
DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Identification of Vaccine Candidates Against Staphylococcus Aureus: An in Silico Reverse Vaccinology Approach

Citation

Sharma, Shekhar. 2019. Identification of Vaccine Candidates Against Staphylococcus Aureus: An in Silico Reverse Vaccinology Approach. Master's thesis, Harvard Extension School.

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:42004133

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Identification of Vaccine Candidates against *Staphylococcus aureus*: An *in silico* Reverse Vaccinology Approach

Shekhar Sharma

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2019

Copyright 2019 Shekhar Sharma

Abstract

The goal of this research was to identify novel antigenic protein targets for designing a vaccine against *Staphylococcus aureus*. *S. aureus* is both a commensal bacterium and a human pathogen. It asymptomatically colonizes the epithelium of 25-50% of the population. The factors that make a benign patch of *S. aureus* virulent are not well understood. However, the event is often preceded by an increase in coagulase secretion. Bacteremia and endocarditis are the serious outcomes of an *S. aureus* infection. Presently, there is no vaccine for the pathogen though SA4Ag, under development, is promising. In this *in silico* experiment, 330 candidate proteins were identified using the principles of reverse vaccinology which uses genomic data for computing the antigenic targets. Custom developed computer code and online bioinformatics tools were used to conduct this experiment. 14,322 whole genome sequences (WGSs) of *S. aureus* were downloaded from GenBank and allied datacenters. The data was curated for completeness and diversity. 6,538 ORFs were extracted from the WGSs using biostatistical algorithms. The ORFs were translated into proteins. In parallel, 1,574 genes were extracted from the WGSs using bioinformatics tools. The location of ORFs were matched with that of known genes. The proteins translated from ORFs were compared with the co-located genes to verify the ORF detection algorithm and identify potential unknown genes. Based on the probability that a protein translated from an ORF might translocate to the cell surface, 1,045 candidate proteins were chosen. Finally, 330 proteins were identified that could be candidates for designing a vaccine against *S. aureus*. The results supported the objective of this research, namely, to find antigenic proteins

expressed by *S. aureus*. The identified proteins must be validated *in vitro*. The future studies could include investigating ORFs in the context of the whole genome, IL-17 and coagulase pathways and membrane bound lipoproteins and glycoproteins.

Dedication

I dedicate this dissertation to my parents who though not having any formal education themselves understood its value and to my family for their unwavering support.

Acknowledgements

I will like to thank my dissertation committee, Dr. Li Hao and Dr. Steven Denkin for their support over the last two years as an amorphous idea took shape into its present form of the completed research. Dr. Hao was there with me patiently as I ran into dead ends while trying to find the solution that worked. I have learned so much through our conversations over this journey. Her suggestions in writing the manuscript were invaluable.

Finally, thanks to all the great professors and teaching assistants at Harvard University who have taught me so much about the field of biotechnology.

Table of Contents

List of Tables

List of Figures and Graphs

Chapter I

Introduction

Staphylococcus aureus: a commensal bacterium and a human pathogen

Staphylococcus aureus, colloquially referred as Staph, is a facultative pathogen that colonizes almost the entire body. It is found in especially high concentrations in the nares (Holtfreter, Kolata, & Broker, 2010). It normally resides asymptomatically as a commensal bacterium in approximately 25-50% of the human population (Read, et al., 2018). The bacteria may colonize a patch of skin silently for a long period, even a lifetime. However, sometimes the same colony can turn virulent. The presentation of the infection can range from a minor skin irritation that resolves without medication to life-threatening bacteremia, i.e., the presence of bacteria in the blood and endocarditis, i.e., the inflammation of the endocardium epithelia (Anderson, et al., 2012).

The factors that transform a benign commensal patch of Staphylococcus into a pathogen are poorly understood. The onset of coagulation cascades in the nasopharynx reservoirs is one of the possible causes of the virulence of Staph (Salazar, et al., 2014). Coagulase contributes to the catalysis of the coagulation cascade. *S. aureus* species are coagulase-positive though they do not express enough coagulase constitutively to initiate coagulation, Upregulation of the coagulase gene (*coa*) portends the onset of pathogenesis. In comparison, most non-pathogenic staphylococci forming the human cutaneous microbiome are coagulase-negative (Chambers, 2012).

Epidemiology of *Staphylococcus aureus*

Staphylococcus aureus normally resides in the commensal microbiome over almost entire epithelia of humans including the epidermis and internal cavities such as pleura and pulmonary tract. The pathological manifestations of a Staph infection range from simple boils to severe impetigo. In extreme cases it can lead to bacteremia and endocarditis, especially in the nosocomial settings. Bacteremia and viremia (presence of bacteria and virus in the blood, respectively) are abnormal events that lead to a severe immune response including sepsis and septic shock. Staphylococcal bacteremia has an added hazard. The tightly bound coagulase on the staphylococcal surface can form a complex with the blood borne prothrombin and cause blood clots. The *S. aureus* bacteremia (SAB) has a high mortality rate. In the industrialized world, the incident rate is 10-30 cases per 100,000 of the population. While the risk groups show ethnic trends, persons on either end of life seem to be most susceptible (Tong, Davis, Eichenberger, Holland, & Fowler, 2015).

The epidemiology of *S. aureus* has gained urgency because the microbe is acquiring resistance to every known antibiotic, starting with penicillin in the 1950s to the more recent antimicrobial drugs including vancomycin and linezolid. The methicillin resistant variant of *S. aureus*, known as MRSA has become widespread in hospitals worldwide (Giersing, Dastgheybb, Modjarradc, & Moorthy, 2016). Relatively fewer cases of MRSA have been also reported for the community acquired infections. Interestingly, methicillin was developed to counter the growing resistance of Staph to penicillin. The high drug resistance among bacteria can be attributed to their short lifecycle and therefore, a high DNA duplication rate. *S. aureus* divides approximately

every half hour *in vitro*. It has approximately 2.8 million nucleotide base pairs in its genome. At the customarily accepted rate of one error per 10^{10} base pairs of the replicating DNA, statistically speaking, the genome of an *S. aureus* specimen can have over 30 mutations in just 30 hours of the post-infection incubation (Pope, O'Sullivan, Mchugh, & Gillespie, 2009).

Vaccines, Toxins, Toxoid Vaccines and Antitoxins

Vaccines are the most effective of means of preventing diseases among large populations of people. Vaccines prevent approximately six million deaths worldwide annually and save an estimated 386 million person-years of life (Vernikos & Medini, 2014).

Traditionally, the antibacterial vaccines target the antigens located on the surface of the bacteria or the toxins released by the microbes during an infection (Fowler & Proctor, 2014). The intracellular components such as the genomic DNA or plasmids are rarely used for designing a vaccine because the cytosolic antigens are inaccessible in an intact bacterium. The bacterial genome comprises a single circular chromosome of DNA, which is very different from the chromosomes of mammals. Therefore, the bacterial chromosome would be antigenic if it were visible to the immunocytes. The circular DNA would constitute a pathogen-associated molecular pattern (PAMP). In that case it would be targeted by the host immune system. The pattern recognition receptors such as the toll-like receptors recognize the allogenic PAMPs and initiate the innate immune response (Akira, Uematsu, & Takeuchi, 2006).

The vaccines utilizing the surface antigens deploy live-attenuated, killed or inactivated bacteria. Vaccines against mumps and measles are the examples of the former

method while the anti-cholera vaccine is made from the killed or inactivated specimen. Alternately, only the antigenic parts of microbe surface can be used for creating a vaccine. These subunit vaccines are better than using the whole organism because they are not virulent. The host is immunized by injecting the intact bacteria or the subunits thereof. It elicits an adaptive immune response in the recipient and builds antibodies against the antigens (Baxter, 2007).

While toxin is an encompassing term meaning a biochemical that is deleterious to an organism, the immunologically significant toxins are released by the bacteria. The exotoxins are secreted by live bacteria while the endotoxins are released during the lysis of the bacteria. Diphtheria and spasmogenic toxins secreted by *Corynebacterium diphtheriae* and *Clostridium tetani*, respectively are two such exotoxins. The toxoids are the biologically inactivated forms of the toxins. Incubation with formalin is the most common method of inactivating the toxin. The toxoids retain the antigenic epitopes of the original toxins without the associated virulence. Therefore, they are suitable for formulating vaccines that can elicit the protective immune response. The toxoid vaccines conjugate with the B cell receptors (BCRs) to produce the antitoxin antibodies (Bröker, Mrochen, & Péton, 2016). The trivalent DTaP (diphtheria, tetanus and acellular pertussis) toxoid vaccine is commonly used for immunizing the children. As opposed to the toxoid vaccines that are the immunogens, the antitoxins are the antibodies created *in vivo* in response to the vaccines. The antitoxins are then harvested from the host, purified and packaged as pharmaceuticals. Similar to other vaccines, the toxoids have a longer gestation period and protection. In comparison, the readymade antibodies of the antitoxins are quick acting, though short lived.

Novel assays are now being developed as a replacement for formalin. Formalin has been a mainstay for producing the toxoids for over a century. The formaldehyde in formalin induces conformational and structural changes in the toxoids in addition to the inactivation. As a result, the antibodies resulting from the toxoids vaccine have a lower avidity to the toxins. Some alternate non-formalin-based assays retain a higher efficacy in the end product. For example, the toxins inactivated by alkylating with iodoacetamide produce antibodies that are orders of magnitude more potent in opsonizing the toxins than those produced by the traditional toxoids (Jones, Liu, Rigsby, & Sesardic, 2008).

Vaccine: A Solution to the Drug Resistance Problem

While the mutations by the microbes resulting in drug resistance is a growing concern for the pharmaceutical industry, the corresponding resistance to vaccines is rare, almost unheard of. There are two key reasons for the resilience of the vaccines. Firstly, the vaccines are used prophylactically. The pathogen is eliminated or neutralized by the host immune system before it can proliferate inside a vaccinee. Since most mutations occur during the DNA replication, the microbe gets less chance to develop mutations that can lead to vaccine resistance (Lipsitch & Siber, 2016). Further, the herd immunity acquired through the widespread immunizations keeps the pathogen population from achieving the critical numbers needed to accumulate diversity and spread through the communal infection. Smallpox was eliminated through this strategy (Kennedy & Read, 2017).

Secondly, the vaccines deploy the host immune response against multiple epitopes while the drugs tend to target a limited number of chemical pathways. The immune system, by its design, is adaptive and can respond to the changing constitution of

the pathogen. A drug on the other hand attacks a predefined biochemical (Kennedy & Read, 2017). For instance, all β-lactam antibiotics including penicillin and methicillin disrupt the synthesis of peptidoglycan by inhibiting the penicillin-binding protein (PBP). Peptidoglycan, a mesh-like polymer surrounding the cell wall is necessary for the structural strength of the bacterium. The weakened mesh causes the microbe to lyse and get scavenged. *S. aureus* became resistant to the β-lactam antibiotics by evolving to express PBP2a, an isoform of PBP (Stapleton & Taylor, 2002).

Vaccine Mediated Protection

Vaccines offer protection against pathogens through multiple and diverse avenues. They do so indirectly by mobilizing the adaptive and innate immune responses. The activated immune system lymphocytes control the proliferation of the microbes, neutralize their toxic secretions or both. The process starts primarily with the creation of pathogen and toxin specific antibodies. The antibodies are formed either from the de novo B cells or from the clonal expansion of the memory B cells. The other effector cells involved are $CD8^+$ and $CD4^+$ T cells. The former, commonly referred as cytotoxic T lymphocytes (CTL), induce apoptosis in the infected cells and bacteria by secreting granzymes and perforins in the extracellular matrix surrounding the pathogens. In a complementary role, the $CD4^+$ T-helper (Th) lymphocytes assist and control the antibodies and T-cell response through their secreted cytokines (Barinov, et al., 2017). There are numerous subgroups of Th cells, each with a unique role in the immune system. T-helper 1 (Th1) cells are the effectors against intracellular bacteria and other monocellular organism while Th2 cells help in removing the extracellular pathogens including bacteria and helminths. Th17 is especially significant in the context of this

research because it defends against the extracellular bacteria and fungi that colonize the epithelia of the skin and mucosa, e.g. *Streptococcus aureus*, *Pseudomonas aeruginosa*, *Mycobacterium tuberculosis* and *Bordetella pertussis*, The Th17 effector cells produce interleukin-17 (IL-17), IL-22 and IL-26 in response to the mucosal inflammation. In addition, the B cells activate the C3a complement cascade of the innate immune system (Guglani & Khader, 2010).

Pathophysiology of *Staphylococcus aureus*

S. aureus and penicillin, the first major antibiotic, were discovered in 1884 and 1941, respectively. The penicillin molecule has a β-lactam ring in its structure. The ring binds to the penicillin binding proteins in the enzymes that catalyze the cross linking of peptidoglycans in a maturing bacterial cell wall. That creates holes in the walls allowing the water to seep in and lyse the cell. Staph developed resistance to penicillin within a decade of its introduction by mutating to encode penicillinase. The enzyme cleaves the penicillin β-lactam ring. Methicillin (1961) containing an additional acyl group in the β-lactam ring has a higher resistance to penicillinase.

The MRSA variants of Staph appeared within 25-30 years. The resistance to methicillin occurred as a result of horizontal transfer of SCC*mec* plasmid (Foster, 2004). A site-specific recombination allows five different variants of SCC*mec* to integrate into the same site of the *S. aureus* chromosome (Figure 1). The *mecA* gene encodes PBP2a, a novel β-lactam–insensitive penicillin binding protein. PBP2a made the methicillin class of antibiotics ineffective against *S. aureus*. PBP2a favors its intended substrate, i.e., peptidoglycan, over the β-lactam antibiotic ring (Mahasenan, et al., 2017).

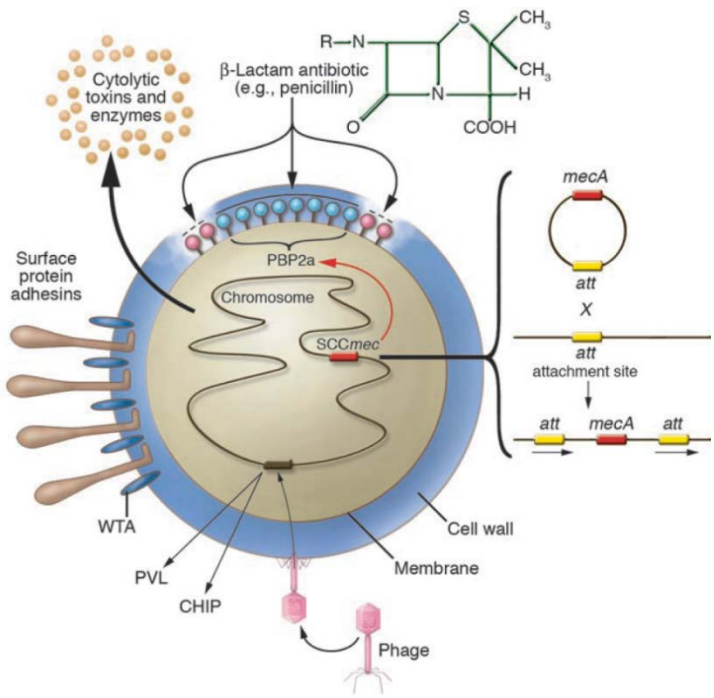

Figure 1. The mechanism of *S. aureus* acquiring resistance to methicillin
From "The Staphylococcus aureus "superbug"", by Foster, 2004
The Journal of Clinical Investigation © 2004

Currently, no antibiotic exists that is broadly effective against the various strains of Staph. *mecA* is a highly mobile motif that belongs to the staphylococcal cassette chromosome (SCC). Of the four major isoforms of SCC, Type 1, 2 and 3 occur primarily in the healthcare-acquired MRSA (HA-MRSA) and Type 4 is in the community-acquired MRSA (CA-MRSA). The latter is a less pernicious variant because it responds to non-methicillin antibiotics and it is less likely to lead to serious pathologies such as bacteremia and sepsis (Patel, et al., 2008).

The surgical site infections (SSIs) are a major part of the nosocomial MRSA morbidity that lead to prolonged hospitalization and mortality. Staph colonizes the surgical implants and intubations such as catheters and various types of feeding tubes. The bacteria anchor themselves to the metallic surfaces by forming biofilm via van der

Waals forces and the stronger electrostatic interactions (Tong, Davis, Eichenberger, Holland, & Fowler, 2015).

Nearly all the CA-MRSA associated skin and soft tissue infections (SSTIs) in the United States (> 97%) are caused by the USA300 strain. It is highly virulent and easily transmissible. The latter factor has significantly increased the overall disease burden due to SSTIs among the regions with high population densities such as the urban areas. Further, in addition to methicillin, USA300 is resistant to fluoroquinolone. Panton-Valentine leukocidin (PVL) toxin genes (*lukS-PV* and *lukF-PV*), the arginine catabolic mobile element (ACME), and SCC*mec* type IV are almost always present in these strains (Alam, et al., 2015). Interestingly, USA300 has failed to establish a foothold in any geographic location other the United States. The pattern of localized spread of *S. aureus* seems to be a common occurrence that has happened often historically (Planet, 2017). Fifty-five strains of USA300 were used in this research including the USA300 TCH1516.

Open Reading Frames

A reading frame is a continuous stretch of genomic DNA or RNA nucleotide sequence that is transcribed into a protein. All organisms other than viruses have DNA genome. Amino acids are transcribed from a triplet of nucleotides called codons. Since the codons use a set of three nucleotides, there are three possible reading frames in the sense (i.e., 5'-to-3') and antisense (i.e., 3'-to-5') directions. Therefore, there are six candidate reading frames for each nucleotide sequence. An open reading frame (ORF) is a segment of reading frame that can be transcribed into a protein. It begins with a start codon and ends with a stop codon. The start and stop codons vary by organisms. In *S. aureus* and humans, ATG is start codon while the stop codon can be TAA, TAG or TGA.

Since an open reading frame represents a sequence of nucleotides that can possibly yield a protein, its length is always in the multiples of three.

Reverse Vaccinology

Traditionally, the vaccines have been developed by the empirical process of isolate, inactivate and inject. In this method, colloquially referred as the Pasteur's Principle, first the causative agent is isolated. A strain of *S. aureus* that is resistant to methicillin and other antibiotics would be the causative agent in the context of this research. In the second step, the agent is inactivated by killing, attenuating or isolating the antigenic subunit. The host develops protective immunity against the antigen when the inactive agent is injected (Baxter, 2007), (Bragazzi, et al., 2018).

In contrast, reverse vaccinology is the science of computational tool to mine genomic data to select antigenic candidates and design vaccines. It blends computational biology with the conventional lab-based assays. It uses patterns in the genomic sequences as a proxy for the expressed proteins. Therefore, the proteins are computed *in silico* in lieu of isolating the same *in vitro* or *in vivo*. The translated proteins are scanned for traits such as cell-surface or extracellular localization, signal peptides and epitopes. The filtered proteins are narrowed down to the ones most likely to elicit an immune response from the host. Eventually, a small subset of the candidate proteins is cloned in the lab and tested on animal models (Rappuoli, Bottomley, D'Oro, Finco, & Gregorio, 2016).

The method was used for creating the first successful vaccine against the serogroup B meningococcus (MenB). The vaccine is marketed by Novartis under the brand name Bexsero. The four distinct antigens forming the quadrivalent vaccine were identified through reverse vaccinology. The antigens are Neisserial adhesin A (NadA),

Neisserial heparin-binding antigen (NHBA), factor H binding protein (fHbp) and PorA P1.4 immunodominant antigen of OMV NZ (strain NZ98/254) (Vernikos & Medini, 2014).

fHbp is also being used in Trumenba, the MenB vaccine developed and marketed by Pfizer. The vaccine is composed of two recombinant lipidated fHbp variants from *N. meningitidis* serogroup B. One of those antigens is from fHbp subfamily A (A05) and the other from subfamily B (B01) (Ostergaard, et al., 2017).

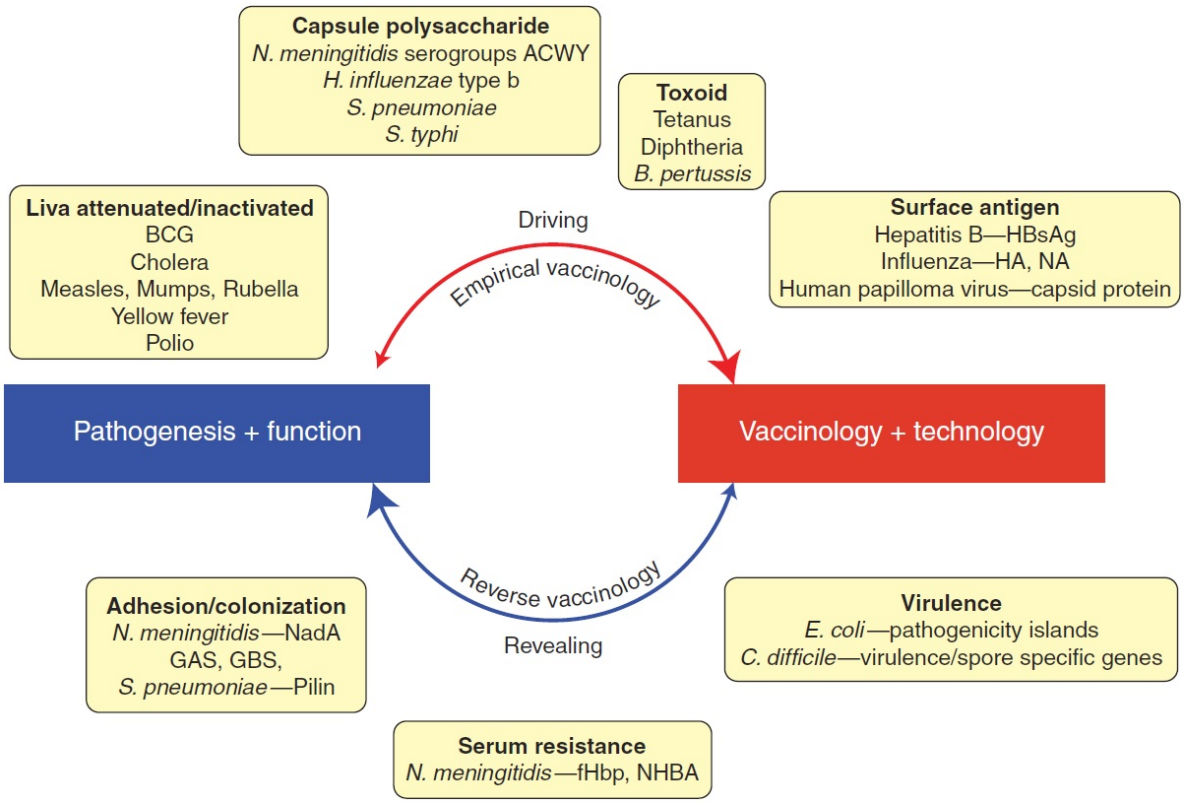

Figure 2 Empirical and reverse vaccinology.
From "Vaccines, Reverse Vaccinology, and Bacterial Pathogenesis" by Delaney et al., 2013, *Cold Spring Harbor Perspectives in Medicine* © 2013, Cold Spring Harbor Laboratory Press.

The potential applications of reverse vaccinology in the drug development are immense, even though it is an emerging technology. In addition to finding potential

vaccine targets, it has increased our understanding of virulence factors, mechanisms of infection, disease propagation, conservation, pathogenesis, etc. (Figure 2). For instance, it has furthered the knowledge of conserved virulence associated surface proteins in *S. aureus* including IsdA, IsdB, SdrD and SdrE (Delany, Rappuoli, & Seib, 2013). In other cases, new genes have been discovered through similar genome mining methods (Choksi, Babu, Lau, & Yu, 2014).

Integral Transmembrane Proteins

Transmembrane proteins are lodged into the cell membrane and are an integral part of it. These proteins have three distinct segments – cytosolic, exoplasmic and embedded. The first two are have hydrophilic moieties that interact with the aqueous solutions of the cytoplasm and the intercellular milieu. In contrast, the embedded membrane spanning segments have hydrophobic amino acids (Figure 3). The sidechains of these amino acids protrude outwards and interact with the hydrophobic core of the phospholipid bilayer. The sidechains wedged between the two leaves of the cell membrane anchor the transmembrane proteins firmly into the membrane (Cossart & Jonquieres, 2000).

These specialized proteins make unique contributions to the bacteria. Some act as gateways for the transport of molecules across the peptidoglycan cell wall and the phospholipid cell membrane while others are used for metagenomic colony forming. The exoplasmic surfaces of the transmembrane proteins in pathogenic bacteria can be antigenic targets since those are visible to the host immunologic cells.

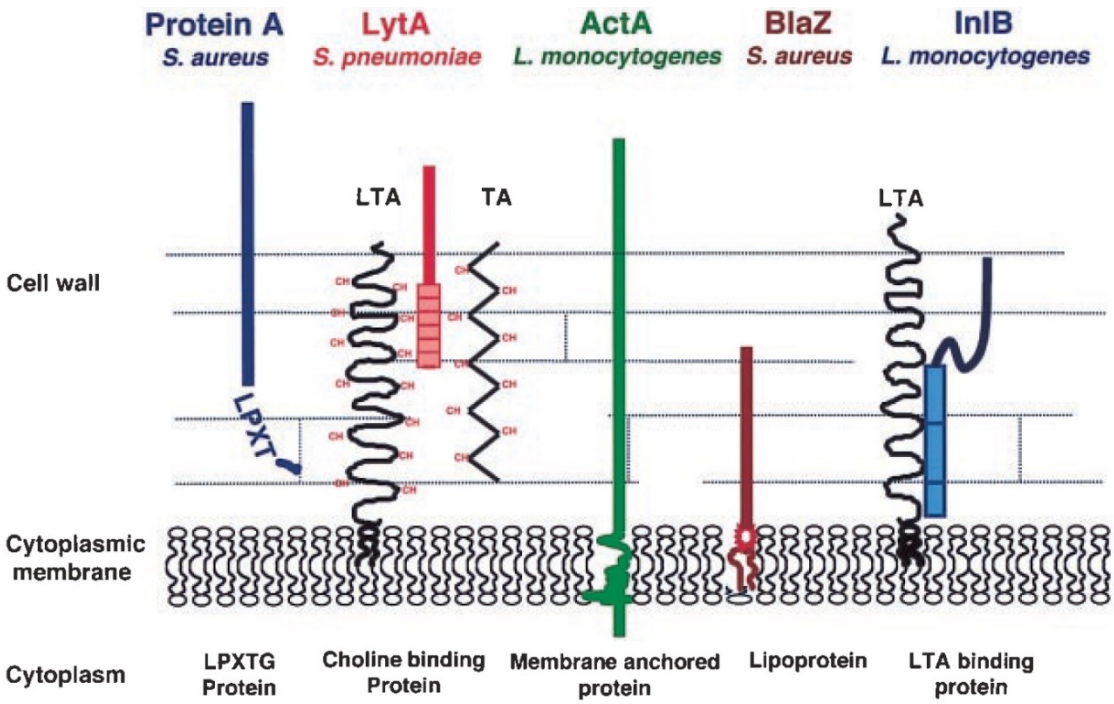

Figure 3. Major types of surface proteins of Gram-positive bacteria.
From "Sortase, a universal target for therapeutic agents against Gram-positive bacteria?" by Cossart & Jonquieres, 2000, *PNAS*, © 2000 Proceedings of the National Academy of Sciences of the United States of America.

Due to their special structure, the ribosomal assembly and posttranslational processing of integral transmembrane proteins differ from other proteins that are soluble in the cytosol. The transmembrane proteins follow a secretory pathway in prokaryotes that is quite similar to the one in eukaryotic cells. The nascent protein is tagged with a signal peptide sequence composed of 15 to 20 hydrophobic amino acids while it is being translated by a ribosome. In one of the pathways, the signal recognition particle (SRP) binds to the signal peptide, arrests the translation and translocates the ribosome to the cell membrane. The translation resumes after the SRP-ribosome complex docks onto the membrane receptor. The protein is inserted into the membrane as it being assembled. Alternately, the mature protein is secreted (Natale, Brüser, & Driessen, 2007).

Further, the membrane-spanning segments are alpha helices comprising hydrophobic amino acids. In contrast, the N and C termini are hydrophilic. Most often they end on the opposite sides of cell membranes. These identifiable patterns coupled with the signal sequence makes the genes encoding the transmembrane proteins unique. Therefore, it is possible to identify known or putative genes from the genome using the reverse vaccinology principles.

Surface antigens of *Staphylococcus aureus*

S. aureus secretes a broad range of adhesive surface proteins that facilitate the anchoring of the bacterium to the host cells and the extracellular matrix. These secretions are collectively known as microbial surface components recognizing adhesive matrix (MSCRAMMs). The group includes clumping factors (Clf) A and B, fibronectin-binding protein (FnBP) A and B, collagen adhesin A (CnA) and coagulase. These proteins are present during Staph infection and are highly immunogenic (Holtfreter, Kolata, & Broker, 2010).

The upregulation of the *mecA* and *fnbA* genes and the resulting expression of fibronectin-binding protein A (FnBPA) enables the *S. aureus* to bind to fibronectin. Fibronectin is a component of the extracellular matrix. The host tissue coats the foreign inorganic material such as the surfaces of the implants with fibronectin in an effort to isolate itself from the material. The opportunistic Staph uses FnBPA in the coating to anchor itself onto the implant and establish a colony. In addition, the extracellular polymers secreted by the colony provides it a protective layer and additional antibiotics resistance. Therefore, sometimes an invasive surgery to replace the infected implant is the only option (Côrtes, Beltrame, Ramundo, Ferreira, & Sá Figueiredo, 2015).

Low Success Rate in Developing a Vaccine against *Staphylococcus aureus*

Currently there are no commercially available vaccines against S. aureus despite the research efforts for half a century. The challenges to developing an effective vaccine against *S. aureus* are numerous and diverse. They include the bacterium's ability to subvert the human immune response learnt during its commensal residence on the skin, lack of a reliable test model, varied clinical presentations and immense geographical diversity (Lee, 2013). The toxicological tests for any vaccine under development require that the commensal bacteria must be spared while destroying the pathogenic ones. Finding the epitopes that can discern between the two populations is non-trivial.

Differentiating between the coagulase-positive and coagulase-negative specimens is yet another predicament. However, *S. aureus* has characteristic 81-bp tandem repeats in the 3'-end coding region of the coagulase (*coa*) gene which are often used for identifying the organism in the epidemiological samples (Dallal, Khoramizadeh, Amiri, Yaraghi, & Fard, 2016).

For the above stated and other reasons, almost all attempts at developing a vaccine have failed, even the well-funded ones. The examples include StaphVax and V710 created by GlaxoSmithKline and Merck, respectively (Lee, 2013). Both vaccines failed in the late stages of Phase III testing. However, V710 showed promise for some subgroups in the post hoc analysis including a higher efficacy against the methicillin-sensitive *Staphylococcus aureus* (MSSA) (infection rate, 19%; 95% CI, 11%-28%) than MRSA (infection rate, 34%; 95% CI, 23%-46%). It was also more effective in controlling the superficial SSIs than the deep-seated ones and in the baseline nasal carriers than the non-carriers (Fowler, et al., 2013).

Even non-vaccine therapies have shown a limited success. The trials for tefibazumab (Aurexis), the monoclonal antibody targeting ClfA in the patients with *S. aureus* bacteremia, were halted in 2006 after it failed to show any significant efficacy in the Phase II trials (Fowler & Proctor, 2014).

Status of the *Staphylococcus aureus* Vaccine Development

Pfizer is developing a quadrivalent vaccine SA4Ag and its trivalent precursor SA3Ag. These two vaccines are the most promising in the industry's pipeline. SA4Ag has shown efficacy by rapidly inducing elevated levels of anti-bacterial antibodies in the Phase I/II trials (Begier, et al., 2017). It comprises the CP5, CP8, ClfA and MntC antigens that are a part of the bacterial extracellular surface. In comparison, the bivalent StaphVAX was formulated with CP5 and CP8 and V710 utilized the *S. aureus* 0657nI iron surface determinate B (IsdB), which is another surface antigen (Giersing, Dastgheybb, Modjarradc, & Moorthy, 2016).

The capsular polysaccharides (CP) are the traditional antigens for the gram-positive bacteria such as *S. aureus.* They are hetero-trimers comprising one mannose and two fucose molecules. CPs are included in the vaccines because they are a characteristic molecule of the bacterial capsule and are mildly antigenic. To improve the host immune response to a vaccine, the CPs are paired with antigenic surface proteins. StaphVAX was formulated with capsular polysaccharides serotypes 5 (CP5) and 8 (CP8) conjugated with recombinant pseudomonal exoprotein A, which is a career protein (Bagnolia, et al., 2015). In comparison SA4Ag uses CP5 and CP8. The remaining two components of the SA4Ag vaccine, namely the protein clumping factor A (ClfA) and manganese transporter protein C (MntC), target the coagulation pathways. MntC was added to SA3Ag to

formulate SA4Ag (Frenck, et al., 2017). MntC is crucial for binding the bacteria to the extracellular matrix and laminar plasminogen and initiating the coagulation cascade. ClfA assists the bacteria in adhering to the host fibrinogen. (Salazar, et al., 2014). Therefore, SA4Ag does not directly target a species, *coa* gene or the coagulase protein. Instead it disrupts the coagulation pathway, thus preventing a bacterial colony expansion initiated by the quorum sensing (McAdow, Missiakas, & Schneewind, 2012). SA4Ag has successfully cleared the Phase I/IIb testing and was granted the Fast Track designation by the U.S. Food and Drug Administration in February 2014 (Giersing, Dastgheybb, Modjarradc, & Moorthy, 2016).

In designing SA4Ag, Pfizer has followed the traditional strategy in vaccinology of targeting the surface antigens. The adaptive immune system produces B-cell antibodies against such antigens. The opsonized alloimmune targets are then phagocytosed by mononuclear leukocytes such as the natural killer cells or killed by the apoptosis induced by cytotoxic T cells. T cells also assist in clonal expansion and affinity maturation of the B cells during the adaptive immune response (Bröker, Mrochen, & Péton, 2016).

Finding a vaccine for *S. aureus* is an unmet pharmaceutical need. The purpose of this research was to use reverse vaccinology to identify proteins that could be potentially antigenic. The strategy was to mine the genome of *S. aureus* to find novel ORFs that could possibly be genes. Thereafter, translate the ORFs and evaluate the probability that the translated proteins could relocate to the bacterium's surface after expression and maturation. The proteins thus identified would be further validated *in vitro*.

Chapter II

Materials and Methods

The details of the data collection and the experiment have been discussed in this section. Genomic data was downloaded from public repositories. It was curated for data integrity and then processed with custom written computer code and freely available bioinformatic tools. Starting with the dataset comprising 14,322 whole genome sequences, a list of 330 candidate proteins was distilled from the data.

Selection of Databases for the Genomic Data

The data collection started with a thorough survey of the nucleotide and protein databases worldwide. A large set of data was downloaded for *S. aureus* from each site and analyzed for completeness and suitability of use. The data was deemed complete and suitable if there was adequate diversity in the bacteria sequenced, both temporal and spatial. That means the data in the repositories should have been collected over a period, from different regions of the world and representing a diverse range of strains. Initially, an explicit size was not set for the usable sequence to further avoid any bias in the core data. Therefore, the genome and proteome sequences of various sizes and types were evaluated. Some of sequences analyzed were cDNA expressed sequence tags (EST), multi-locus sequence typing (MLST) and whole genome sequence (WGS). Both EST and MLST sequences represent the proteins and hence the coding segment of genes. ESTs are generated from the cloned complementary DNA (cDNA) while the MLSTs are the internal fragments of the housekeeping genes. Both types of sequences are a few hundred

base pairs in length. In comparison, the lengths of WGSs for *S. aureus* range between approximately 2,600,000 and 3,200,000 nucleotides.

Over 10,000 test sequences from various databases were analyzed for the aforementioned criteria such as diversity and completeness of data. ESTs and MLSTs were found lacking because they are sub-sequence segments of genes. They do not retain the location and the context of the sequences in relation to the whole genome. Databases such as MLST.ORG, UniProt and The Sanger Institute did not have adequate data for *S. aureus*. Therefore, WGSs from International Nucleotide Sequence Database (INSD) were chosen as the primary source of data. INSD comprises the following regional datacenters.

- DNA Data Bank of Japan (National Institute of Genetics)
- EMBL (European Bioinformatics Institute)
- GenBank (United States National Center for Biotechnology Information)

Software Modules Used for the Processing and Analyzing the Data

The project was done entirely *in silico*. Computer code was written for every aspect of the research, ranging from data acquisition and processing to genome mining and analysis. A short informative description of various third-party software modules and their usage is provided in this Section. The actual application of the modules is enmeshed in the 10,000+ lines of code written for the project. Therefore, the specifics of how a tool was used has not been discussed. For instance, NumPy was used wherever array processing or Fourier Transform was needed, which was at multiple instances during analysis.

The code was developed in the C/C++ and Python/BioPython programming languages. Python and BioPython were augmented with the following extensions. Again, only a few brief introductory statements about the extensions have been listed here.

- NumPy: a portmanteau of Numerical and Python, NumPy is the fundamental library for mathematical operations in Python.
- SciPy: built on NumPy, SciPy (stands for Scientific Python) contains essential modules for scientific computing including calculus, statistics and linear algebra.
- SciKit-Learn: based on NumPy and SciPy, the package provides algorithms for data mining and machine learning tasks such as model selection, regression and dimensionality reduction.
- StatModels: used for computing inferential statistics, e.g. confidence intervals and ANOVA
- MatPlotLib: a low-level API for plotting and visualizing data
- Seaborn: a higher-level library based on MatPlotLib containing a rich gallery of visualization tools for complex datasets.

Further, statistical analysis was done with R. While there are several Python interfaces to R such as rpy2 and rPython, working directly with the R programming language was the most efficient.

The computer code developed for this project was supplemented with the public domain bioinformatics tools. The tools were valuable in testing the proofs of concept and checking intermediate results. Some of the major tools used in this experiment are as follows.

BLAST (Basic Local Alignment Search Tool): is a suite of algorithms that compares a biological sequence, nucleotide or peptide, with another biological sequence. It also searches various National Center for Biotechnology Information databases for biological sequences that closely match a queried sequence. Both functionalities of the BLAST toolset were used extensively in this research to qualitatively and quantitatively evaluate the difference between isoforms and strains and to find close matches for the proteins of interest.

ExPASy (Expert Protein Analysis System): is another large suite of tools with applications in proteomics, genomics, transcriptomics, etc. Mostly the Translate Tool of

ExPASy was used here because in addition to translating nucleotides into proteins on all six reading frames, the tool highlights the putative open reading frames.

Clustal/Ω, Kalign, MAFFT (Multiple Alignment using Fast Fourier Transform), MUSCLE (MUltiple Sequence Comparison by Log-Expectation) and T-Coffee (Tree-based Consistency Objective Function for Alignment Evaluation): are the multiple sequence alignment programs for aligning three or more sequences in a biologically relevant manner. Clustal also produces cladograms and phylograms which show the phylogeny of a group of sequence in a tree format. The evolutionary phylogenetic distances can be then computed from the edge length in the tree. These programs were used extensively in this research to study the relationship among strains and their motifs.

EMBOSS (European Molecular Biology Open Software Suite) Needle: uses the Needleman–Wunsch algorithm for creating global alignment of biological sequences. The tool was used as an ancillary to the BLAST alignment modules.

The materials and methods had to be adapted for some of the third-party tools while working in different areas of the project. For example, Clustal/Ω is available in both command line script and web-enabled graphical user interface (GUI). In general, the scripts are complex to set up and run. They were written when Unix was the dominant platform for computing. However, they have fewer limitations. For instance, the GUI version of Clustal/Ω can accept only 4 million amino acids. It was not feasible to use the web version of Clustal/Ω for *S. aureus* which has a genome size of ~2,800,000 that translates into ~900,000 amino acids. Therefore, the GUI was used for analyzing the smaller proteins such as the genes and the command line was used for the larger datasets.

The computer code was developed with gcc 8.2 under Oracle Linux 7.1 and Visual Studio 2017 C/C++, Python 3.7.2 and BioPython 1.7.3 under Windows 10.

Collection of Genomic Data

The data was downloaded from INSD both interactively and programmatically. In the former process, queries and downloads were done manually on the websites while in the latter method, computer code was used to access the ftp servers for queries and downloads.

An unrestricted query for *Staphylococcus aureus* yielded 1,818,834 nucleotide and 23,694,580 protein samples. The sequence lengths for nucleotides varied from 14 bp (contigs created during shotgun sequencing) to 3,367,972 (WGS). Correspondingly, the proteins lengths varied between 20 aa (phenol-soluble modulin PSM-alpha-4) and 10,545 aa (hyperosmolarity resistance protein Ebh).

Noticeably, there were no whole genome protein sequences in the database. That can be explained by the fact that unlike the eukaryote DNA, the bacterial DNA is being modified constantly by the horizontal gene transfer and incorporation of plasmids. Therefore, there is no 'standard' genome for *S. aureus* or any other bacterium. By corollary, the nucleotide whole genome sequences too varied in length. The typical length was approximately 2,800,000 bp.

The experiment was conducted with the 14,322 whole genome sequences.

Table 1. Stats of the Whole Genome Sequences

Statistics	Length (in base pairs)
Mean	2,832,496
Median	2,829,071
Range	2,563,627 to 3,095,697
Quartiles	2,773,461; 2,829,071; 2,878,610

The length of WGSs used in the experiment varied. The whole genome sequences are assembled by shotgun sequencing. The sequencing starts with the collection of multiple copies of the genome under study. The chromosomes are then broken into random small sequences, amplified through polymerase chain reactions and reassembled into the whole genome sequence using computer algorithms. Artifacts can be introduced at each stage of this process. It is especially so in the organisms with large chromosomes such as *S. aureus*. That results in the generation of sequences with varying lengths, as was the case in the WGSs collected for this research. The lengths varied between 2,563,627 and 3,095,697 with a median value of 2,829,071 (Table 1).

Finding Open Reading Frames

The varying lengths of WGS illustrated that there is no single representative sequence of *S. aureus*. Instead, there are a large number of sequences with included artifacts. Therefore, the remaining research focused on the open reading frames (ORFs) and genes embedded in the whole genome sequences. In the absence of a standard chromosome for the bacterium, narrowing the investigation to ORF and gene regions of the genome was the logical choice for studying the potentially antigenic proteins.

The WGSs were parsed for the ORFs using algorithms that looked for start and stop codons in the sequences since an ORF is bookended by those two codons. The candidate ORFs smaller than 75 nucleotides (= proteins with 25 amino acids) were not processed because that could be noise in the data. The computer code was augmented with bioinformatics tools including ExPASy, NCBI ORFfinder and Sequence Manipulation Suite (Stothard, 2000). The bioinformatics tools were used for verifying the results of the computer code and to discover additional ORFs.

The ORF mining was computational very intensive. It required multiple networked computers working in tandem round-the-clock for days. A set of 6,538 ORFs were extracted from the WGS using biostatistical algorithms developed for this project. The biostatistical algorithms, as the name implies, used statistical analysis while accounting for the microbiology of the translated proteins. The ORFs were analyzed from multiple facets such as the length, location in the sequence and matching proteins. The matching proteins were discovered using BLASTP. The nucleotide sequences in the ORFs were first translated into proteins using BioPython. The resulting proteins were queried against the non-redundant protein sequences (nr) database of the GenBank using BLASTP.

Gene Extraction

Next de novo genes were predicted from the WGS using bioinformatics tools, which included Artemis (Carver, Harris, Berriman, Parkhill, & McQuillan, 2012), EuGene (Sallet, Gouzy, & Schiex, 2014), GeneMark (Besemer & Borodovsky, 2005) and GLIMMER (Delcher, Bratke, Powers, & Salzberg, 2007). A total of 1,574 genes were consolidated from various WGS. The known genes of *S. aureus* were excluded from this list.

Corelating Open Reading Frames and Genes

The ORFs denote an area of the genome that could possibly encode a protein. In other words, it could be the location of a known or unknown gene. Therefore, the next logical step was to correlate the coordinates of computed ORFs and genes. An ORF was considered as a part of a gene if it occupied the same location as the gene. Conversely, an

ORF not sharing a location with a known gene could be an unknown gene. Lastly, it could also be a meaningless pattern of nucleotides that matches the algorithms of ORF detection.

Prediction of Surface Proteins

The previous steps helped find the known and unknown proteins that could be coded by the genome of *S. aureus*. In an intact bacterium, the surface antigens are the easiest targets for the host immune system. Therefore, the surface proteins too are ideal vaccine targets. During transcription, a short signal peptide is appended to the N-terminus of the newly synthesized protein that is targeted to the secretory pathway. Such proteins are either secreted or embedded on the cell surface after the cleaving of the signal peptide. In either case, they are possible vaccine targets. In prokaryotes, the signal peptide directs the protein towards the SecYEG protein-conducting channel present in the plasma membrane (Natale, Brüser, & Driessen, 2007).

The prediction of surface proteins was done with TMHMM bioinformatics tool (Krogh, Larsson, von Heijne, & Sonnhammer, 2001). The tool evaluates both the signal peptides and trans-membrane motifs. Proteins translated from 1,045 ORFs were used for this segment of the study. The nucleotides of these ORFs demonstrated the patterns consistent with the presence of a gene. In order to seek hitherto unknown targets only the ORFs that did not overlap with the known genes were used at this stage. The proteins translated from the ORFs were evaluated for factors such as the probability of their N-terminus being on the cytosolic side, number of membrane-spanning domains and the likelihood of having a signal peptide. Finally, 330 protein sequences were chosen that could be possible antigens.

Chapter III

Results

The objective of this research was to identify candidate proteins that could be antigenic targets for the development of a vaccine for *S. aureus*. This *in silico* experiment used the whole genome sequences of the bacterium as the primary input into the model. The sequences were parsed for open reading frames and genes. The ORFs and genes were correlated to verify the integrity of the data. Finally, 330 protein sequences were identified that have the potential to relocate to the cell surface or be secreted, making them candidates for the host immune response and the components of a vaccine. The results at each of these steps dictated the design of downstream experiments.

Whole Genome Sequences

A large number of genomic segments such as the expressed sequence tags and multi-locus sequence typing, are available in the databases. However, they do not provide the continuity of whole genome sequences or the possibility of extracting ORFs, which was the cornerstone of this research. Interestingly, during the two years of this research, the number of whole genome sequences of *S. aureus* available in the databases increased from a few hundred to over 20,000. There are still more raw shotgun sequences in the pipeline waiting to be processed into whole genome sequences. 14,322 whole genome sequences for *S. aureus* were selected after checking for completeness, duplications and whether the headers were curated by the contributors of the sequences. A few sample WGS are listed in Table 2.

Table 2. A Sample List of *S. aureus* Whole Genome Sequences

Accession	Length (in base pairs)	Description
AM990992	2,872,582	Staphylococcus aureus subsp. aureus ST398 complete genome
AP009324	2,880,168	Staphylococcus aureus subsp. aureus Mu3 DNA, complete genome
AP009351	2,878,897	Staphylococcus aureus subsp. aureus str. Newman DNA, complete genome
AP014942	2,775,733	Staphylococcus aureus DNA, complete genome, strain: FDA209P
CP000730	2,968,760	Staphylococcus aureus subsp. aureus USA300_FPR3757, complete genome
CP000730	2,968,760	Staphylococcus aureus subsp. aureus USA300_TCH1516, complete genome
CP011685	2,848,095	Staphylococcus aureus strain ZJ5499, complete genome
CP017091	2,833,430	Staphylococcus aureus subsp. aureus strain ISU926, complete genome
CP018629	3,095,697	Staphylococcus aureus strain MRSA107 chromosome, complete genome
CP027788	3,044,721	Staphylococcus aureus strain CMRSA-6 chromosome, complete genome
CP029172	2,902,681	Staphylococcus aureus strain PTDrAP2 chromosome, complete genome
CP029629	2,808,798	Staphylococcus aureus strain MOK063 chromosome, complete genome
CP029685	2,958,212	Staphylococcus aureus strain CMRSA-3 chromosome, complete genome
FN433596	3,043,210	Staphylococcus aureus subsp. aureus TW20, complete genome
FR714927	2,729,540	Staphylococcus aureus subsp. aureus ECT-R 2 complete genome
HE579073	2,759,328	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 18583
LR134085	2,846,632	Staphylococcus aureus strain NCTC12233 genome assembly, chromosome: 1
LS483309	2,951,503	Staphylococcus aureus strain NCTC9944 genome assembly, chromosome: 1

Open Reading Frames and Translated Proteins

The ORFs were searched on all six frames because there was no way to ascertain the polarity of the sequences. The ORFs were then translated using the Translation Table 1 of the genetic code (Jukes & Osawa, 1993). The number of ORFs and therefore, the number of translated proteins varied with each WGS sample. The variation could have been due to the artifacts introduced during shotgun sequencing and the inclusion of plasmids in the genome that is common among bacteria.

Table 3. Distribution of Translated Proteins

Protein Length (in amino acids)	Number of ORFs			
	Mean	Median	Minimum	Maximum
< 25	Not computed			
25 - 50	17,836	17,870	16,763	19,052
51 - 75	3,136	3,125	2,964	3,415
76 - 100	823	830	733	928
101 - 125	386	390	340	432
126 - 150	253	250	223	298
151 - 175	198	198	182	217
176 - 200	148	148	136	172
201 - 225	159	157	145	177
226 - 250	162	163	150	176
251 - 275	171	171	156	183
276 - 300	139	141	130	145
301 - 325	146	146	139	155
326 - 350	129	129	121	143
351 - 375	108	109	98	113
376 - 400	82	83	73	105
401 - 425	99	99	94	102
426 - 450	73	72	67	90
451 - 475	75	75	70	79

476 - 500	54	54	49	59
501 - 750	23	22	9	53
751 - 1,000	19	5	1	9
1,001 - 1,250	2	2	0	5
1,251 -	1	0	0	8
Total	24,489	24,540	22,981	26,298

All ORFs smaller than 75 nucleotides (= 25 amino acids) were ignored because they were unlikely to yield any meaningful results (Table 3). The number of ORFs detected in each WGS varied. The count ranged between 22,981 and 26,298. The number seems large considering that there are ~2,500 known genes for *S. aureus*. Therefore, it was verified that the ORF algorithm was not simply treating an occurrence of ATG as the start of an ORF. An ATG in the DNA sequence does not necessarily flag a start codon, it can also code methionine. Therefore, the occurrences of ATG on all six reading frames were counted. The number of ATGs found in each WGS ranged from 82,376 to 132,487. The number of computed ORFs was substantially lower (22,981 to 26,298) than the ATGs in the sequences.

The length of ORFs and their translated proteins also varied. For example, the number of proteins found in various WGS strains that are 25 to 50 aa (amino acids) in length ranged between 16,763 and 19,052, with a median value of 17,870.

The variation in the protein lengths and the number of detected ORFs could be attributed to numerous reasons. The bacterial genome codes numerous proteins of different lengths. Also, artifacts could be introduced into the WGS when the sequences are assembled from the shotgun fragments. Since ORF detection algorithm looks for patterns in the protein sequences to estimate a putative ORF, it could be spoofed by these artifacts.

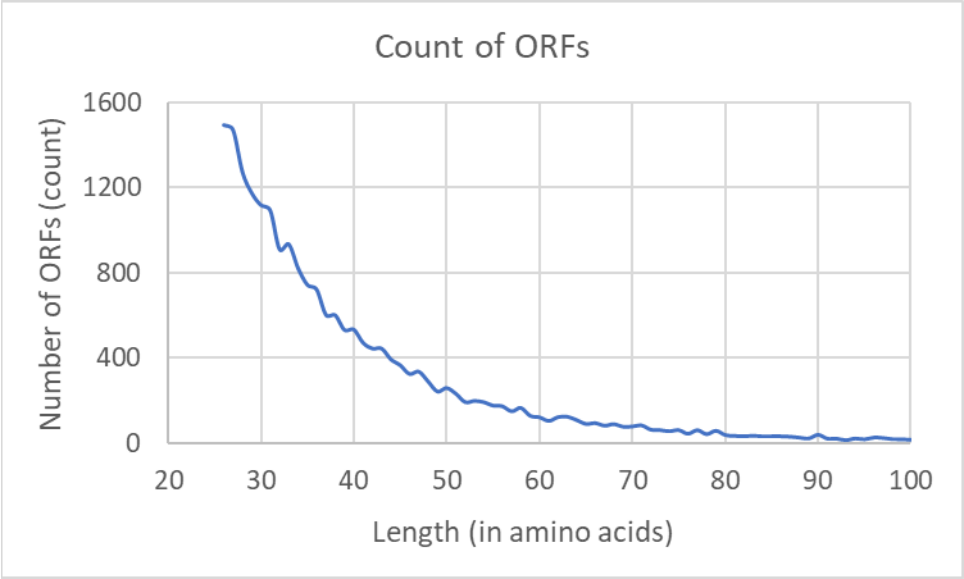

Figure 4. Count of Computed ORFs

Further, the translated proteins smaller than 50 aa were not processed. The reasons were two-fold. Firstly, the number of ORFs increased rapidly as the length of the ORFs reduced below 50 aa (Figure 4). The numerical processing required for each ORF was immense. Secondly, the smaller ORFs are more likely to be noise than the longer ones because the ORF detection algorithm is making a judgement on short runs of the DNA in those cases. Therefore, an arbitrary though reasonable, lower threshold was set for the ORF length at 50 amino acids. That reduced the number of ORFs used for further processing to 6,528. The number is congruous, considering that there are about 2,500 known genes in the *S. aureus* chromosome. Including 6,528 ORFs in the dataset could help identify unknown proteins beyond those coded by the known genes.

The translated proteins, and not the original ORF DNA sequences, were used for further analysis.

Table 4. A Sample List of the Proteins Translated from ORFs

Index	ORF Coordinates (in bp)		Length of Translated Protein (in aa)	Protein Matches from GenBank	
	Start	End		Number of Matches	Matched Protein
1	477,911	478,081	56	1	glutamate synthase [NADPH] large subunit [Staphylococcus aureus]
2	10,217	10,417	66	1	glycosyltransferase family 2 protein [Flavobacterium sp. LB2P30]
3	2,394,743	2,394,961	72	1	Uncharacterized protein [Klebsiella pneumoniae]
4	404,064	404,324	86	15	GlsB/YeaQ/YmgE family stress response membrane protein [Staphylococcus aureus]
5	851,509	851,841	110	16	pathogenicity island protein [Staphylococcus aureus]
6	1,139,236	1,139,583	115	8	unresolved
7	2,237,288	2,237,761	157	14	multidrug efflux transporter SepA [Staphylococcus aureus]
8	1,697,052	1,697,561	169	21	unresolved
9	1,877,865	1,878,398	177	22	enterotoxin [Staphylococcus aureus]
10	24,339	25,046	235	8	DNA-binding response regulator [Staphylococcus aureus]
11	2,584,367	2,583,204	387	0	no matches found
12	464,136	465,476	446	20	sodium-dependent transporter [Staphylococcus aureus]
13	2,659,725	2,658,379	448	28	unresolved
14	44,935	46,284	449	12	recombinase RecA [Staphylococcus aureus]
15	26,860	28,212	450	14	hypothetical protein [Staphylococcus aureus]
16	124,364	125,722	452	23	tetracycline efflux MFS transporter Tet (38) [Staphylococcus aureus]
17	2,472,362	2,473,801	479	27	transport system protein [Staphylococcus aureus]

Other reasons for translating the ORFs into proteins were to verify the authenticity of the ORFs and to investigate the putative products of the compute ORFs. The translated proteins were matched against the ones in GenBank nr database using BLASTP. A sample list of the results is in Table 4. The BLASTP queries returned zero or more matches. The returned matches were parsed for the protein names. A protein was considered a match with the queried protein sequence (i.e. the one translated from the ORF) if majority of the proteins returned by BLASTP (i.e. > 80%) were the same. Partial matches were accepted at this step if there was a consensus among matched proteins.

In some cases, numerous matches were returned but there was no clear consensus among the returned proteins, e.g. #6, #8 and #13 in the table. Those have been marked as "unresolved" in the table. Most often there was no discernable pattern among the returned proteins in those cases. Still in other queries, the returned proteins were conserved across multiple bacterial species, as in the item #13. Some queries were also flagged as unresolved for other reasons. For example, the GenBank IDs and descriptions of the eight matches to item #6 are listed below. Even though the matched proteins were all ABC transporter permease, conserved across multiple species, it was marked unresolved because there was no specific reference to *S. aureus*.

	GenBank ID	Description
1.	WP_113641387	ABC transporter permease [Prauserella sp. PE36]
2.	WP_110334801	MULTISPECIES: ABC transporter permease [Prauserella]
3.	WP_106181863	ABC transporter permease [Prauserella shujinwangii]
4.	WP_024876359	MULTISPECIES: ABC transporter permease [Saccharomonospora]
5.	WP_091802079	ABC transporter permease [Prauserella marina]
6.	WP_005461325	ABC transporter permease [Saccharomonospora glauca]
7.	WP_101591230	hypothetical protein [Bacillus sp. M6-12]
8.	WP_012795881	ABC transporter permease [Saccharomonospora viridis]

Some of the salient patterns observed in the results were,

- A large number of smaller proteins (< 100 aa in length) were unresolved (41.68% of the strains in this group), had no matches (25.93%) or matched with bacterial genera other than Streptococcus (17.62%). Only 14.77% of the smaller proteins matched conclusively with *S. aureus*.
- In contrast, most of the larger proteins (≥ 100 aa in length) had a positive match with the known proteins of *S. aureus* (85.94% of the strains in this group). While there were still unresolved (11.14%) and no-match (1.02%) cases in this group, the matches with species other than *S. aureus* were rare (1.89%).

Table 5. Protein Matches for Larger ORFs

Index	ORF Coordinates (in bp)		Length of Translated Proteins (in aa)	Protein Matches from GenBank	
	Start	End		ID	Description
1	262,946	264,640	574	WP_000975357	MULTISPECIES: glycosyltransferase family 2 protein [Staphylococcus]
2	335,972	338,044	690	WP_000943830	MULTISPECIES: YSIRK domain-containing triacylglycerol lipase Lip2/Geh [Staphylococcus]
3	360,377	361,606	409	WP_000186217	deferrochelatase/peroxidase EfeB [Staphylococcus aureus]
4	419,081	420,622	513	WP_117231992	glutamine-hydrolyzing GMP synthase [Staphylococcus aureus]
5	526,454	528,547	697	WP_114288247	ATP-dependent metallopeptidase FtsH/Yme1/Tma family protein [Staphylococcus aureus]
6	637,574	638,989	471	ABX28617	hypothetical protein USA300HOU_0591 [Staphylococcus aureus subsp. aureus USA300_TCH1516]
7	654,119	655,780	553	WP_065315517	arginine--tRNA ligase [Staphylococcus aureus]
8	674,591	676,633	680	WP_000402172	MULTISPECIES: sodium:proton antiporter [Staphylococcus]
9	772,115	773,896	593	WP_078104574	DNA helicase RecQ [Staphylococcus aureus]

The metrics of matching such as e-values, percent identities and gaps were not evaluated exhaustively for all ORFs because the objective at this stage was to assess the computed ORFs rather than identify any specific protein. However, the matches for numerous larger ORFs (> 450 aa in length) were investigated individually using BLASTP. In all cases, the proteins translated from the ORFs perfectly matched the ones in GenBank. A perfect match was defined as having 100% identities and zero gaps. A sample list of ORFs and their perfectly matching proteins is in Table 5.

The pattern in the smaller proteins in Table 4 can be attributed to the noise or less reliability of the smaller ORFs. However, the results for the larger proteins is remarkable in both Tables 4 and 5. To recall, the ORFs were detected using the bioinformatics algorithms on nucleotide sequences of DNA and then translated. There was neither a presumption of the species nor any bacterial/prokaryotic specific processing. Still the proteins, especially the larger ones, unequivocally matched with those of *S. aureus*. Considering that there are thousands of species in the databases, that is extraordinary.

Gene Prediction

ORFs are the nucleotide sequences delimited by start and stop codons. They are candidates for transcription that can eventually produce a protein. On the other hand, genes code for mRNA that are translated into proteins. While there were numerous *de novo* genes, a large number of predicted genes matched the known genes. The location of ORFs were investigated for overlaps with the known genes.

Table 6. Open Reading Frames Overlapping with Known Genes

Index	ORF Coordinates		Protein Length	Matching Protein for the ORF	Overlap Gene	Protein Coded by the Overlap Gene
	Start	End				
1	1,063,568	1,063,568	55	hypothetical protein [Staphylococcus aureus]	purD	phosphoribosylamine-glycine ligase
2	1,414,617	1,414,617	56	Dihydrodipicolinate reductase [Staphylococcus aureus]	dapA	dihydrodipicolinate synthase
3	2,430,638	2,431,069	143	DUF2871 domain-containing protein [Staphylococcus aureus]	trxB	pyridine nucleotide-disulfide oxidoreductase
4	2,759,381	2,760,073	230	polysaccharide biosynthesis tyrosine autokinase [Staphylococcus aureus]	cap1C	capsular polysaccharide biosynthesis protein Cap1C
5	2,162,658	2,163,386	242	F0F1 ATP synthase subunit A [Staphylococcus aureus]	atpF	F0F1 ATP synthase subunit B
6	1,141,387	1,141,387	271	succinate dehydrogenase iron-sulfur subunit [Staphylococcus aureus]	sdhA	succinate dehydrogenase flavoprotein subunit
7	1,121,670	1,122,548	292	heme ABC transporter substrate-binding protein IsdE [Staphylococcus aureus]	isdD	Heme transporter ABC transporter ATP-binding protein IsdD
8	38,644	38,644	322	beta-lactam sensor/signal transducer MecR1 [Staphylococcus aureus]	mecA	penicillin binding protein 2 prime
9	695,225	695,225	341	iron ABC transporter permease [Staphylococcus aureus]	fhuA	ferrichrome ABC transporter ATP-binding protein
10	894,890	894,890	465	Fe-S cluster assembly protein SufB [S aureus]	nifU	FeS cluster formation protein

The units of the coordinates in the above table are in base pairs of nucleotides and the lengths of the translated protein are in amino acids. The overlapping genes were found by comparing the locations of the ORFs and the genes. Two groups of proteins were compared in this step. The first set was obtained by BLASTP queries of the proteins translated for the computed ORFs. The second set comprised the products of the known genes at the same location as the ORFs. The objective was to find how well the proteins from the ORFs and the co-located genes match.

In a significant number of cases, there was no match between the proteins from the genes and the ORFs, as in Items #1 and #3 of Table 6. There were matching or closely matching proteins in other cases, e.g., Items #8, #9 and #10. However, in some cases such as Item #2 and #5, #6, there were inconsistencies in the descriptions between the proteins translated from the ORFs and those coded by the genes. These could be a result of the discrepancies in labeling the proteins in the databases.

The results showed that once again the length of the ORFs mattered. The proteins from the longer ORFs matched that of the genes better. That further bolstered the case for using longer ORFs to obtain more reliable results. By corollary, an ORF that does not share its location with a known gene has the potential to be an undiscovered gene. In addition, the gene prediction tools Artemis, EuGene, GeneMark and GLIMMER were used to find genes in the WGS.

Putative Surface Proteins

The proteins that are likely to be embedded into the cell surface must have anchoring motifs (Figure 3). In addition, they are prepended with the signal peptides. The

bioinformatics tools look for the patterns in the biological sequences that yield proteins with such properties.

TMHMM was the primary bioinformatics tool used for this step. The list of 6,528 proteins translated from the ORFs were evaluated for their probability of translocating to the cell surface after assembly. After eliminating the proteins that had no chance of being a surface protein, the remaining 1,045 candidate proteins were evaluated with TMHMM.

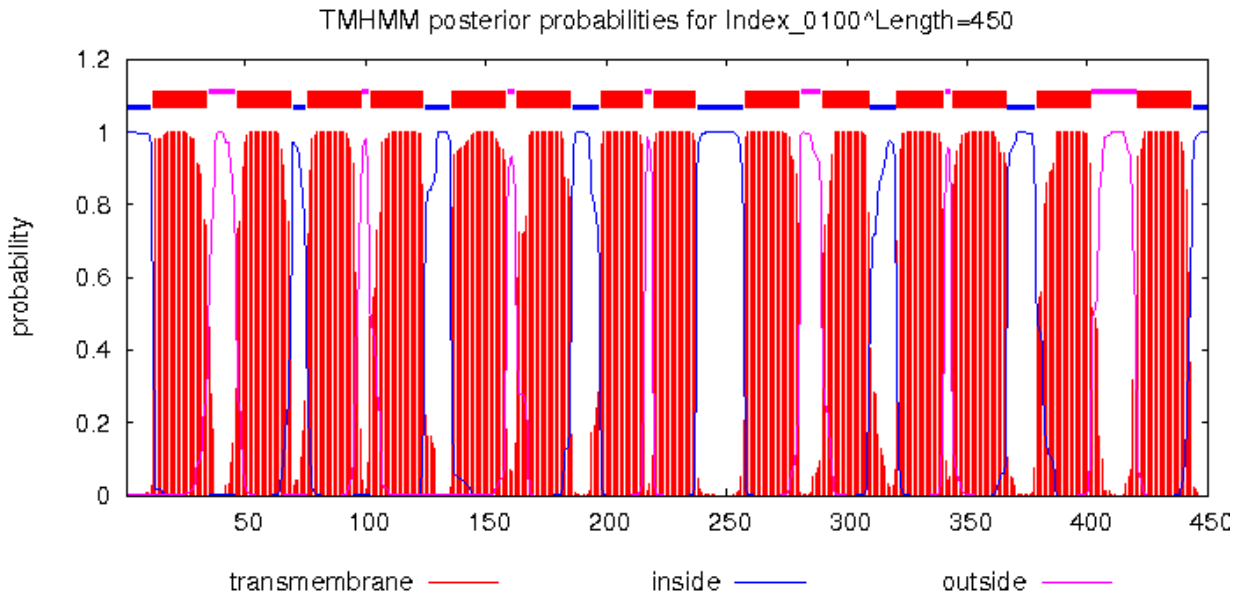

Figure 5. Sample output of TMHMM

A sample output of TMHMM is in Figure 5. It shows the probability that segments of a protein sequence can span membrane and hence anchor the protein. Based on the probability values, 330 protein sequences were chosen that could be antigenic targets and therefore, candidates for incorporating into a vaccine. A sample list of the proteins is in Table 7. The IDs in the list are the control numbers used in this research. The length is measured in the number of amino acids in the protein and The sequences denote the composition of the protein.

Table 7. A Sample List of Putative Antigenic Proteins

Protein		Sequence
ID	Length	
0087	135	MSSIIGKIAIWIGIVAQIYFSVVFVRMISINIAGGSDYETIFLLGLILALF TVLPTIFTAIYMESYSVIGGALFIVYAIIALCLYNFLSSILWLIGGILLIW NKYSKDESTDENEKVDIESTENQFESKDKITKE
0141	119	MNTIDTHTKEQQFSNLVRSYRKEYVGKGPNSIRVSFKDNWAIAHMT GVLSKVESFYLNDKRNESMLHYTRTEKIKQMYKEIDVNEMESLVGA KFVKLFTDIDLNDDEVISIFVFDKSIE
0186	114	MTTQMKIKTYLVAGIKAALLDTTGIKLASKSETTSHTYQHQALVDQ LHELIANTDLNKLSYLNLDAFQKRDILAAHYIAKSAIRTKNLDQMTK AKQRLESIYNSISNPLHSQNN
0370	135	MADITVVNDTGELYNVINQKKSEGYLESELTIISKSKLHLNDLHDSEI SLISTSGTFSDRMTKLLTGEDGEHAVLSRYNLAPDELEKYKQLILDD KMLVVAVRDKSSHQEVHENNSAYEEIDITHFAEASKGPKA
0904	121	MTTTPYDIIGKEALYDMIDYFYTLVEKDERLNHLFPGDFAETSRKQK QFLTQFLGGPNIYTEEHGHPMLRKRHIDFTITEFERDAWLENMQTAI NRAAFPQGVGDYLFERLRLTANHMVNS
1350	110	MGKKMGLGLSIALVVIGIAVVCLMIFSSQKTTYFGYMNSNTNAEKV VSEKDGLVKHNIKVEPSNDFKPKKGDFVKLVSKDDGKTFYKQEIVK HDDVPHGLMMKIHDMHMN
1771	104	MEQKKLSEMSEPELRHEIQLYKEKMRKAEMNGILNEYDVYQSKVIV AESYLVDRKKIENGKIYKLTDGSNQYFKVERLKGIFAWGFRFNSDEP EEGLPIALLQL
2299	178	MTKTMIRLASEQDAEKLQQLMHEAFTPLRELGIDWPSVNANLDAVK ENIDKNTTFVMTIDDEIISTITVRYPWGSVKSISGYPFVWWFATNPEY EGKGYGSQLLTYVEEAFLRDTLKSAAVTLGTSARLHPWLLKIYEKR GYEIYSEHENDDGDLGVIMRKVLIPERFDETILGQPPF
2375	140	MNKKHVFIIIGVILCICIVASVIYLKVKYDEKEKQKAIYYKEQQERITL YLKHNTKEPNTIKTVHFTSLKRGPMGDAVIEGYINENKEDDFVAYGS PEHNYQFGGSLIKSKNLSTLLKPVHQTKSPDEIKKELESKKNDR
2505	138	MRLKVLFHFIAAIFISFMLLWMTMLFDLISNQSHLKALLLNLDFLIPS DNTPYILEIICHLLIGSVIYFVFVLLFHTSKRLYYLCYIPLFFLFIALYPF LVFIAQRPIFQFSVTELIGWIITHIFFMSLMALVIPRIK

Chapter IV

Discussion

Humans have a dichotomous relationship with *Staphylococcus aureus*. We harbor it as a commensal bacterium. In contrast, the virulent strains of Staph cause serious illnesses. The benefit of its existence on the epithelia of numerous organs is poorly understood. However, sometimes the seemingly benign bacteria turn pathogenic. Once again, the factors leading to the virulence are not well known but the initiation of coagulation cascades seem to precede the onset of the symptoms (Dallal, Khoramizadeh, Amiri, Yaraghi, & Fard, 2016). The methicillin-resistant *Staphylococcus aureus* (MRSA), especially the healthcare-acquired variant (HA-MRSA), is particularly virulent. Currently, there is no commercially available vaccines against *S. aureus* despite 50 years of well-funded efforts. SA4Ag, under development by Pfizer, could be the first viable vaccine (Begier, et al., 2017)

The purpose of this study was to find antigenic targets that can be used for designing an *S. aureus* vaccine. Unlike the conventional approach of "isolate, inactivate and inject" to develop a new vaccine (Bragazzi N. , et al., 2018), this project used reverse vaccinology. Reverse vaccinology is the science of using genomic data to identify antigenic proteins that could relocate to the surface of the pathogens upon maturation or be secreted. The surface antigens elicit antibody mediated host immune response that kill the pathogen while the secreted factors can be targets for the vaccines that neutralize the effects of the toxins on the host and disrupt the lifecycle of the pathogenic microbes.

Reverse vaccinology is a relative new field. There are no established practices for the technology. As in any drug development, it is a long arduous process. It was successful in creating the first viable vaccine for meningococcus serotype B (MenB). The researchers at Chiron Corporation (now Novartis Vaccines and Diagnostics) analyzed the genome sequence of the MC58 strain of MenB. Of all the surface-associated proteins identified through reverse vaccinology, approximately 600 were cloned in *E. coli*. Of those, about half (~350) were expressed, purified and used for producing antisera in mice. The sera were used in the standard bactericidal assay where each serum was incubated with the MC58 bacteria and complement. Of the 29 proteins that elicited sufficient bactericidal antibodies in mice, five were found to be adequately conserved to offer broad immunization coverage. Eventually, the vaccine was formulated with those five proteins combined with a single adjuvant (Giuliani, et al., 2006). Reverse vaccinology is also being used for formulating vaccines against parasitic pathogens including the malaria-causing Plasmodium (Tuju, Kamuyu, Murungi, & Osier, 2017), Leishmania, Schistosoma and Theileria (Lew-Tabor & Rodriguez, 2016).

Using the MenB design paradigm, this research has completed the first step of identifying the surface-associated proteins. It started with the downloading and curating of the whole genome sequences. The WGS were chosen over the segments of DNA such as expressed sequence tags because the WGS retain the contextual information of the reading frames and also enough well-curated data is now available. The suggested future work using the contextual information has been discussed below.

6,538 ORFs and 1,574 putative genes were computed from the WGS. The genes did not include known genes. Out of these, 1,045 ORFs were selected whose translated

protein had a potential to relocate to the surface. The ORFs had a full or partial overlap with the unknown genes. The ORFs were translated into proteins and the probability of those proteins translocating to the cell surface were calculated. Finally, the results were narrowed down to a list of 330 proteins that could be candidates for a vaccine.

The results supported the objective of this investigation. Starting with the raw genomic data, a methodology was established to identify the proteins that could be included in a vaccine for *S. aureus*. The proteins must be further investigated in a lab, as was done by the MenB vaccine development team.

Interestingly, the proteins translated from the ORFs matched hypothetical proteins and uncharacterized proteins of *S. aureus* in 10.66% and 2.75% cases, respectively. A label of hypothetical or uncharacterized signifies that the existence of the protein has been predicted but it has not been isolated *in vivo*. It can be deduced from the results that a large number of unexplored proteins still exist that could be antigenic targets.

A secondary objective of this research was to establish a procedure for using reverse vaccinology for developing vaccines against the microbes. There was no *S. aureus* specific processing at this stage of the project. The computer code and algorithms created thus far could be ported to other protein expressing pathogens including bacteria and fungi. Notably, this method is not suitable for viruses because they use the host's genome replication machinery to create the DNA, RNA, proteins and other building blocks needed for their progenies.

There are numerous experiments that could logically follow the work done so far. While the results presented here traversed the pathway of finding ORFs and genes and then selecting the candidate proteins that have the potential for localizing to the cell

surface, many other avenues were explored as a part of this research that used WGS as the primary input. These included the study of the pathogenicity islands in the virulent strains and the Interleukin-17 (IL-17) pathway as an adjuvant to a vaccine such as SA4Ag. Disrupting the coagulase-initiated cascade, discussed earlier, is another viable solution. Targeting IL-17 and coagulase will be a strategy similar to anti-toxin vaccines. Those are some of the pertinent experiments to continue.

The ORFs were extracted and processed in isolation, discarding any contextual information. The gene expression is a complex process that is modulated by other structures such as the promoters and inhibitors. These and other modulators may flank the gene or could be hundreds of thousands of base pairs away. There are bioinformatics tools for finding the modulators of gene expression (Babur, Demir, Gönen, Sander, & Dogrusoz, 2010). Future work could include the modulators along with the genes and ORFs to more accurately predict the target proteins.

There are genomic signatures that differentiate the virulent strains from the non-pathogenic ones. Further, only the ORFs that do not overlap with a known gene were investigated here. This was done to explore novel targets. Including the known genes will augment the pool of target proteins. Virulence sensitive processing and including the known genes will enhance this investigation.

While this research focused on pure proteins, *S. aureus* also produces a rich repertoire of membrane bound lipoproteins and glycoproteins. The future studies could be extended to the lipoproteome and glycoproteome of the bacterium. About 2-3% the *S. aureus* genome is devoted to the lipoproteome, coding about 70 lipoproteins. The resulting lipoproteins attach to the outer leaflet of the membrane by di- or tri-acylglyceryl

moieties. In addition to being linked with the bacterial virulence factors, they perform surface associated functions such as cell wall synthesis, signal transduction and uptake of nutrients (Graf, et al., 2018).

Some of the follow-up experiments mentioned above are the works-in-progress. The intermediate results of those studies have not been presented here to limit the scope of this document.

References

Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen Recognition and Innate Immunity. *Cell, 124*(4), 783-801.

Alam, M., Read, T., Petit, R., Boyle-Vavra, S., Miller, L., Eells, S., . . . David, M. (2015). Transmission and Microevolution of USA300 MRSA in U.S. Households: Evidence from Whole-Genome Sequencing. *mBio, American Society of Microbiology, 6*(2), 1-10.

Anderson, A., Miller, A., Donald, R., Scully, I., Nanra, J., Cooper, D., & Jansen, K. (2012). Development of a multicomponent Staphylococcus aureus vaccine designed to counter multiple bacterial virulence factors. *Human Vaccines & Immunotherapeutics, 8*(11), 1585-1594.

Babur, O., Demir, E., Gönen, M., Sander, C., & Dogrusoz, U. (2010). Discovering modulators of gene expression. *Nucleic Acid Research, 38*(17), 5648–5656.

Bagnolia, F., Fontanaa, M., Soldainia, E., Mishraa, P., Fiaschia, L., Cartoccia, E., . . . Grandia, G. (2015). Vaccine composition formulated with a novel TLR7-dependent adjuvant induces high and broad protection against Staphylococcus aureus. *PNAS, 112*(12), 3680-3685.

Barinov, A., Galgano, A., Krenn, G., Tanchot, C., Vasseur, F., & Rocha, B. (2017). CD4/CD8/Dendritic cell complexes in the spleen: CD8+ T cells can directly bind CD4+ T cells and modulate their response. *PLoS ONE, 12*(7).

Baxter, D. (2007). Active and passive immunity, vaccine types, excipients and licensing. *Occupational Medicine,, 57*(8), 552–556.

Begier, E., Seiden, D., Michael Patton, E. Z., Severs, J., Cooper, D., Eiden, J., . . . Gurtman, A. (2017). SA4Ag, a 4-antigen Staphylococcus aureus vaccine, rapidly induces high levels of bacteria-killing antibodies. *Vaccine, 35*, 1132–1139.

Besemer, J., & Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruse. *Nucleic acids researc, 33*, 451-454.

Bragazzi, N., Gianfredi, V., Villarini, M., Rosselli, R., Nasr, A., Hussein, A., & Behzadifar, M. (2018). Vaccines Meet Big Data: State-ofthe-Art and Future Prospects. From the Classical 3is ("isolate–inactivate–inject") vaccinology 1.0 to vaccinology 3.0, vaccinomics, and Beyond: A Historical Overview. *Frontiers in Public Health, 6*(62), 1-9.

Bragazzi, N., Gianfredi, V., Villarini, M., Rosselli, R., Nasr, A., Hussein, A., . . . Behzadifar, M. (2018). Vaccines Meet Big Data: State-of-the-Art and Future Prospects. From the Classical 3Is ("Isolate-Inactivate-Inject") Vaccinology 1.0 to

Vaccinology 3.0, Vaccinomics, and Beyond: A Historical Overview. *Frontiers in Public Health, 6*(62).

Bröker, B., Mrochen, D., & Péton, V. (2016). The T Cell Response to Staphylococcus aureus. *Pathogens, 5*(1).

Burge, C., & Karlin, S. (1998). Finding the genes in genomic DNA. *Current Opinions in Structural Biology, 8*(3), 346-54.

Carver, T., Harris, S., Berriman, M., Parkhill, J., & McQuillan, J. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England) , 28*(4), 464-9.

Chambers, H. (2012). Staphylococcal Infections. In L. Goldman, & A. Schafer, *Goldman's Cecil Medicine* (pp. 1815-1820). Philadelphia: Elsevier/Saunders.

Choksi, S., Babu, D., Lau, D., & Yu, X. R. (2014). Systematic discovery of novel ciliary genes through functional genomics in the zebrafish. *Development, 141*(1), 3410-3419.

Côrtes, M., Beltrame, C., Ramundo, M., Ferreira, F., & Sá Figueiredo, A. (2015). The influence of different factors including fnbA and mecA expression on biofilm formed by MRSA clinical isolates with different genetic backgrounds. *International Journal of Medical Microbiology, 305*(1), 140-147.

Cossart, P., & Jonquieres, R. (2000). Sortase, a universal target for therapeutic agents against Gram-positive bacteria? *PNAS, 97*(10), 5013–5015.

Dallal, M. M., Khoramizadeh, M. R., Amiri, S. A., Yaraghi, A. A., & Fard, R. M. (2016). Coagulase gene polymorphism of Staphylococcus aureus isolates: A studyon dairy food products and other foods in Tehran, Iran. *Food Science and Human Wellness, 5*, 186-190.

Delany, I., Rappuoli, R., & Seib, K. (2013). Vaccines, Reverse Vaccinology, and Bacterial Pathogenesis. *Cold Spring Harbor Perspectives in Medicine, 3*(a012476), 1-17.

Delcher, A., Bratke, K., Powers, E., & Salzberg, S. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer . *Bioinformatics, 23*(6), 673-679.

Foster, T. (2004). The Staphylococcus aureus “superbug”. *The Journal of Clinical Investigation, 114*(12), 1693-1696.

Fowler, V., & Proctor, R. (2014). Where does a Staphylococcus aureus vaccine stand? *Clinical Microbiology and Infection, 20*(5), 66-75.

Fowler, V., Allen, K., Moreira, E., Moustafa, M., Isgro, F., Boucher, H., . . . Betts, R. (2013). Effect of an Investigational Vaccine for Preventing Staphylococcus aureus Infections After Cardiothoracic Surgery: A Randomized Trial. *The Journal of the American Medical Association, 309*(13), 1368-78.

Frenck, R., Creech, B., Sheldon, E., Seiden, D., Kankam, M., Baber, J., . . . Girgenti, D. (2017). Safety, tolerability, and immunogenicity of a 4-antigen Staphylococcus aureus vaccine (SA4Ag): Results from a first-in-human randomised, placebo-controlled phase 1/2 study. *Vaccine, 35*, 375–384.

Giersing, B., Dastgheybb, S., Modjarradc, K., & Moorthy, V. (2016). Status of vaccine research and development of vaccinesfor Staphylococcus aureus. *Vaccine, 34*, 2962–2966.

Giuliani, M., Adu-Bobie, J., Comanducci, M., Arico, B., Savino, S., Santini, L., . . . Pizza, M. (2006). A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences of the United States of America, 103*(29), 10834-10839.

Graf, A., Lewis, R., Fuchs, S., Pagels, M., Engelmann, S., Riedel, K., & Pané-Farré, J. (2018). The hidden lipoproteome of Staphylococcus aureus. *International Journal of Medical Microbiology, 308*(6), 569-581.

Guglani, L., & Khader, S. (2010). Th17 cytokines in mucosal immunity and inflammation. *Current Opinions in HIV AIDS, 5*(2), 120-127.

Guigó, R. (1998). Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology, 5*(4), 681-702.

Holtfreter, S., Kolata, J., & Broker, B. (2010). Towards the immune proteome of Staphylococcus aureus – The anti-S. aureus antibody response. *International Journal of Medical Microbiology, 300*, 176-182.

Huang, E., Pillai, S., Bower, W., Hendricks, K., Guarnizo, J., Hoyle, J., . . . Meaney-Delman, D. (2015). Antitoxin Treatment of Inhalation Anthrax: A Systematic Review. *Health Security, 13*(6), 365-377.

Jones, R., Liu, Y., Rigsby, P., & Sesardic, D. (2008). An improved method for development of toxoid vaccines and antitoxins. *Journal of Immunological Methods, 337*(1), 42-48.

Jukes, T., & Osawa, S. (1993). Evolutionary changes in the genetic code. *Comparative Biochemistry and Physiology, 106*(3), 489-94.

Kang, S., Yang, J., Kim, K., Yun, C., Holmgren, J., Czerkinsky, C., & Han, S. (2013). Anti-bacterial and anti-toxic immunity induced by a killed whole-cell-cholera toxin B subunit cholera vaccine is essential for protection against lethal bacterial infection in mouse pulmonary cholera model. *Mucosal Immunology, 6*, 826–837.

Karst, S., Dueholm, M., & Mcilroy, S. (2018). Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nature Biotechnology, 36*(2), 190-195.

Kennedy, D., & Read, A. (2017). Why does drug resistance readily evolve but vaccine resistance does not? *Proceedings. Biological Sciences, 284*(1851).

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology, 305*(9), 567-580.

Lee, J. (2013). Overcoming Challenges in Staphylococcus aureus Vaccine Development . (pp. 1-18). Rockville, Maryland: NIH, National Institute of Allergy and Infectious Diseases . Retrieved from www.niaid.nih.gov/sites/default/files/staphworkshop.pdf

Lew-Tabor, A., & Rodriguez, V. (2016). A review of reverse vaccinology approaches for the development of vaccines against ticks and tick borne diseases. *Ticks Tick Borne Diseases, 7*(4), 573-85.

Lipsitch, M., & Siber, G. (2016). How Can Vaccines Contribute to Solving the Antimicrobial Resistance Problem? *mBio, American Society of Microbiology, 7*(3), 1-8.

Mahasenan, K., Molina, R., Bouley, R., Bauecas, M., Fisher, J., Hermoso, J., . . . Mobashery, S. (2017). Conformational Dynamics in Penicillin-Binding Protein 2a of Methicillin-Resistant Staphylococcus aureus, Allosteric Communication Network and Enablement of Catalysis. *Journal of the American Chemical Society, 139*(5), 2102-2110.

McAdow, M., Missiakas, D., & Schneewind, O. (2012). Staphylococcus aureus Secretes Coagulase and von Willebrand Factor Binding Protein to Modify the Coagulation Cascade and Establish Host Infections. *Journal of Innate Immunity, 4*, 141-148.

Natale, P., Brüser, T., & Driessen, A. (2007). Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. *Biochimica et Biophysica Acta, 1778*(1), 1735-1756.

Ostergaard, L., Vesikari, T., Absalon, J., Beeslaar, J., Ward, B., Senders, S., . . . Harris, S. (2017). A Bivalent Meningococcal B Vaccine in Adolescents and Young Adult. *The New England Journal of Medicine , 377*, 2349-2362.

Patel, A., Calfee, R., Plante, M., Fischer, S., Arcand, N., & Born, C. (2008). Methicillin-resistant Staphylococcus aureus in orthopaedic surgery. *The Journal of Bone & Joint Surgery, 90-B*(11), 1401-6.

Planet, P. (2017). Life After USA300: The Rise and Fall of a Superbug. *The Journal of Infectious Diseases, 215*(1), 71-77.

Pope, C. F., O'Sullivan, D. M., Mchugh, T. D., & Gillespie, S. H. (2009). A Practical Guide to Measuring Mutation Rates in Antibiotic Resistance . *Antimicrobial Agents and Chemotherapy, 52*(4), 1209-1214.

Proctor, R. (2012). Challenges for a Universal Staphylococcus aureus Vaccine. *Clinical Infectious Diseases, 54*(8), 1179–1186.

Rappuoli, R., Bottomley, M. J., D'Oro, U., Finco, O., & Gregorio, E. D. (2016). Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *The Journal of Experimental Medicine, 213*(4), 469.

Read, T., Petit, R., Yin, Z., Montgomery, T., McNulty, M., & David, M. (2018). USA300 Staphylococcus aureus persists on multiple body sites following an infection. *BMC Microbiology, 18*(206), 1-12.

Salazar, N., Castiblanco-Valencia, M., Bezerra da Silva, L., Arantes de Castro, I., Monaris, D., Masuda, H. P., . . . Mattos Areas, A. P. (2014). Staphylococcus aureus manganese transport protein C (MntC) is an extracellular matrix- and plasminogen-binding protein. *Plos One*, e112730.

Sallet, E., Gouzy, J., & Schiex, T. (2014). EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics, 30*(18), 2659–2661.

Stapleton, P., & Taylor, P. (2002). Methicillin resistance in Staphylococcus aureus: mechanisms and modulation. *Science Progress, 85*(1), 57-72.

Stothard, P. (2000). The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques, 28*(6), 1102-1104.

Tong, S., Davis, J., Eichenberger, E., Holland, T., & Fowler, V. (2015). Staphylococcus aureus Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management. *Clinical Microbiology Reviews, 28*(3), 603-661.

Tuju, J., Kamuyu, G., Murungi, L., & Osier, F. (2017). Vaccine candidate discovery for the next generation of malaria vaccines. *Immunology, 152*(2), 195–206.

Vernikos, G., & Medini, D. (2014). Bexsero chronicle. *Pathogens and Global Heal, 108*(7), 305-311.