4.0 Discussion

Streptococcus pneumoniae continues to be an important cause of death especially amongst the very young and the elderly. These mortalities are mostly concentrated in low-income countries primarily in sub-Saharan Africa, and Asia [34]. The currently licensed vaccines have many limitations, including serotype specificity and the coverage of only a subset of serotypes. This leads to serotype replacement by non-vaccine type serotypes in carriage and a subsequent increase in diseases caused by these serotypes [15]. Current vaccines are also relatively expensive making it difficult for resource-limited countries, who are most affected by *S. pneumoniae* disease, to purchase. As a result, a great deal of research into trying to find vaccine candidates that are well conserved across all serotypes, immunogenic and cheap to make has been undertaken.

My research has looked at a specific class of *S. pneumoniae* proteins, the lipoproteins. Some of the lipoproteins in this dataset have already been mooted as vaccine candidates by various experimental methods [23, 153]. Here I have utilised both the largest sample collection of *S. pneumoniae* isolates as well as the highest number of serotypes of all the pneumococcal protein screening studies to date [80, 98]. Together, this dataset has enabled me to gain unprecedented insight into the conservation and level of diversity of these lipoproteins within different lineages of the pneumococcus.

The scoring method explained in the methods section and illustrated in Table 3.6 was used to rank the proteins, where more weight was given to larger proteins, proteins with good immunogenicity results (percentage of protein predicted as epitope), and more conserved (prevalent) proteins.

Lmb is highly ranked in this dataset. The gene encoding this protein has been identified recently to encode for a second zinc transporter lipoprotein called AdcAII and I will refer to it as AdcAII from now on [154]. Zinc has both catalytic and structural roles in many proteins but as it doesn't passively traverse the cell wall, the pneumococcus has to utilise specific transporters to internalise the zinc especially during invasive disease, where zinc availability is restricted [155]. Indeed, both zinc

and manganese (transported by PsaA) are crucial to the bacteria but must be regulated to maintain homeostasis as both an excess or a lack of them is detrimental to the pneumococcus [156]. Even before the discovery of AdcAII as a zinc transporter, researchers speculated that there must be at least one other lipoprotein involved in zinc and/or manganese transport as in-vitro growth of an *S. pneumoniae* PsaA and AdcA (zinc transporter lipoprotein) double-mutant was restored with the addition of zinc and manganese in their right proportions [156, 157]. Here, I have evaluated both zinc transporters, AdcA and AdcAII as potential vaccine candidates.

AdcAII has been predicted to be immunogenic with many linear and discontinuous epitopes predicted. The size of the protein and its relatively small number of alleles further enhances its potential as a vaccine candidate. The major alleles 1, 2 and 3 (Fig. 3.23), cover almost all the disease lineages, however, unless the alleles can induce cross-protective antibodies, inclusion of at least these three alleles in a vaccine may be required.

AdcA is also possess important qualities with overall score of 83. This protein is both immunogenic and is a larger protein than AdcAII with 501 residues. However, its allele count of 82 makes it a less attractive option. The fact that both AdcA and AdcAII are involved in zinc transport in the pneumococcus makes them functionally redundant and any successful vaccine must include both proteins as well as their disease associated alleles or cross-protective alleles where possible. If this can be achieved, these proteins will make interesting vaccine candidates since a study has shown a complete loss of virulence in an AdcA/AdcAII double mutant in mouse models of infection [158]. Intriguingly, single mutants of either of these genes have been shown to be significantly more invasive than wild-type T4R strain, meaning inclusion of only one of these proteins in a vaccine is not an option [159].

As previously mentioned, PsaA transports manganese and is essential for full virulence of the pneumococcus [157, 160]. PsaA was previously thought to be involved in adhesion because a *psaA* mutant *S. pneumoniae* had reduced adhesion to endothelial cells, hence affecting carriage. But reduced adhesion is now thought to be a secondary effect on surface adhesion molecules due to the ensuing manganese deficiency [68, 153]. Further, a recent study in mice has shown an increase in the IgG levels of 3 proteins including PsaA following colonisation to be partially protective

against non-invasive lung disease [161]. However, inconsistent results about its effect on sepsis have been reported [162, 163]. Consistent with previous findings, I found PsaA to be highly conserved across all serotypes in this study [164]. Also, it has one of the fewest number of alleles and is predicted to be immunogenic by several of the prediction methods used here. These findings and the fact that it is the only prominent manganese transporter in *S. pneumoniae* further supports its suitability as a protein vaccine candidate. However, it may have to be used in combination with other candidates to offer protection against certain serotypes (especially those producing a lot of capsule) in invasive disease, where it may be buried beneath the capsule.

AmiA achieved the fourth highest overall score mainly because of its sequence length and the number of linear epitopes predicted. Due to the fastidious nature of the pneumococcus, it depends on external sources for various amino acids (used as nutrients) and AmiA plays an important role in amino acid uptake as well as in the recycling of cell wall peptides [165]. AmiA is encoded by a member of a five-gene operon, which comprised of genes encoding 2 transmembrane proteins, 2 ATP binding proteins and AmiA as the substrate binding protein [166]. Mutations to this locus have been shown to increase resistance to aminopterin, methotrexate and Celiptium, however, mutation of AmiA alone does not confer full resistance to these molecules suggesting that other factors may also be important in the resistance mechanisms [167]. Furthermore, it was shown that the ami permease comprised of two other lipoproteins (AliA and AliB) with high protein similarity to AmiA. These proteins are also involved in oligopeptide transport as oligopeptide deficiency was observed only when all three lipoproteins were mutated [168]. Consistently, in this study, both *amiA* and *aliA* were seen to be missing in *S. pneumoniae* isolates associated with disease. AmiA was missing in a serotype 9V strain recovered from CSF and AliA was missing in several serotype-3 disease isolates. This apparent functional redundancy for both AmiA and AliA has significant consequences for a vaccine candidate as it means that these proteins are dispensable to the pneumococcus and it can potentially lose either of them to evade any vaccine designed to target them.

MalX is another important protein suggested to be involved in the uptake of maltodextrines such as maltotetraose but not in maltose transport itself [169]. Also,

among several proteins involved in α-glucan degradation and transport, MalX is one of 6 suggested to play a role in pneumococcal virulence [170, 171]. Inconsistently, this protein was found to be absent or truncated in 14 isolates in this dataset including disease isolates thus indicating that the pneumococcus can cause disease in its absence. This also indicates that another protein may also play a role in α-glucan degradation. If this is true, a vaccine targeted to MalX may result in the pneumococcus losing this protein to escape the vaccine and continue to cause disease.

The pneumococcal lipoproteins involved in iron transport have been identified as pneumococcal iron uptake A (PiuA), pneumococcal iron acquisition A (PiaA), pneumococcal iron transport A (PitA) and the recently identified pneumococcal iron transporter, SPD_1609 [172, 173]. Interestingly, these lipoproteins were included for evaluation in my dataset. Although PitA has a low allele count, which is desirable, its short amino acid sequence length, low number of predicted linear and discontinuous epitopes have led to it scoring the lowest amongst all the lipoproteins in my dataset. This is consistent with the fact that a PitA mutant *S. pneumoniae* showed no difference in iron acquisition or virulence when compared to the wildtype [173]. The recently defined iron transporter SPD_1609 has a similar iron acquisition potential as PitA [172], therefore, an SPD_1609 mutant would likely have no effect on iron acquisition or virulence. This lipoprotein also has 103 alleles and an overall score of 76.7, which is one of the lowest scores in this dataset.

Conversely, both PiaA and PiuA have been identified as the major iron acquisition lipoproteins and are essential for full virulence in mouse models of invasive pneumococcal disease. Mice were also protected against systemic and respiratory disease when immunized with recombinants of both PiaA and PiuA and these protections were serotype independent [23, 24, 174]. Interestingly, both lipoproteins achieved good overall scores in my ranking. Consistent with a previous study, PiuA was found in all the genomes I screened, but PiaA was missing in a few NTs [175]. Although these proteins are highlighted as potential candidates both in previous studies [23, 176] and this current study, the fact that they are functionally redundant means that unless all iron transporter lipoproteins are included in a vaccine, with time, the pneumococcus may lose them and evolve to use the other iron transporters more efficiently to evade vaccines targeting only these, PiaA and PiuA.

The YesO_2 protein is also one of the highly ranked lipoproteins in my dataset with an overall score of 92.5, which is the 8th best score. This protein belongs to the extracellular solute-binding protein family 1 and is involved in sugar transport [3]. It is 100% present in my dataset suggesting its importance to the pneumococcus. Despite the seemingly long branches of the phylogenetic tree, this protein has only 19 alleles suggesting a lot of the variation is caused by synonymous mutations. Although it has 19 alleles, allele 2 (Fig. 3.31) was found in the majority of the genomes screened, including almost all the disease isolates. Epitopes were predicted across the entire protein (Fig. A26). Even though it may be argued that some of these proteins may encounter immune cells because of their small size, which means the capsule will completely cover them, this protein is larger than PsaA, which has been shown to come into contact with the immune system [163]. Therefore, YesO_2 is also expected to come into contact with these immune cells especially during carriage, where most strains have been shown to express less capsule.

PrsA_1 is a foldase protein annotated to be involved in protein folding and transport [177]. It has an overall score of 91.8 on my scoring algorithm thereby placing it amongst the top ten ranked proteins in this data set. This protein is expressed on the cell surface and, although its sequence length is similar to that of PsaA, it has 4 chains making it a very large protein perhaps capable of protruding through the cell wall. The protein has 24 alleles but like YesO_2, 3 alleles (1, 2 and 4) represent almost all the genomes screened. Allele 1 is the most prevalent of the three while allele 2 is only found in serotype 1 lineages Fig. 3.26. Allele 1 and 4 have a single amino acid substitution at position 50, from asparagine (N) to serine (S) both of which are hydrophilic [178]. Indeed, the same is true for allele 1 and 2, with valine (hydrophobic) at position 38 in allele 1 substituted by isoleucine (hydrophobic) in allele 2. Because both substitutions involve amino acids with similar properties, a subtle or no change to the protein folding is expected, hence it is highly likely that antibodies against one allele will cross-react with the other. In fact, the most divergent alleles only differed by 6 amino acids (97.476% identical). The entire surface of PrsA_1 is predicted by ElliPro to be immunogenic (Fig. 3.42). To my knowledge, this is the first time this protein has been evaluated as a potential vaccine candidate and taking together its attributes, it has good potential especially if used alongside other proteins.

Group_2005 lipoproteins are carbohydrate substrate-binding proteins belonging to the newly classified sub-class G transport proteins [179]. Due to the pneumococcus' dependence on carbohydrates as a source for carbon, approximately one-third of uptake systems are dedicated to carbohydrate transport, and 7 of these are ABC transporters hence, this lipoprotein is functionally redundant [179, 180]. Group_2005 is a large protein found in both monomeric and dimeric states [179] with a relatively small allele count of 26. Nonetheless, there is only a single dominant allele (1) that is present in almost all the genomes and together the alleles are less than 3% divergent (12 amino acid substitutions) suggesting that they may produce cross-reactive antibodies. The epitope predictions gave strong indications that this lipoprotein is immunogenic and its size gives confidence that it encounters immune cells, at least during colonisation. These qualities enabled this protein to attain the highest score in my ranking. Taking these factors into account, this protein maybe considered for inclusion in a multi-protein vaccine, due to its functional redundancy.

TauA is one of the most interesting proteins in this dataset with very short branch lengths aside from a group of NTs and a low number of alleles (15). This lipoprotein belongs to the periplasmic binding protein-like II family and functional family 84595 [181]. It is 100% present in all the genomes screened and it has a bigger size than PsaA, suggesting contact with immune cells. Further, the epitope predictions also suggest that it is capable of inducing sufficient immune response. The divergence of this protein at the amino acid level is less than 2% (98.214%) driven by only 6 amino acid substitutions. This means that unless homologs can be found in other species, the only source of divergence will be SNPs. This lipoprotein therefore possesses most of the characteristics of a potentially successful vaccine and should be investigated further.

Group_2056 lipoproteins have a large single chain and belong to the extracellular solute-binding family 1 functional family. Like Group_2005 proteins, they are also carbohydrate transporters [181]. With an overall score of 86.3, this is indicative of a good vaccine candidate. It has 35 alleles but it is clear that allele 1 is by far the most dominant allele (Fig. 3.17). Both the linear and discontinuous epitope counts are indicative of immunogenicity. Nonetheless, the fact that it is functionally redundant

means that it will most likely fail as a single vaccine antigen but, may be considered in a multiple-protein vaccine.

MetQ is smaller than PsaA but also ranked high on my list. It is a D-methionine binding lipoprotein involved in the biosynthesis of phospholipids [177]. It is present in all the genomes indicating its importance to the pneumococcus and it has only 28 alleles. It has few dominant alleles with a serotype 1 specific allele (Fig. 3.24). Nevertheless, only a 7-amino-acid difference exists between the most divergent alleles (97.544% identical) suggesting that antibodies against one may protect against other alleles.

PstS_2 is also a prospective candidate. It plays a role in phosphate ion transport by binding phosphate in the pstSCAB and phoU operon, although phoU does not play a role in phosphate transport [182]. It is present in all the genomes indicating the importance of phosphate to the pneumococcus. Interestingly, overexpression of this gene correlates with penicillin resistance while inactivation confers up to a two-fold susceptibility to penicillin [183]. This protein is also immunogenic based on the epitope predictions here and affinity of human sera in another study [176]. It also has few alleles, 23. The amino acid divergence is less than 3% (6 amino acid deletion). Together, these findings suggest that it is a promising candidate to be included in a vaccine. Otherwise, it may be a good drug target especially drugs used in combination with penicillin.

TcyJ and TcyA are both substrate binding proteins predicted to be involved in amino acid transport [181]. Although TcyJ has relatively many alleles, 48, only a small number of alleles were found in the majority of genomes and the least identical alleles were only 11 amino acid dissimilar. TcyA had less alleles and a higher number of ElliPro predicted discontinuous epitopes than TcyJ. Both have good characteristics for a vaccine candidate including good immunogenicity predictions, TcyJ is also dimeric and both have high prevalence in the screened genomes (TcyJ was missing in a single serotype 6B strain and TcyA was missing in only 5 genomes). However, the fact that both are functionally redundant means that they can only be considered for inclusion in a multi-protein vaccine alongside other proteins with similar functions.

VanYb is the name given by Roary but this is 100% identical to DacB of the pneumococcus which encodes LD-carboxypeptidase and it shall be referred to as DacB henceforth [184]. This protein works in concert with another protein called DacA to preserve cell shape and also plays an important role in cell division [185]. Here, DacB is absent in only two carriage strains of serotype 6A. It has 3 chains and 3 discontinuous epitopes predicted by ElliPro, which are very good qualities for any vaccine candidate. However, it also has a high allele count of 53, which are more than 15% amino acid divergent. With so much diversity, the alleles may not induce cross-reactive antibodies against each other meaning more than one allele must be included in a vaccine. Also, recombination between different alleles may drive vaccine escape. This makes VanYb a less attractive vaccine candidate.

TmpC is a well-conserved lipoprotein, present in all but one serotype 7F carriage strain. For a 350AA protein, its allele count of 21 is relatively small. The alleles are less than 5% divergent and are distributed evenly with no allele found uniquely in one lineage (Fig. 3.29). Also, the epitope predictions suggest it is immunogenic (Fig. A68 and A69). This protein is most likely involved in nucleoside transport because it has similar domain structure to purine nucleoside receptor A (PnrA), formerly called TmpC of *Treponema pallidum* [181, 186]. This protein is therefore a potential vaccine candidate especially if the alleles can induce cross-reactive antibodies.

Although Group_953 lipoproteins have undefined function, they were present in all but a single serotype 9V isolate recovered from carriage. This protein has 28 alleles but allele 1 was found in approximately 70% of the genomes. The epitope prediction results and the size of the protein are also favourable. However, because the function of this protein is unknown, it may be functionally redundant meaning the pneumococcus could lose it to escape vaccines against it. Further investigations must be made to determine its role before it can be a genuine vaccine candidate.

GlnH is an amino acid ABC transporter lipoprotein involved in glutamine transport [177]. Like many proteins in this dataset, it has interesting characteristics including a good immunogenicity prediction but a great many alleles (67), some of which are more than 5% divergent. Similarly, ArtP_1 proteins are also involved in glutamine transport [181]. Although both proteins have almost the same overall score, ArtP_1 had

significantly less alleles (38) and was present in all the genomes. Nonetheless, both proteins are functionally redundant so unless all the amino acid transporters are included in a vaccine, any vaccine targeting them singly will likely fail.

Another lipoprotein involved in amino acid transport is LivJ, which has a high affinity for branched-chain amino acids [181]. Although this protein has good immunogenicity predictions both in this study and another [176], the fact that this protein was absent in more than 20 isolates including disease isolates suggests that it may not be essential for *S. pneumoniae* pathogenesis, and therefore an unlikely vaccine candidate.

Group_6587 is ranked in amongst the upper half of proteins with an overall score of 91. These are lipoproteins predicted to be involved in protein folding [181]. Despite its small size, its allele count of 28 and good immunogenicity predictions are good characteristics for a vaccine candidate. Furthermore, it was present in all the genomes screened. However, like many other proteins in this dataset, it may only be successful being part of a multi-protein vaccine because of its functional redundancy.

Group_1655 lipoproteins have a short amino acid sequence (165) with a relatively high allele count of 36. Nonetheless it is 100% present in all the genomes and has very good immunogenicity results (Fig. A80-A81). It is an uncharacterised lipoprotein assigned functional family 247 [181]. Although it will be interesting to know its function, its relatively small size makes it a less attractive candidate. Group_2298, Group_2074 and Group_510 all fall under the same category of small proteins with amino acid sequence lengths of 185, 188 and 164 respectively. Group_510 lipoproteins are even less attractive as vaccine candidates, missing in more than 4% of the genomes and having a high allele count of 36. Both Group_2074 and Group_2298 lipoproteins are thioredoxin proteins (called Etrx1 and Etrx2 respectively) involved in oxidative stress resistance and redox homeostasis. A loss of both proteins affects virulence [187]. They were both at least 98% present in the genomes. Etrx1 (Group_2074) has the lowest allele count (11) of all the proteins in this dataset and the linear epitope predictions as well as the ElliPro predictions are positive. DiscoTope2 however did not predict a single discontinuous epitope for this protein (Fig. A60). Since both proteins must be targeted to affect virulence, both must be included in an effective vaccine but their

small size and the fact that no discontinuous epitopes were predicted for Etrx1 by DiscoTope2, makes them less attractive candidates [187]. The possibility of using these two proteins as drug targets could be explored because they are well conserved and play a vital role in the pneumococcus.

Although 30 proteins have been evaluated here, many of them have characteristics that make them less attractive candidates. That is not to say that the ones possessing better qualities are going to be any good *in vivo*. However, it is reassuring that this dataset includes previously studied lipoproteins. Immunization studies in mice have shown recombinant PiuA and PiaA to be protective against respiratory and systemic challenges [23, 24]. Antibodies to these two lipoproteins were also shown to promote opsonophagocytic removal of *S. pneumoniae* in human cell lines [83]. Furthermore, antibodies to these two proteins were recovered from convalescing septicaemia patients suggesting that they are both expressed in disease and also in healthy children suggesting immunogenicity in children as well [188].

Interestingly, Wizemann *et al* [80] utilised reverse vaccinology to screen the genomes of *S. pneumoniae* isolates for potential vaccine candidates. Of the 108 cloned products of 97 unique genes, none was protective against *S. pneumoniae* N4 in a mouse sepsis model, however, 5 of the products were shown to be protective against serotype 6B and 4 of the 5 products were also protective against serotype 6A [80]. However, none of these protein products were from lipoproteins.

Another study used a combination of genomics and human sera recovered from convalescing patients as well as healthy individuals exposed to pneumococcal infection. This study identified many epitopes belonging to many proteins including lipoproteins AmiA and MalX, however, only 6 (PspC, PspA, StkP, PcsB, SP0368 and SP0667) were identified as promising candidates [189]. None of these is a lipoprotein and StkP and PcsB showed the highest potential [189]. Furthermore, a study utilising reverse vaccinology, identified and analysed 13 conserved proteins initially thought to be unique to the pneumococcus for their potential as vaccine candidate [98]. These proteins included 4 lipoproteins including 2 thioredoxin family proteins, iron transporter PiuA, a glutamine ABC substrate binding protein and a lipoprotein of unknown function [98]. However, this study only evaluated these proteins for their conservation and diversity within different serotypes but no immunogenicity tests or predictions were performed. Also, some of the proteins identified as antibody binding targets in a study

utilising pan-genome wide immunological screening with human sera correlated very well with the candidate proteins in this dataset [176]. 208 antibody binding targets were identified based on their high affinity for adult human sera of which 16 were classified as substrate binding proteins. Of these 16, 10 are also present in my dataset, these include PsaA, PiuA, PiaA, PstS2, LivJ, GlnH, AmiA, MalX, PnrA (called TmpC here) and AliA [176]. Together, these experimental results support the fact that lipoproteins are immunogenic during disease and in carriage making lipoproteins interesting candidates for a protein vaccine.

Limitations

The limitations of this study include the fact that all the samples were retrieved from the Gambia, hence the findings may not be representative at the global scale. However, these findings will be relevant in the sub-Saharan context, where the burden of IPDs is enormous. Additionally, lipoproteins undergo post-translational lipidation of the conserved cysteine residue and this may extend to neighbouring residues to enable attachment to the cell membrane, therefore even though these regions may be predicted as epitopes, they would not be accessible to immune cells in-vivo meaning that they cannot be considered as true epitopes.

Conclusions and Future work

S. pneumoniae is an opportunistic bacterium that colonizes the nasopharynx of many people without causing disease. For its survival, it must obtain nutrients from its environment and it achieves this through various transporters, especially lipoproteins. Here, I have identified numerous vaccine candidates some of which could be further explored for inclusion in a protein vaccine. Some of these proteins have been previously studied, including iron transporters, PiuA and PiaA, manganese transporter PsaA, zinc transporters AdcA and AdcAII. Others including TauA, PrsA_1, and YesO_2, have not been evaluated until now. These proteins are serotype independent, which is an important characteristic of a prospective pneumococcal vaccine. An important caveat of most of these proteins, with the exception of PiaA,

however, is their presence in other non-pathogenic streptococci. An ideal vaccine candidate would target all pathogenic pneumococci and allow the non-pathogenic streptococci to fill the niche. Also, perhaps due to the importance of the nutrients transported by some of these proteins for pneumococcal survival and virulence, it has evolved to use several proteins to do the same job thus rendering some of them functionally redundant. This leaves PiaA as the single best candidate in this dataset but may be used with PiuA because of their synergistic effect on *S. pneumoniae* virulence. Therefore, for a successful protein vaccine, I believe multiple proteins especially of the same function must be included in the vaccine.

Moving forward, it will be essential to evaluate the impact of a multi-protein vaccine using the proteins in this dataset in mouse models of infection and carriage. Proteins with similar roles must be included in such vaccines. Furthermore, a multi-protein vaccine targeting at least two sets of proteins with different functions may work even better. However, for these to work effectively, it is essential to verify and choose alleles capable of inducing cross-reactive antibodies from each protein.

Furthermore, taking advantage of the wealth of genomes at our disposal, the techniques of evaluation described here could serve as a platform for future evaluations of other pneumococcal proteins as well as proteins of other bacterial pathogens. These will give valuable insight about the proteins with a better chance of making a good vaccine thus saving time and money doing animal studies on less suitable proteins.