the read length and number (Lander et al. 2001). It can be calculated from the length of the original genome (G), the number of reads (N), and the average read length (L) as

$$\boldsymbol{N \times L / G}$$

Equation 1

Coverage must be taken into account in order to generate high-quality, unbiased and interpretable data; however, it must also be balanced against the consideration of sequencing cost, which can be substantial (Sims et al., 2014). Key considerations in assembly will be discussed in Section 1.7.2. Complexity of the genome is also a main

factor that affects the quality of the resulting genome sequence. In this project, we aim to produce a genome sequence of at least 'Non-contiguous-finished' quality.

Table 1.3. Categories of community-defined standards of assembled genomes. Adapted from (Chain et al. 2009)

Category	Description
Standard Draft	• Minimally or unfiltered data, from any number of different sequencing platforms, that are assembled into contigs. • Minimum standard for a submission to the public databases. • Likely harbour many regions of poor quality and can be relatively incomplete; may have contain contaminating sequence data. • Least expensive to produce and still possesses useful information.
High-Quality Draft	• Overall coverage representing at least 90% of the genome or target region. • Efforts should be made to exclude contaminating sequences. • Still considered a draft assembly with little or no manual review of the product. • Sequence errors and mis-assemblies are possible, with no implied order and orientation to contigs. • Appropriate for general assessment of gene content.
Improved High-Quality Draft	• Additional work has been performed beyond the initial shotgun sequencing and High-Quality Draft assembly, by using either manual or automated methods. • Contains no discernible mis-assemblies and should have undergone some form of gap resolution to reduce the number of contigs and supercontigs (or scaffolds). • Undetectable mis-assemblies are still possible, particularly in repetitive regions.

Table 1.3. (Continued)

Category	Description
	• Low-quality regions and potential base errors may also be present. • This standard is normally adequate for comparison with other genomes.
Annotation-Directed Improvement	• May overlap with the previous standards, but the term emphasizes the verification and correction of anomalies within coding regions, such as frameshifts, and stop codons. • Used in cases involving complex genomes where improvement beyond this category fails to outweigh the associated costs. • Gene models (gene calls, including intron-exon determination for eukaryotes) and annotation of the genomic content should fully support the biology of the organism and the scientific questions being investigated. • Exceptions to this gene-specific finishing standard should be noted in the submission. • Repeat regions at this level are not resolved, so errors in those regions are much more likely. • This standard is useful for gene comparisons, alternative splicing analysis, and pathway reconstruction as most gene annotation at this stage is curated and of good quality.
Noncontiguous Finished	• High-quality assemblies that have been subjected to automated and manual improvement and where closure approaches have been successful for almost all gaps, mis-assemblies, and low-quality regions. • Attempts have been made to resolve all gap and sequence uncertainties, and only those recalcitrant to resolution (e.g. repetitive or intractable gaps) remain (with notations in the genome submission as to the nature of the uncertainty). • Appropriate for most analyses.

Table 1.3. (Continued)

Category	Description
Finished	• Refers to the current gold standard; genome sequences with less than 1 error per 100,000 base pairs and where each replicon is assembled into a single contiguous sequence with a minimal number of possible exceptions commented in the submission record. • All sequences are complete and have been reviewed and edited, all known mis-assemblies have been resolved, and repetitive sequences have been arranged in order and correctly assembled. • Remaining exceptions to highly accurate sequence within the euchromatin are commented in the submission. • The Finished product is appropriate for all types of detailed analyses and acts as a high-quality reference genome for comparative purposes.

1.7.1 Data quality assessment

Once data have been obtained, it is important to assess the quality of the reads that have been generated. The measurement of quality varies from platform to platform, but in the Illumina platform, quality scores measure the probability that a base is called incorrectly. With sequencing by synthesis (SBS) technology, each base in a read is assigned a quality score by a phred-like algorithm, similar to that originally developed for Sanger sequencing experiments (Ewing and Green 1998). The quality score of a given base, Q, is defined by the following equation:

$$\mathbf{Q = -10\log_{10}(e)}$$

Equation 2

where e is the estimated probability of the base call being wrong (Ewing and Green 1998). Thus, a higher quality score indicates a smaller probability of error. As shown in Table 1.4, a quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99%.