# Genetic diversity and distribution of the pneumococcal surface lipoproteins and implications on potential protein-based vaccines

**Ebrima Bojang**
**University of Cambridge**
**Wellcome Trust Sanger Institute**

**This dissertation is submitted for the degree of Master of Philosophy**
**August 2017**

**Hughes Hall College**

## Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This thesis did not exceed the prescribed word limit by the Faculty of Biology.

## Abstract

*Streptococcus pneumoniae* causes life-threatening diseases such as meningitis, sepsis and pneumonia. Over half a million children under 5 years die annually of pneumococcal disease. However, most of these deaths occur in resource-limited countries mostly in sub-Saharan Africa and Asia. Based on the antisera binding pattern of the capsules, the pneumococcus has almost 100 serotypes and the currently licensed vaccines are serotype specific and target only a subset of these serotypes. The 23-valent polysaccharide vaccine is not immunogenic in young children and the conjugate vaccines, which are immunogenic in young children cover only a small number of serotypes and are expensive to manufacture. Furthermore, there is serotype replacement with non-vaccine type serotypes in both carriage and disease.

Consequently, there has been much interest in finding alternative vaccine candidates that are serotype independent, less expensive to produce and most importantly, can induce sufficient immune response. Several pneumococcal proteins have been evaluated for their potential as vaccine candidates with mixed results.

Using reverse vaccinology, I have taken a holistic approach to look at the level of diversity and distribution of core (≥90% presence in my dataset) pneumococcal surface lipoproteins and predicted their immunogenicity. First, I screened all the genomes for surface exposed lipoproteins using established patterns. The candidate proteins also underwent immunogenicity screening and these proteins were ranked based on their potential as vaccine candidates.

The final candidate proteins include previously evaluated lipoproteins PsaA, AdcA, AdcAII, PiuA, PiaA as well as several new candidates that have not been evaluated in detail thus far, including YesO_2, TauA and PrsA.

## Acknowledgement

First, I would like to thank Allah (SWT) for giving me life, health and the aptitude to undertake this dissertation.

Second, I am much grateful to my supervisor Professor Stephen Bentley who has been extremely helpful and supportive throughout this year. I would also like to thank Dr Simon Harris and Dr Lucy Weinert both of whom were on my thesis committee and helped guide this project by sharing with me their expertise. I would also like to say thank you to Chrispin Chaguza who took his time to introduce me to the tools available within the Infectious Genomics group. I would like to acknowledge everyone in E204 especially Rebecca Gladstone for always being there to answer my questions.

I would also like to thank my employer, MRC Unit, The Gambia for giving me the opportunity to undertake this prestigious degree. I am indebted to Dr Brenda Kwambana-Adams, Professor Martin Antonio and Dr Assan Jaye for seeing the potential in me and nominating me for this scholarship.

Thank you to the Wellcome Trust Sanger Institute for funding my studies and to the Graduate office for always assisting with all my needs

Finally, I would like to thank all my family and friends for their support and words. I dedicate this degree to my mum and dad. Both of them were instrumental in instilling discipline in me and making me realise that with hard-work, I can achieve all my dreams. Special thank you to my beautiful wife for the unconditional love and always being there for me. She has been my rock through-out this journey.

## Table of Contents

## List of Figures

## List of Tables

## List of Abbreviations

| Abbreviation | Full name |
|---|---|
| BCG | Bacillus Calmette-Guérin |
| CASP | Critical assessment of protein structure prediction |
| CBP | Choline binding protein |
| I-TASSER | Iterative threading assembly refinement |
| IL | interleukin |
| IPD | Invasive pneumococcal disease |
| NVT | Non-vaccine type |
| PAF | Platelet-Activation Factor |
| PDB | Protein Data Bank |
| Phyre | Protein Homology/AnalogY Recognition Engine |
| PI | Protrusion Index |
| SNP | Single Nucleotide Polymorphism |
| STGG | Skim milk-Tryptone-Glucose-Glycerol |
| VT | Vaccine type |
| WGS | Whole genome sequencing |

# 1 Introduction

## 1.1 The pneumococcus

*Streptococcus pneumoniae,* or simply “the pneumococcus”, is a Gram-positive diplococci that colonizes the upper respiratory tract (nasopharynx) of many healthy individuals (up to 80% in some settings) without causing disease [1]. However, it is also a bacterial pathogen able to breach the host defences thus causing disease. The mechanisms by which *S. pneumoniae* causes disease are not fully understood but often occur secondary to another respiratory tract infection making it an opportunistic pathogen [2, 3]. The pneumococcus causes a wide range of diseases in its host including less serious but more frequent diseases such as otitis media and sinusitis to life-threatening diseases such as meningitis, bacteraemia and sepsis [4, 5]. It is therefore, a very important cause of mortality and morbidity globally especially in children under the age of 5, patients with cardiopulmonary disease, immunocompromised patients as well as elderly people [6, 7].

*S. pneumoniae* has almost 100 known serotypes based on the antisera binding pattern [8]. The capsule is one of the most important virulence determinants of the pneumococcus and some capsular types are known to be more important than others in causing invasive disease [6]. Much of the diversity of these immunogenic capsules is believed to be caused by the selective pressure exerted by the host immune system [9]. It is on the basis of this knowledge that the currently licenced vaccines have been developed. These vaccines currently target only a subset of the most virulent capsules. This has now led to a reduction in carriage and disease of serotypes included in the vaccines (vaccine type (VT) serotypes) and an increase in carriage and disease of non-vaccine type (NVT) serotypes [10-12]. This phenomenon known as serotype replacement is apparent in many vaccinated populations including The Gambia [9, 12-15]. Further, serotype switching (strains acquiring a different set of capsule synthesis genes), has also been observed in vaccinated populations and vaccine pressure is thought to play some role in this [16, 17], since the currently licensed vaccines all target the capsule. Consistently, Croucher *et al*. [18] have shown that recombination hotspots seems to be concentrated around antibiotic resistant genes (*tetM*) and surface exposed proteins, which are potential vaccine candidates such as *pspA, psrP* and *pspC* as well as the capsule locus.

However, serotype switching is a natural part of pneumococcal evolution and has been known to have occurred decades before the introduction of vaccines [19]. Indeed, a study has revealed that the serotype switching event to serotype 19A, by a 23F lineage not previously found to have serotype 19A capsule, is thought to have occurred more than 10 years prior to the introduction of the PCV7 vaccine which targets 23F [20]. This shows that the 19A variants existed prior to PCV7 and have expanded to detectable levels after the selective reduction of VT serotypes by the vaccine.

The pneumococcus has a single circular chromosome that is approximately 2.1Mb in size with a G + C content of about 40% [21, 22]. Genome annotations have identified over 2000 protein coding genes in both TIGR4 and the un-encapsulated R6 strain but only over 60% of these have been assigned a biological function, leaving a great number of genes that could have vital roles in both disease and colonisation yet to be discovered [21]. Non-coding RNAs are also present with both TIGR4 and R6 having 4 rRNA operons and several other tRNAs [21, 22]. The pneumococcus has many insertion sequences (IS), which make up approximately 5% of the TIGR4 genome. Additionally, they possess a wide array of ATP-dependent transporters including the iron transporters, zinc transporters and manganese transporters. However, the most abundant transporters are the sugar transporters, which make up about 30% of all the transporters [21]. Some of these proteins are essential for full virulence of the pneumococcus therefore, they are being investigated as potential vaccine candidates [23, 24].

Also, the capsule synthesis genes, which determine the serotype are flanked by two conserved genes, *aliA* and *dexB* [25].

## 1.2 Pneumococcal colonisation

Colonisation is a prerequisite for infection and is more prevalent in younger children (<5 years of age) and can be as high as 80% in some countries but reduces to less than 10% as they reach adulthood [1, 26]. Most children in developing countries, would have been colonised with the pneumococcus at some stage of their childhood [27].

Carriage strains can be horizontally transferred from one individual to another and some of the risk factors for this include crowded areas such as day-care centres, hospitals, and prisons. Most of this horizontal transfer is believed to occur within children who are the main reservoirs of carriage [4]. Serotypes can be carried singly or simultaneously with other serotypes in the nasopharynx and this may last from a few weeks to months before being cleared and replaced by another type [1]. Although the length of carriage of different serotypes varies with some serotypes found much less often in carriage than others [28], this trait however does not affect a serotype's invasiveness, with some serotypes rarely found in colonisation studies shown to be amongst the most invasive [28, 29]. Furthermore, strains do compete against one another for colonisation. Some pneumococci produce strain-specific pneumocins which are inhibitory to other pneumococcal strains thus out competing them during colonisation [30]. This is why, when these more prevalent strains are cleared out due to vaccination, they are replaced by the previously suppressed strains [31].

Even though the capsular polysaccharide is the main determinant of immunogenicity, it is less important during carriage, thus the observed prevalence of transparent strains (strains with less capsule expression) in carriage [1, 4]. Consequently, the pneumococcus is known to express an array of proteins beneath the capsule that are essential for adhesion and thus colonisation [4]. These proteins interact with the host epithelial cells and ensure that the bacteria are anchored sufficiently to prevent innate immune clearance by ciliary movement [32].
Conversely, strains with increased expression of capsular polysaccharides are more often isolated in invasive disease because the capsule helps protect the pneumococcus against phagocytosis [1, 4]. The importance of capsule in invasive disease is further supported by the fact that un-encapsulated pneumococci are rarely if ever seen in invasive disease [1, 4].

It is worth mentioning that even though colonisation precedes disease, most colonised individuals do not go on to develop disease. The reason for this is not fully understood but transition to disease often requires the generation of local inflammation factors including tissue necrotic factor F and interleukin 1 [33]. Subsequently, this increases the number of receptors on their target cell (host cells) including the platelet-activation factor (PAF) receptor. The pneumococcus takes advantage of this scenario to bind to the PAF receptor which facilitates internalisation *en route* to causing invasive disease as depicted in Fig. 1.1 [1, 4, 33]. The host immune system may also play a role as disease burden is greatest among young children whose immune system hasn't fully developed, the elderly and immunocompromised individuals [34].

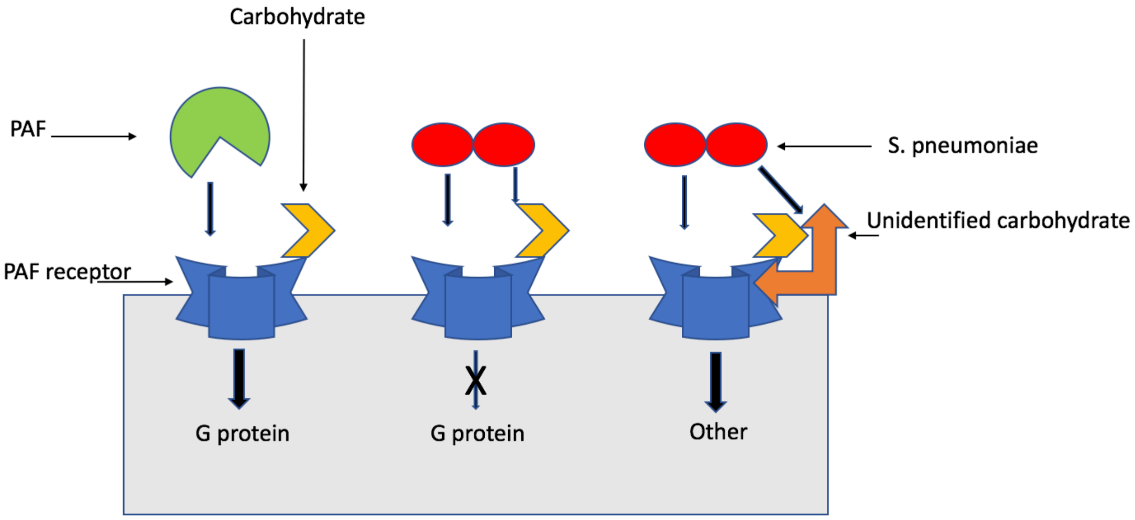


**Figure 1.1 Interaction of the pneumococcus with PAF receptor.**
The left illustrates PAF binding to the PAF receptor in a choline-depending fashion, thereby eliciting a G protein signal. The middle and the right diagram illustrate two proposed mechanisms of pneumococcal binding to the PAF receptor. The middle proposes that it engages both the PAF receptor and a carbohydrate on the PAF receptor or as depicted on the diagram on the right, it binds to the PAF receptor and another carbohydrate from an unidentified receptor that co-caps with the PAF receptor.
Picture adapted from [33].

## 1.3 Natural immune response to pneumococcal colonisation

There is evidence that the cytokine IL17A, which activates neutrophils, plays a significant role as an effector of rapid pneumococcal clearance in mice when challenged with non-encapsulated pneumococcal cells [35]. This clearance is antibody independent and it has also been demonstrated in *in vitro* studies that human IL17A cytokines were independently sufficient to induce pneumococcal killing by neutrophils [36, 37]. Further, there is proof of the role of CD4+ T cells in providing protection against colonisation and invasive pneumococcal disease in mouse models [38, 39]. Malley *et al*. argued that CD4+ T cells are the main source of protection against recolonization rather than antibodies and that this protection is serotype independent [39]. They further disputed the absolute necessity of antibodies in protecting against colonisation due to the fact that children often build resistance against invasive pneumococcal disease (IPD) from all serotypes before the appearance of measurable levels of anti-capsular antibodies [39].

## 1.4 Epidemiology and Burden

Pneumococcal diseases are a major problem worldwide but more so in resource limited countries. In 2000, the burden of serious pneumococcal diseases was estimated at 14.5 million cases worldwide which resulted in about 826,000 deaths in children between the ages of 1 and 59 months [40]. Furthermore, mortality due to pneumococcal diseases is estimated to account for about 11% of all deaths in HIV-negative children under the age of 5 globally and unsurprisingly, more than 60% of deaths occurred in sub-Saharan Africa and south Asia [40]. Nasopharyngeal colonization precedes invasive disease and pneumococcal carriage in healthy Gambian children under the age of 5 was shown to be 80% [26]. However, other studies have shown that the prevalence of carriage can vary between countries or even between cities of the same country [41, 42].

The most prevalent IPD is pneumonia, accounting for over 95% of pneumococcal disease/cases with pneumococcal meningitis reported to account for only 0.7% of all IPD worldwide [40]. IPD is responsible for most of the mortality caused by the pneumococcus however, some serotypes have a higher propensity to cause invasive disease than others [43]. Before the introduction of conjugate vaccines, PCV-7 serotypes accounted for about 90% of IPD in American and Canadian young children and at least 60% in all other regions except Asia, where they account for only 45% of IPD [6]. Nevertheless, the contribution of these serotypes to IPD was significantly lower in Europe than in USA and Canada or in Oceania, perhaps due to the significantly higher prevalence of serotype 1 and 5 in Europe than in these other regions [6].

## 1.5 Clinical Disease

As mentioned above, the pneumococcus causes a wide range of diseases, also depicted in Fig. 1.2.

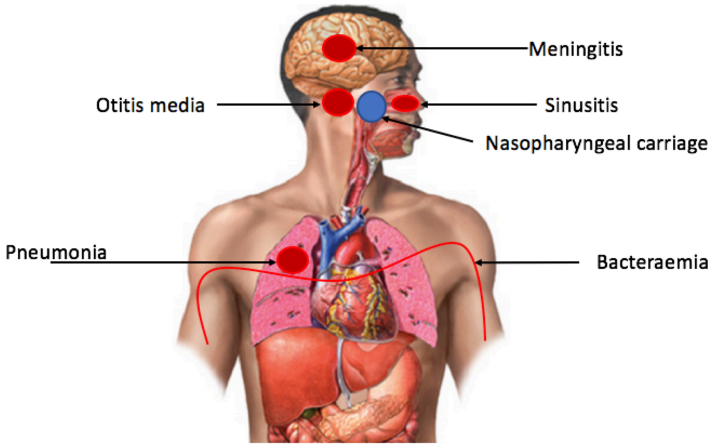


**Figure 1.2 Pneumococcal diseases and their anatomic sites.**
Schematic diagram showing the diseases caused by the pneumococcus and their site of infection.
Adapted from (https://www.slideshare.net/meningitis/human-28372932)

### 1.5.1 Pneumonia

Pneumonia is an infection of the lower respiratory tract often caused by bacteria but also caused by viruses [44]. *S. pneumoniae* has been implicated as the leading bacterial cause of pneumonia and this may be accompanied with bacteraemia in some instances [45]. Clinically, the signs and symptoms of pneumonia include fever and chills, rapid or difficult breathing, coughs and chest pain [46].

Pneumococcal pneumonia is by far the most common form of IPD globally [34]. Before the introduction of the PCVs, pneumococcal pneumonia was responsible for about 53% of all IPD in the USA [7]. *S. pneumoniae* is also the leading bacterial cause of community acquired pneumonia in European adults [47]. Furthermore, etiological studies of lower respiratory tract infections done in Zimbabwe and The Gambia both showed *S. pneumoniae* to be the commonest bacterial cause of pneumonia with incidence rates of 46% and 61% of patients respectively [44, 48]

### 1.5.2 Meningitis

Meningitis, as the name implies is the inflammation of the meninges and *S. pneumoniae* was the second most prevalent bacterial cause of meningitis before the introduction of conjugate vaccines against *Haemophilus influenzae* type b. However, it is now the number 1 cause in many countries [49]. Pneumococcal meningitis is the most devastating IPD that leads to death in up to 50% of patients without treatment. Also, long term consequences such as hearing loss, learning difficulties, seizures as well as brain damage may occur [7, 50]. In a 2010 meningitis outbreak in a Ghanaian district, *S. pneumoniae* was the leading cause of meningitis accounting for almost half (49%) of all bacteria isolated in the study [50]. Indeed, similar observations have been made previously in Ghana and neighbouring Burkina Faso between 2000- 2003 and 2002-2005 respectively, as well as in Malawi [51-53]. In all these studies, the case-fatality rate was approximately 40% with serotype 1 being the most prevalent serotype [51, 52]. In a more recent meningitis outbreak in Ghana (2015-2016), *S. pneumoniae* was again the leading cause accounting for 77% of all bacterial isolates of which 80% were serotype 1 [51, 54]. This recent outbreak is cause for concern because it occurred three years post-PCV13 introduction in Ghana [54]. Nonetheless, less than

5% of cases occurred in children under 5 years, which is the vaccinated group [54]. It is interesting to note however that these outbreaks resemble meningococcal meningitis serogroup A outbreaks, showing high levels of seasonality with peak incidences occurring between March and April. The fact that these countries fall within the meningitis belt, which runs from Senegal in the West to part of Ethiopia in the East further suggests that climate may be a factor [51, 52, 55]. Unfortunately, in Burkina Faso, the current licensed conjugate vaccines are not adequate for most of the serotypes implicated in pneumococcal meningitis [55]. PCV7 would cover only 33% of paediatric meningitis serotypes and a meagre 10% of adult meningitis serotypes [55]. Together, these results further support the need for vaccines with broader serotype coverage because the serotypes implicated in meningitis might be different from those mostly found in pneumonia *per se*.

### 1.5.3 Bacteraemia

Bacteraemia in its simplest term means the isolation of bacteria from blood and it is often secondary bacteraemia as a consequence of severe pneumonia or meningitis as they progress to causing death [45, 56]. When bacteraemia occurs without a known anatomic source of the infecting bacteria, it is called primary bacteraemia [57]. Bacteraemia is a prominent cause of death especially in young children and has been implicated as the cause of death in one-third of infants (<60 days) and a quarter of children older than 1 year in sub-Saharan Africa [58]. Interestingly, in that study and another from the same region, *S. pneumoniae* was the most commonly isolated bacteria from the blood cultures [56, 58]. Additionally, *S. pneumoniae* was amongst the most important pathogens in bacteraemia in studies outside Africa including Australia, England, USA and Denmark, which consistently showed it to be amongst the top 5 most commonly isolated pathogens in each study and overall the third most common pathogen [57].

Host genetics have been recently implicated to play a role in susceptibility to pneumococcal bacteraemia. Through genome-wide association studies (GWAS), an association with polymorphisms in two neutrophil expressed long intergenic non-coding RNA genes (lincRNA) was found in Kenyan children [56].

### 1.5.4 Otitis Media

Furthermore, the pneumococcus also causes non-invasive diseases, including otitis media (OT), which is a middle ear infection. Although this infection is less severe than IPDs, it is much more prevalent causing high morbidity in children. Prolonged untreated OT may result in ear and development problems [59, 60]. The burden of OT can be as high as 1.5 million annual cases in some regions of the world [61] and studies have shown *S. pneumoniae* to be the most important pathogen in otitis media [61, 62].

## 1.6 Antibiotic resistant pneumococcus

Antibiotic resistance is a global problem with many pathogens acquiring resistance to various classes of antibiotics and threatening to lead us to a post-antibiotic era. The data on resistant pneumococcal strains varies between continents and indeed between countries on the same continent [27, 63, 64]. In the Pre-conjugate vaccine era, some countries observed an increase in antibiotic non-susceptible *S. pneumoniae*. In a study in South Africa, as much as 95% of their hospital-acquired strains were resistant to penicillin [64]. Also, this study reported that as much as 9%, 12% and 4% of all their isolates were resistant to chloramphenicol, tetracycline, and erythromycin respectively [64]. An increase in cefotaxime resistance was also reported [64]. Of further concern is the observed increase in resistance to other antibiotics unrelated to penicillin such as vancomycin, which is the last line drug used in pneumococcal diseases [65]. In The Gambia, there was a slight but insignificant increase in penicillin, chloramphenicol or trimethoprim-sulfamethoxazole resistance amongst invasive pneumococcal isolates between 1996-2003 [66]. However, the increasing trend will most likely continue because of antibiotic abuse in many countries including The Gambia, where antibiotics can be obtained without prescription.

## 1.7 Licenced vaccines

Because of the large disease burden caused by the pneumococcus coupled with its rapidly decreasing susceptibility to the available antibiotics, it is necessary to explore the benefits of vaccination. The current WHO-recommended vaccine schedule of the conjugate vaccines is either three primary doses (the 3+ 0 schedule) or the 2 + 1 schedule, which includes two primary doses and a single booster dose [67]. The choice of schedule is completely dependent on the pneumococcal disease epidemiology of the population, the coverage as well as the timeliness of the vaccine doses [67].

Initially, a 14-valent polysaccharide vaccine was used until it was replaced in 1983 by the still in use 23-valent polysaccharide vaccine. This vaccine contained 23 of the most common serotypes implicated in invasive disease, accounting for about 88% of invasive disease [68].  However, this vaccine is only about 60% efficacious and it is not immunogenic in children younger than 2 years and the elderly (>65 years) who comprise the main risk groups [69]. Consequently, in 2000, the first conjugate vaccine was introduced in the USA, which included 7 serotypes conjugated to a non-toxin form of the diphtheria toxin protein [69].

Although this is more immunogenic in younger children and was licensed to be used in that age group, the conjugation to the protein meant that less serotypes could be incorporated in the vaccine. Nonetheless, this vaccine can induce a T-cell dependent immune response and studies including PCV-7, -9 and -11 have shown a disease risk reduction range between 62 and 89% in children less than 24 months old [68, 70]. Also, PCV13 has been shown to reach approximately 93% efficacy against VT invasive disease in children (<15 years) in Madrid in 2014-2015 [71].

To further improve vaccine coverage in all regions especially Asia, where PCV7 was less efficacious, another vaccine was formulated to include all the serotypes in PCV-7 plus serotype 1 and 5 to make the 9-valent vaccine. Together these serotypes accounted for approximately 66% of all IPD in Asia and >82% of IPD in all the other regions [6]. Furthermore, another two serotypes (3 and 7F) were added to the 9-valent formulation to make the 11-valent vaccine and this further improved the vaccine coverage of IPD in most regions by 2.6-6.5% and a substantial increase of 9% in Asia [6]. The 10-valent vaccine, which has eight serotypes conjugated to non- *Haemophilus influenzae* protein D and 2 other serotypes conjugated to either tetanus or diphtheria toxoids is also in use [72]. This vaccine has all the PCV9 serotypes plus serotype 7F [72].

Most countries, including The Gambia, have now included the PCV13 to their national immunisation programmes because it improves serotype coverage even further with two additional serotypes (6A and 19A) added to the 11-valent formulation [73-75]. Studies carried in vaccinated populations have seen significant reductions in the rates of IPD incidence as well as a reduction in VT serotype carriage in both vaccinated and unvaccinated individuals [76, 77]. The reduction of VT-carriage in non-vaccinated individuals is due to herd immunity, however, this is masked by the increase in carriage of NVT serotypes thus keeping carriage rates fairly constant [76, 78, 79].

## 1.8 Limitations of the current licenced vaccines

As indicated above, the currently licensed vaccines are limited in several ways including, low serotype coverage, serotype replacement and an increase in NVT IPD. All these limitations have prompted renewed efforts to identify conserved surface exposed proteins across all serotypes that are capable of inducing sufficient immune responses. Several identified proteins are in advanced stages of vaccine development [65, 68]

Further, the expensive cost of the polysaccharide conjugate vaccines is also a hindrance especially for low-income countries who are most affected by pneumococcal diseases. This makes the development of cheaper vaccines even more important. In the conjugate vaccines, every serotype is conjugated separately to the protein and this is the reason these vaccines are costly and can contain only a limited number of serotypes in one formulation [68]. To circumvent this problem, it is essential to manufacture protein-based vaccines which are relatively cheaper to produce.

## 1.9 Surface Proteins

*Streptococcus pneumoniae* like many other bacteria possess several surface exposed proteins that interact with host tissues and are believed to be essential for pathogenicity and survival *in vivo* by helping to conceal the bacterial surface from host defence mechanisms [69]. These proteins can be differentiated and grouped together based on their mechanism of attachment on to the cell surface. These groups include Choline Binding Proteins (CBP), proteins covalently bound to the peptidoglycan, histidine triad family macromolecules and those directly attached to lipids of the cytoplasmic membrane [80]. After sequencing the genome of TIGR4 by Tettelin *et al*., [21], they predicted a total of 36 lipoproteins in that genome. Generally, the pneumococcus possesses many lipoproteins performing many different roles with some expressed on the outer surface of membranes and others remain within the inner membrane [81]. Common to all lipoproteins are their identifiable signal peptides linked to their amino termini, which is essential for their transport through the cell membrane to the outer membrane but it does not exclusively determine which lipoproteins get to the outer membrane [69, 81]. Subsequent to its transport to the outer membrane, it undergoes further modification, which generally occurs in three steps. First, diacylglyceride is transferred to the cysteine sulphydryl group of the unmodified prolipoprotein. Second, signal peptidase II, which is specific for lipoproteins cleaves the signal peptide thus forming an apolipoprotein and finally, there is acetylation of the $\alpha$-amino group of the conserved N-terminal cysteine residue (Fig.1.3) [81, 82]. This leads to a highly hydrophobic amino terminus, believed to be firmly associated with the hydrophobic membrane [81].

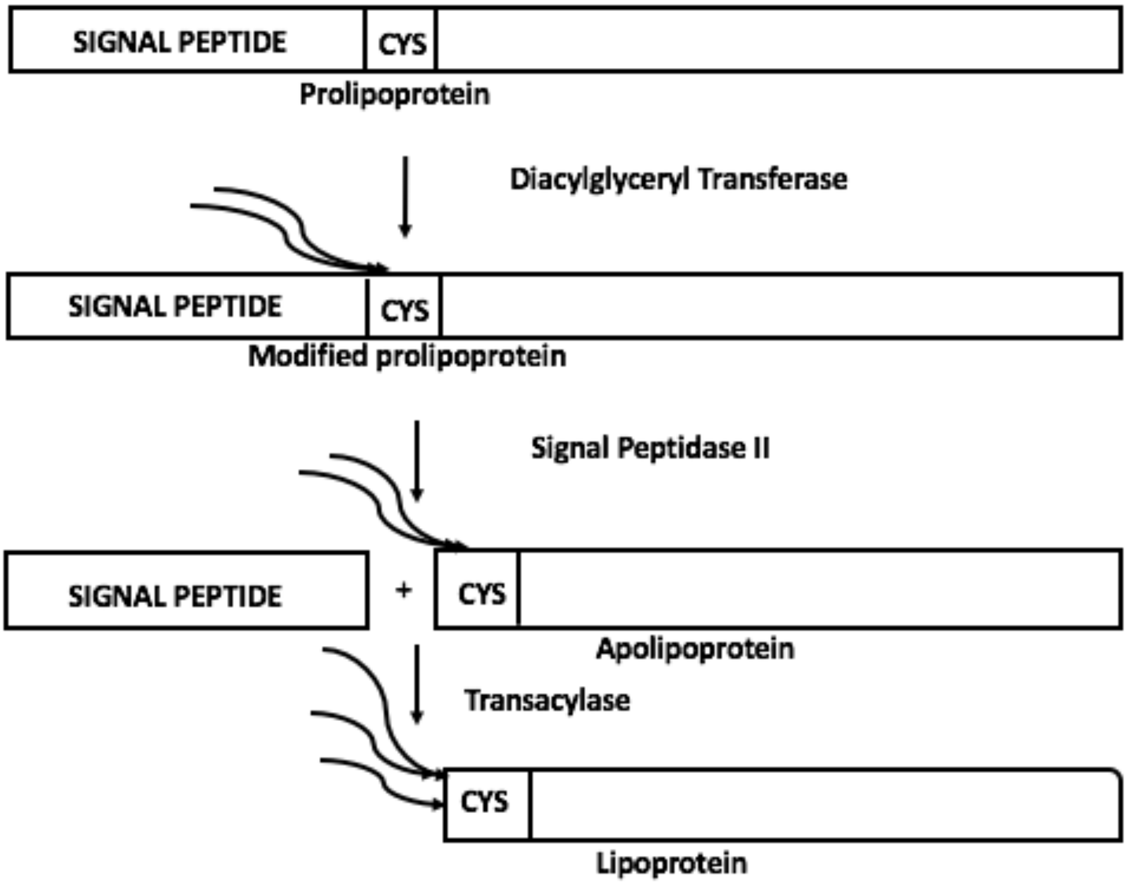


**Figure 1.3 Post-translation modification of lipoproteins.**
Modification process of the lipoprotein post export to the outer membrane. The signal peptide is to the left of the cysteine and the catalytic enzymes are written to the right of the reaction arrows. Adapted from Juncker *et al*. [82]

## 1.10 Some Protein Vaccines in the pipeline

As mentioned above, the limitations of the currently licenced vaccines have prompted research into finding alternative vaccine candidates. The most interesting candidates with such potential seem to be the pneumococcal proteins themselves. A number of these proteins have been promising as they have been shown to be essential for full virulence of the pneumococcus and vaccination with these proteins have been protective in mouse models of infection [23, 24, 83]. Two of these proteins were recently examined in a phase-II clinical trial in The Gambia to determine their efficacy against pneumococcal carriage but have been shown to be ineffective [84]. These proteins, pneumococcal enzyme pneumolysin (Ply) and pneumococcal histidine triad protein D (PhtD) were given in two different doses (10μg or 30μg) alongside pneumococcal 10-valent vaccine conjugated to non-typable *Haemophilus influenzae* protein D (PHiD-CV). The efficacy was measured by the prevalence of non-PHiD-CV serotypes in the nasopharynx but rather than seeing a reduction in these serotypes, there was an increase due to the void left by the reduction of the 10-valent vaccine serotypes [84]. Contrariwise, prior studies involving these proteins in mouse models were very promising with a study showing that passively vaccinated mice with human antibodies raised against these two proteins conferred protection against nasopharyngeal colonisation [85] and vaccinating with recombinant proteins protected against pneumonia [86].

Another pneumococcal surface protein that has been studied extensively with great promise is the pneumococcal surface protein A (PspA) [87, 88]. Recently, mice vaccinated with recombinant PspA (rPspA) and genetically modified pneumolysin (PdT) with recombinant BCG as an adjuvant and subsequently giving a booster of rPspA-PdT has been shown to protect these mice against lethal challenges [87]. The antibodies raised against both proteins were sufficiently high with a favourable shift in antibody class from IgG1 to IgG2. Vaccination also enhanced complement deposition and nullified the cytotoxic effect of Ply [87]. However, it has to be mentioned that even though there have been no experimental data to support this, there is fear that anti-PspA may react with myosin thus leading to an autoimmune disease [89].

Furthermore, pneumococcal choline-binding protein A (PcpA) has been studied for its immunogenicity and protection in animal models. Briefly, this candidate was shown to be protective against invasive pneumococcal diseases and has been shown to be both immunogenic and safe in a Phase I clinical trial when conjugated to PhTD or given alone (monovalent) [90, 91].

Together these are all encouraging steps towards finding an effective vaccine that is cost-effective, serotype independent and immunogenic in the main risk groups. Some candidates are shown to be more protective against colonisation [92], some more effective against systemic challenges [93] and some effective in both models [89]. Some are protective when used alone while some have shown synergistic effects when used in combination formulas [23, 89]. However, there remains a need to explore other potential proteins that may offer better coverage and might be more immunogenic to use as vaccines.

## 1.11 Application of Whole Genome Sequencing

The decreasing cost of next generation sequencing has led to more whole genome sequencing (WGS) projects. WGS of pathogens has enabled us to do several simultaneous studies on the same pathogen thus increasing our capacity to answer many biological questions. Since the sequencing of the pneumococcal genome by Tettelin *et al*. [21] in 2001, many studies utilising WGS have been done to improve our understanding of this important pathogen. Studies focused on the evolution of this pathogen have enhanced our understanding of its interaction with its host and other bacteria in the same niche. Further, we have better understanding about the mechanisms that contribute to pneumococcal evolution. Random mutations through point mutations and recombination are the major contributors to evolution, which are then subject to mechanisms such as natural selection and genetic drift [18, 94]. The larger size of DNA involved in recombination means that recombination introduces more diversity than point mutations [95]. A classic example of recombination due to selective pressure is in the event of resistance to antibiotics as these loci have been identified as recombination hotspots [17, 96].

Furthermore, WGS has been utilised in comparative genomic studies to identify virulence determinants that contribute to the different invasive propensities observed between similar serotypes found in different geographical locations [29]. WGS has also been used extensively to study the mechanisms of antibiotic resistance through horizontal gene transfer from other successful bacteria in the same niche [97].

The use of WGS in reverse vaccinology has also been explored. Wizemann *et al*. [80] used surface exposed protein motifs to screen for pneumococcal surface proteins. The identified proteins were cloned and recombinant protein used to immunise mice before being challenged with lethal doses of *S. pneumoniae* to look for protective candidates. The limitation of this study however was the low number of isolates with every serotype represented by only one isolate. This makes it impossible to evaluate the level of conservation of these proteins or even the diversity within the same serotype or between serotypes.

Furthermore, another study used a similar approach to scan for proteins present in only *S. pneumoniae* and absent in other Streptococcus species [98]. Although they chose 13 proteins that were present in all their isolates (51) covering 29 serotypes, these were also found in other Streptococcus species. Similar to the earlier study, their sample size of 51 was too small for conservation studies and they may have missed a lot of potential candidates in their study [98].

## 1.12 Thesis Aims and Objectives

With the above in mind, I have designed my MPhil to make a holistic evaluation of a particular family of pneumococcal surface proteins, 'the lipoproteins', as potential vaccine candidates.

The interest in lipoproteins is due to the many important roles they play such as in cell signalling, substrate binding and transport, antibiotic resistance as well as protein export and folding [99]. Some lipoproteins have been shown to play a role in both bacterial colonisation and pathogenesis. Pneumococcal lipoprotein PsaA has been shown to have an indirect role in colonisation and it is also essential for full virulence [100]. Furthermore, several other pneumococcal lipoproteins especially metal transporters have also been implicated in the virulence of this pathogen. These include the zinc transporters AdcAI and AdcAII and the ABC iron transporters PiuA and PiaA [23, 101]. Lipoproteins' role in virulence have also been demonstrated in other Gram-positive bacteria such as the LpK of *Mycobacterium leprae,* which induces human interleukin 12 during infection [102]. Also, their role in conjugation, protein secretion and localisation, sensing of their environment as well as their involvement in the spore cycle of *Bacillus subtilis* have also been speculated [103].

In this MPhil thesis, I will utilise the whole genome sequences of *S. pneumoniae* isolated from the Gambia through the Bill and Melinda Gates Foundation-funded Global Pneumococcal Sequencing (GPS) Project available to me at the Sanger Institute to evaluate lipoprotein genes. These samples include both invasive and carriage samples therefore, this will give me a broader perspective of the protein distribution in both sets of data. The aims are to:

1. Identify all the surface exposed lipoprotein genes in these genomes.
2. Determine their level of conservation (prevalence) and diversity
3. Determine if there is association between alleles of these genes and pathogenic potential (i.e. ratio of prevalence in disease vs. carriage)
4. Use bioinformatics techniques to predict surface exposure and antigenicity of these proteins, which should aid in identifying those that have a greater potential to be successful vaccine candidates.

# 2 Methods

## 2.1 Ethical Approval

Written informed consent was sought and obtained from all adult study participants and from guardian/parent of all participants under the age of 18 in the form of a signature or a thumb-print from participants who could not write. This study (GPS) was approved by the joint MRC/Gambia government ethics committee under the study number SCC1188.

## 2.2 Global Pneumococcal Sequencing Project

GPS is a multi-site study which aims to understand the population genetics of the pneumococcus in response to vaccinations with a particular focus on developing countries. This study officially started in October 2011 and deep sampling was carried out in 12 developing countries in sub-Saharan Africa, Asia and South America. Publicly available pneumococcal genome datasets were also included in this study. The aim is to sequence 20, 000 genomes spread across about 50 countries and including isolates from pre-PCV, during vaccination and post-vaccination (Fig 2.1). The founding partners are Wellcome Trust Sanger Institute, Emory University, Bill and Melinda Gates Foundation, Centre for Disease Control and Prevention, MRC-The Gambia, National Institute for Communicable Diseases (NIDC, South Africa) and Malawi-Liverpool-Wellcome Clinical Research Programme, however, as mentioned above my project is only focused on those isolates isolated in The Gambia from MRC-The Gambia.

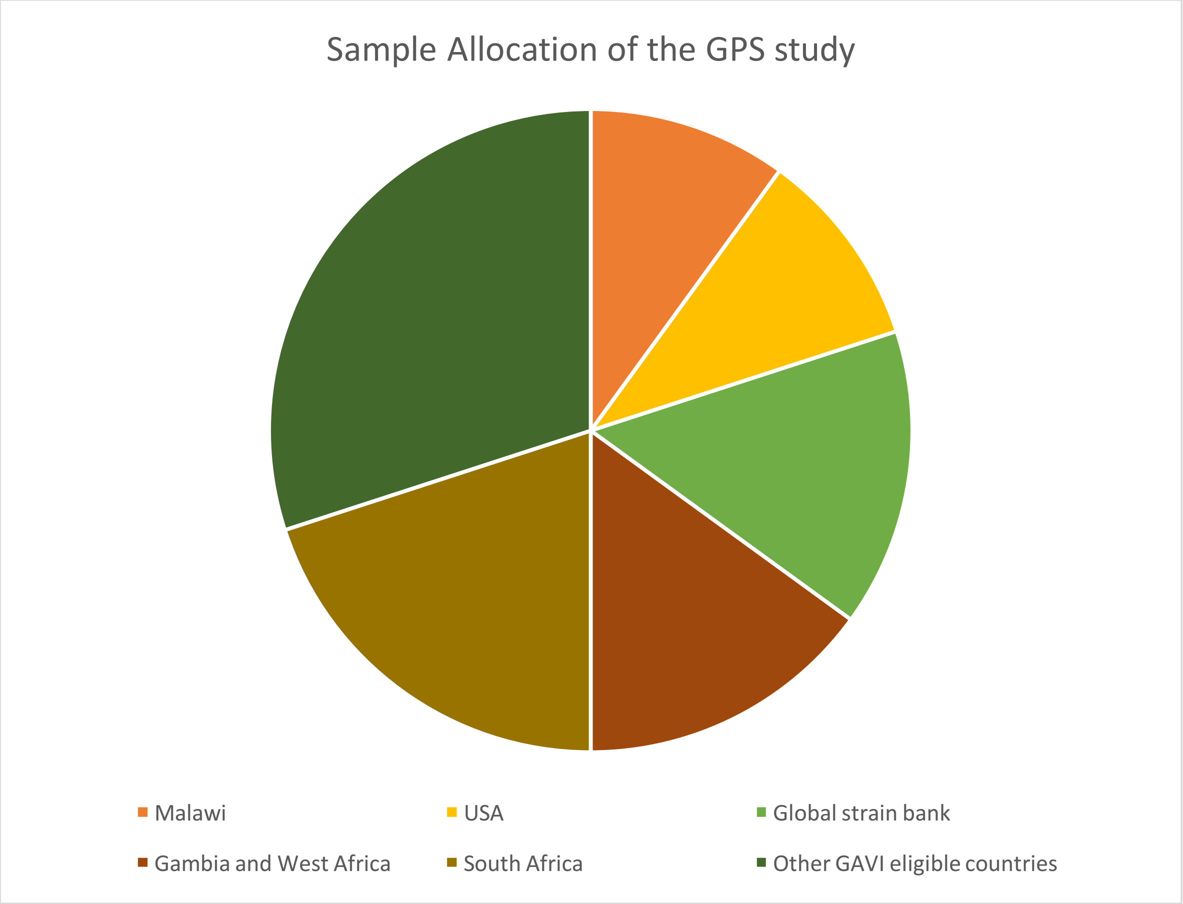


**Figure 2.1 GPS sample collection sites.**

The places from which the GPS samples are being obtained. USA is among the countries despite being a developed country because it was the first country to introduce a PCV vaccine.

Adapted from: (http://www.pneumogen.net/gps/project_outline.html)

## 2.3 Sampling

The Gambia is a small West African state with a population of approximately 2 million people [104]. Nasopharyngeal swabs (NPS) were collected from healthy adults and children in Sibanor, located in the Western River Region of The Gambia (Fig. 2.2). This region shares borders with Casamance, southern Senegal. Most people in this area are of the Jola and Mandinka ethnic group who are mostly subsistence farmers. Most of the other samples were collected from Fajara and the remaining from the

Central and Upper river regions of the Gambia. There are only two seasons in the Gambia, a shorter rainy season that runs from late June to October and a dry season. The Gambia introduced PCV7 in its expanded programme of immunisation in August 2009, applying WHO's 3+0 protocol where vaccines are administered to babies at 2, 3 and 4 months. PCV7 was later replaced with PCV13 in May 2011 [75].

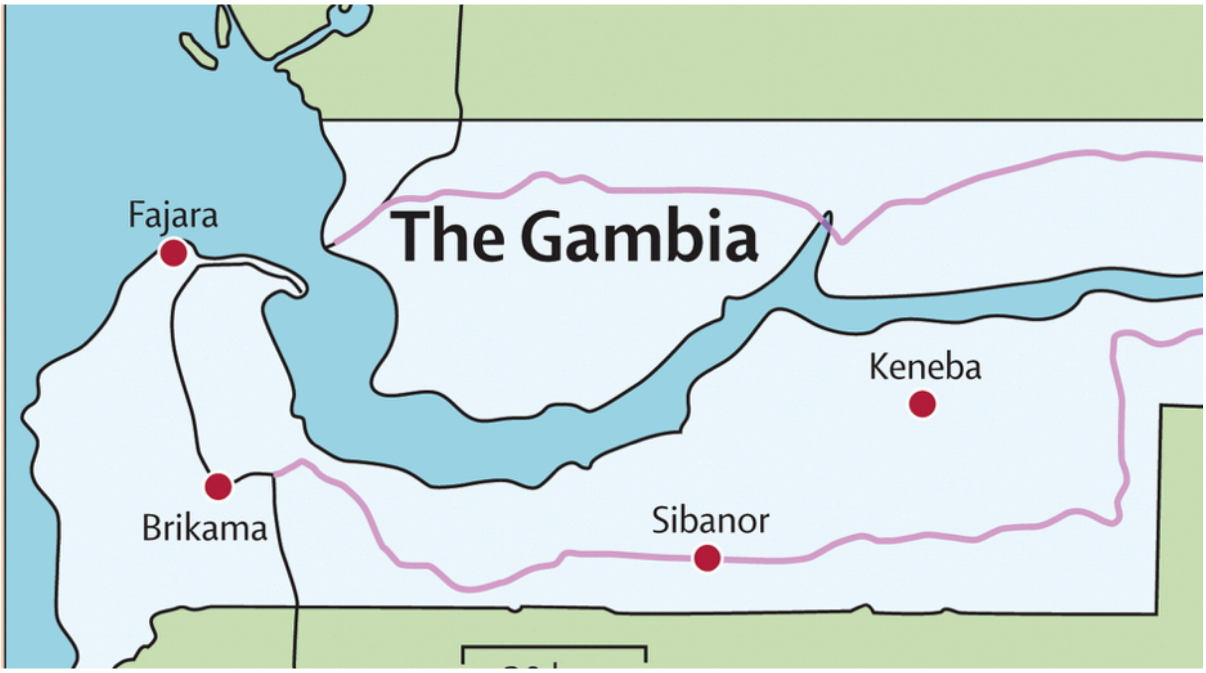


**Figure 2.2 Partial map of The Gambia.**

Partial map of The Gambia showing the two main sample sites, Fajara and Sibanor in the Western River Region. Samples were also collected from Central and Upper river regions of The Gambia, which are the middle part and the east-most part of the country respectively (Not shown in the map).

Adapted from Ceesay *et al*. [105]

Because of sampling constraints, to meet the quota, all the isolates that were isolated in the Gambia were sent for sequencing. These samples were divided into three age groups including, children aged ≤2 years, those aged more than 2 years but ≤5 years and those children and adults aged more than 5 years.

For the carriage study, a cotton swab was inserted through the nostrils up to the nasopharynx of both participating adults and children. The swab is then gently swabbed against the walls of the nasopharynx and samples stored in vials containing Skim milk-Tryptone-Glucose-Glycerol (STGG) and transported on ice to a -80 degrees Celsius storage facility within 8 hours prior to microbiological culture and isolation of *Streptococcus pneumoniae* as per WHO protocol [106]. Additionally, all invasive isolates of *Streptococcus pneumoniae* recovered from the MRCG ward were also included in this study.

Together with the swab samples, meta-data was also collected including the vaccination status of the individual, sex and age.

## 2.4 Dataset

The sample collection years in this dataset range from 1993 to 2014. They were all isolated from The Gambia and had passed post-sequencing quality control (QC). These include 1268 from carriage, 4 unknown and 497 isolates from invasive disease, which amounts to a total of 1769 genomes. The source of the invasive isolates is as follows; Ascetic fluid 2(0.4%), Blood 367 (73.8%), cerebrospinal fluid (CSF) 50 (10.1%), Knee aspirate 2(0.4%), lung aspirate 62(12.5%), Pleural aspirate/fluid 5 (1%), and 9(1.8%) are unknown. Although some isolates are missing data on gender, overall, there were more males than females in the study population with about 820 (43%) males and 650 (35%) females and the rest had no data on gender as illustrated in Figure 2.3.

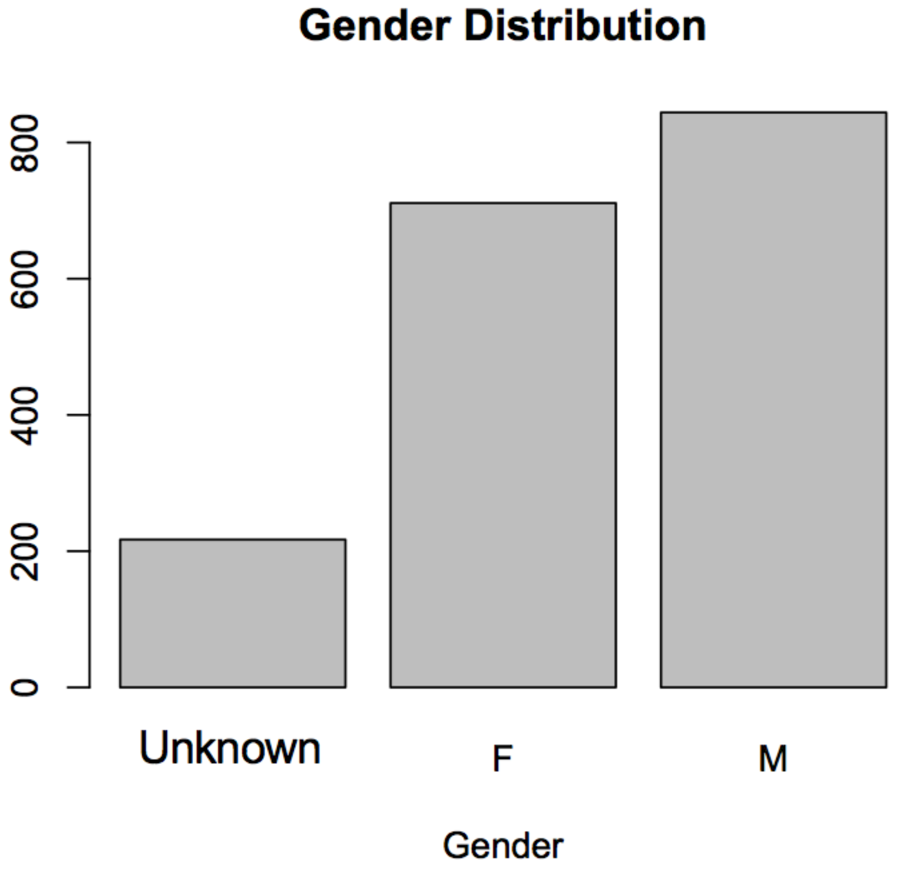


**Figure 2.3 Gender distribution of the isolates in this dataset.**

## 2.5 Microbiological isolation, DNA extraction and Quantification

All the genome samples sequenced in this dataset were isolated from MRC-The Gambia laboratories. *S. pneumoniae* was isolated from nasopharyngeal swab samples in the case of carriage and other sites including blood, CSF or lung aspirates for the invasive samples. The samples were isolated using conventional microbiology techniques as detailed in document identification code ASSAY-RML-123, version 1.0. Subsequently, a confluent growth of a sub-cultured single colony from each isolate was harvested and DNA extracted using QiaAmp DNA mini kit, also detailed in document number ASSAY-MML-003, version 4.0. The aim of the extraction is to isolate >1μg/mL of RNA free double stranded DNA. DNA quantification was performed using the Pico green technique also detailed in version 1.0. of document code ASSAY-MML-005.

## 2.6 Sequencing and Assembly pipeline

WGS was done on all the isolates at the Sanger Institute using the Illumina HiSeq platform. The mapping and assembly was automated using Sanger Institute customised pipelines [107]. 150bp short reads and paired-end reads were either mapped to ATCC 700669, serotype 23F (ST81) strain using BWA/SMALT or assembled *de novo* as illustrated in Fig 2.4 below. Initially, VelvetOptimiser (https://github.com/tseemann/VelvetOptimiser) was used to determine the kmer size prior to assembly by Velvet [108]. Scaffolding of the assembled contigs using paired-end reads to assess the orientation, order and distance was then achieved by SSPACE [109] and then GapFiller [110], which also uses paired-end reads was used to fill the gaps within scaffolds. Next, the assemblies underwent automatic annotation using Prokka [111].

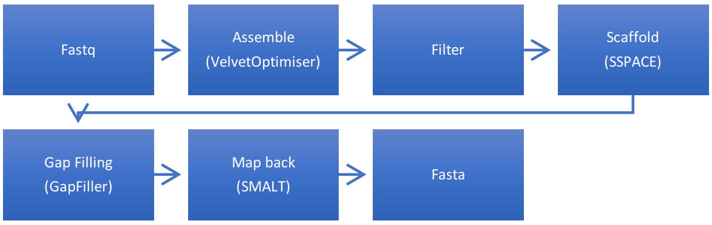


**Figure 2.4 *De novo* assembly with the Sanger Institute pipeline.**
An overview of *de novo* assembly of the Sanger Institute assembly pipeline. VelvetOptimiser determines the optimal kmer size then Velvet assembles the contigs. Subsequently, SSPACE is used for Scaffolding the contigs and GapFiller to fill the gaps within scaffolds.
Adapted from Page *et al*. [107]

## 2.7 Post-Sequencing Quality Control (QC)

The parameters for QC were developed to prevent the exclusion of good quality data but also importantly, to avoid the inclusion of bad quality data in the final dataset. Accordingly, all sequenced data must fulfil all the following conditions to be considered to pass the QC;

1. Sequence reads must map to >60% of the genome of pneumococcal reference strain ATCC 700669
2. The average coverage depth must be greater than 20x
3. Only <1% of reads assigned to another taxon other than the pneumococcus by Kraken [112] is allowed
4. Assembly length must be between 1,900,000 and 2,300,000bp long and
5. The reads must assemble to less than 500 contigs.

## 2.8 In-silico MLST and Serotyping

An *in-silico* MLST was performed using a script [113]. This script used seven house-keeping genes (*aroE, gdh, gki, recP, spi, xpt,* and *ddl*) to assign sequence types (STs) to all the genomes. Furthermore, to confirm the serotyping done by conventional Quellung method, an *in-silico* serotyping was also performed using pneumoCAT [114]. PneumoCAT maps reads to 92 known pneumococcal capsule loci and an addition two subtypes. When the reads match >90% to a single locus then the call is made immediately and the run terminates, otherwise, a second step is undertaken when the reads match >90% to more than 1 locus using a capsular type variant database to distinguish serotypes between serogroups [114].

## 2.9 Bayesian Analysis of Population Structure (BAPS) clustering

Further, hierBAPS [115] clustering was performed on a subset of all the GPS samples. This comprised of ~13,000 genome alignments and the cluster of the rest of the samples were inferred from their ST. The hierBAPS separates lineages in a dataset by clustering sequences of the same lineages together.

## 2.10 Whole genome phylogeny (FastTree)

In this study, FastTree [116] was used for the reconstruction of the whole genome phylogeny. This was chosen especially because of its speed with many sequences and because it produces trees close enough to trees produced by other precise maximum-likelihood methods [116]. Prior to building the tree, all the sequence reads and the reads from a non-typable strain from USA as the outgroup were mapped to *S. pneumoniae* reference strain ATCC700669 to create aligned pseudogenomes. Then the SNP sites were extracted using SNP-sites [117] excluding the reference strain. Finally, the SNP alignment was used to reconstruct the phylogenetic tree in FastTree.

## 2.11 Lipoprotein genes identification

Since, lipoproteins are the main focus of my analysis, I developed a multi-step algorithm to extract my genes of interest from the genomes and also use further bioinformatics tools to verify true lipoproteins. These steps are illustrated in Fig. 2.5. Initially, Roary [118] was run with the option to not separate paralogs. Roary produces a pan-genome reference file in fasta format containing a reference sequence for every gene in the pan-genome and gene-presence/absence file for every genome in comma separated value (CSV) format. The sequences in the pan-genome file are produced as nucleotides and therefore was translated into amino acids to enable querying with my lipoprotein specific patterns.

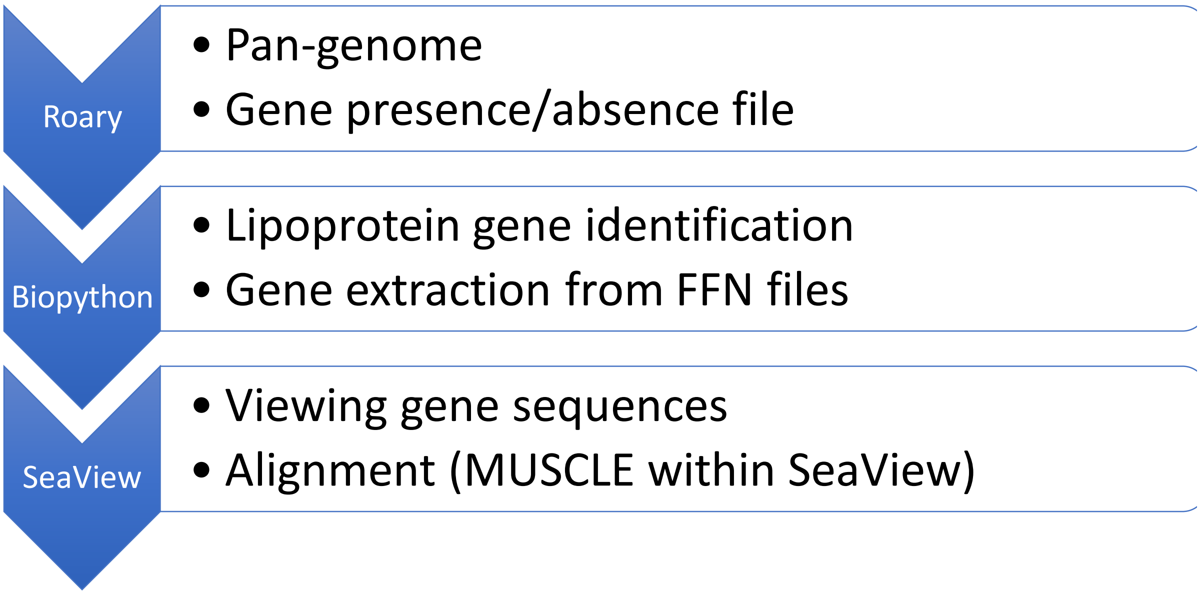

Roary
• Pan-genome
• Gene presence/absence file
Biopython
• Lipoprotein gene identification
• Gene extraction from FFN files
SeaView
• Viewing gene sequences
• Alignment (MUSCLE within SeaView)

**Figure 2.5 Steps taken to select candidate lipoproteins.**
This diagram shows the steps that were taken to identify lipoprotein gene, extraction of the genes and visualisation and alignment of the genes.

Subsequently, a Biopython [119] script was used to parse the translated pan-genome reference file, identifying lipoprotein genes using three different patterns. Briefly, the Prosite pattern PS52157 [120], the G+LPP[121] and the G+LPPv2[122] patterns were used (Table 2.1). The sensitivity and specificity of the G+LPPv2 have been showed to be 1.000 and 0.891 respectively when tested against a known Gram-positive lipoprotein dataset [122]. The G+LPP, G+LPPv2 and Prosite patterns produced 127, 136 and 167 lipoprotein hits respectively which resulted into a total of 169 unique hits. For lineages that seem to be missing a particular protein, I built a local blast database using their assembled genomes and using the specific gene sequence from a reference genome (D39) as query for the blast search [123]. This added step was performed to verify if they really lacked the protein or have a more divergent protein to the others. A further mapping step was performed on all the genomes missing a gene of interest to ascertain true absences. The reads of the genomes were mapped on to a reference gene (D39) with about 100bp flanking regions.

**Table 2.1 Patterns used for lipoprotein search.**

This table shows the patterns used to do the lipoprotein search with their pattern expressions. Adapted from Rahman *et al*. [122]

| Pattern | Pattern Expression |
|---|---|
| **G+LPP** | <[MV]-X(0,13)-[RK]-[^DERKQ](6,20)-[LIVMFESTAG]-[LVIAM]-[IVMSTAG]-[AG]-C |
| **G+LPPv2** | <[MV]-X(0,13)-[RK]-[^DERK](6,20)-[LIVMFESTAG**PC**]-[LVIAM**FTG**]-[IVMSTAG**CP**]-[AG**S**]-C |
| **Prosite (PS51257[a])** | [^DERK](6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C |

[a] The Prosite pattern has an additional rule that there must be a K or R in the first 7 amino acids and the conserved cysteine must appear between amino acid position 15 and 35.

Only those present in ≥90% of the genomes were selected for further testing. The selected lipoproteins were further investigated by using the online available bioinformatics tools; SignalP 4.0 a neural network-based method [124], Phobius, based on a hidden Markov model and predicts both transmembrane topology and the signal peptide of a protein [125], DOLOP [126, 127] and LipoP, which although developed for predicting lipoprotein of Gram-negative bacteria, still correctly predicts about 93% of Gram-positive lipoproteins [82]. It has also been described by Rahman *et al*. as being the single best performing tool for lipoprotein confirmation in their study [122].

Furthermore, the LipoP prediction server also gives the residue at position +2 of the conserved cysteine residue. It has been shown that having aspartate (D) at that position is associated with lipoproteins attached to the inner membrane, therefore, not expressed on the outer surface of the cell. However, this is not entirely straightforward as other positions also seem to play a part in it [81, 128].

By combining the search results of the Prosite, G+LPP and G+LPPv2 patterns, and subsequently verifying them with the online tools mentioned above, the chances of missing any lipoprotein from this screening are low.

## 2.12 Gene Extraction

With a high level of certainty, I set out to extract the nucleotide sequences of the selected protein from all the genomes. First, the annotation ffn files from Prokka of all the genomes were concatenated to create a database. Then, I developed a Biopython script that used the gene-presence/absence file to get the unique gene identifiers and used that information to extract the genes from the database.

## 2.13 Gene Visualisation, Alignment and Phylogenies

All the gene sequences were visualised for insertions, deletions and polymorphisms using SeaView [129]. These sequences were all aligned using MUSCLE [130] within SeaView.

The alignment files were subsequently used to build phylogenetic gene trees using Rapid Axelerated Maximum Likelihood (RAxML) [131]. This was run with the option to omit sequences less than 80% of the reference sequence length to avoid the addition of truncated genes in my tree.

## 2.14 Gene Allele assignment

Allele assignment was performed in two simple steps. First, the gene nucleotide sequences were translated to amino acids using SeaView [129]. This was done to exclude the effect of synonymous polymorphisms. Second, the amino acid sequence alignments of all the lipoprotein genes in this dataset were individually parsed using a script to assign alleles. This script takes the first sequence in the alignment and assigns it an allele number (i.e. 1), then iteratively, assigns a new allele number to any new allele variant found.

## 2.15 Tree Annotation

The whole genome tree was annotation with the serotype, BAPS cluster, disease status and gene-presence/absence information of all the candidate proteins using Phandango [132]. Each protein tree was also annotated with their serotype information, BAPS clusters, disease status and allele information using both Phandango and interactive Tree of Life (iTOL) [132, 133].

## 2.16 Protein Antigenicity

For a successful vaccine design, proteins are selected that are capable of inducing sufficient immune response through their antibody binding sites called epitopes [134]. Epitopes are recognised by the immune system and hence causes B-cells of the immune system to produce antibodies against the protein [135]. Epitopes that are formed by different parts of the polypeptide but are within spatial proximity of each other due to protein folding are called discontinuous epitopes while epitopes which are from a single stretch of the polypeptide are known as continuous or linear epitopes [135]. Since a desirable quality of a potential protein vaccine will be the possession of both types of epitopes, I sought to identify these in my protein dataset, however, bearing in mind that possession of an epitope does not completely explain what will happen *in vivo* but a very important step in identifying those that are most likely to make a good vaccine candidate.

I used four linear epitope prediction methods that make use of different propensity scales based on the physio-chemical properties of amino acids to assign them numerical values. Also, a sliding window rule is applied to determine the overall score of segments of the sequence at a time. A few groups used amino acid hydrophilicity to develop their propensity scale [136, 137], others used antigenicity [138], secondary structure [139], β-turn scale [140] as well as accessibility  to develop their propensity scales. The four prediction methods used here are the Bepipred, which is a combination of the Parker hydrophilicity scale and a hidden Markov model is the most sensitive amongst the tools used here, Parker hydrophilicity prediction, Chou and

Fasman Beta-Turn, and the Karplus and Schulz flexibility prediction method were also used [135, 136, 141, 142].

Additionally, due to the fact that only approximately 10% of epitopes are linear and that all the methods mentioned above are trained to detect linear epitopes, I went on to predict discontinuous epitopes in my protein dataset by using two defined tools called ElliPro and DiscoTope2 [134, 143]. DiscoTope is one of the first tools developed to predict discontinuous epitopes and it uses a combination of amino acid statistics, spatial information and surface accessibility to predict discontinuous epitopes [134]. Conversely, ElliPro  predicts epitopes using the concept of Thornton and colleagues [144], who showed correlation between regions protruding from a protein's globular structure and known continuous epitopes in three different proteins. While the Thornton method is based on two steps including, predicting the ellipsoid structure of the protein, and calculating residue protrusion index (PI) using the $\alpha$-C atom, ElliPro calculates PI using the residue's centre of mass and added another step where it uses residue PI to cluster neighbouring residues [143, 144]. Both these prediction methods use the protein structure to predict discontinuous epitopes. Accordingly, I searched the protein data bank (PDB) [145], which has over 130,000 experimentally verified protein structures using the amino acid sequences of my proteins. However, when there wasn't any appropriate structure (i.e. a structure with ≥50% amino acid identity with my protein) in PDB, I used *de novo* protein structure modelling tools I-TASSER (Iterative Threading ASSEmbly Refinement) [146-148] or the Phyre2 (Protein Homology/AnalogY Recognition Engine) server [149] to model my proteins with high confidence. Initially, Phyre2 was used to model the proteins and when it fails to produce a model with >90% confidence then I went on to use I-TASSER, which has been shown to perform better than all the modellers in all aspects of the critical assessment of protein structure prediction (CASP) [149].

## 2.17 Protein 3D structure

The idea behind using a protein model to predict discontinuous epitopes is due to the fact that despite low sequence identity (as low as 40%) of two proteins in the same protein family, their structures can still be similar. This is because protein structures are more conserved than their primary sequences in the family [150]. Nonetheless, the higher the sequence identity and the lesser the alignment gaps between two proteins the more likely their structural similarity [150]. Even though most models will produce similar results at ~40% sequence similarity, I chose a more stringent cut-off of 50% identity to improve the quality of my models [150].

## 2.18 Presence in other non-pneumococcal streptococci

An ideal protein vaccine will be a vaccine that clears the *S. pneumoniae* without affecting non-pneumococcal streptococci. In that regard, I searched both the NCBI non-redundant protein database and UniProt to investigate if the proteins in this dataset are also present in other streptococcus species or not.

## 2.19 Rank order

Finally, I developed a simple algorithm to rank my candidates by order of their potential as vaccine candidate. I used simple criteria to assign them scores and used their overall score to rank them. The criteria used here are:

1) Size of the protein, (i.e. bigger meaning better)
2) Level of Diversity (less alleles scored higher)
3) Proportion of protein predicted as immunogenic
4) Number of protein chains and,
5) Level of conservation (% presence in genomes)

First, the proteins with sequence lengths between 100-150 amino acids were scored 1 (the lowest), those between 151-200 were scored two and so on. Second, proteins with allele counts between 121-130 were scored 2, those between 111-120 were scored 4 and so on. Third, the proteins scored exactly as the percentage of the protein predicted to be an epitope by ElliPro. Fourth, they are scored double the number of chains they have. Finally, proteins are scored based on their level of conservation. 100% scored a maximum 10 points, 99.9 scored 9.9, and 94% scored 4 points. The total scores were used to rank the lipoproteins with the lipoprotein with the highest overall score ranked first.

# 3.0 Results

## 3.1 Serotyping

The capsule of the pneumococcus is its single most important virulence determinant and it is the basis for assigning serotypes. The *in-silico* serotyping performed in this study yielded 68 serotypes including all the serotypes in the presently licensed vaccines. A recently described serotype 35 variant assigned serotype 35D was also seen in this dataset [151].

In this study, serotype 1 and 5 were the leading causes of invasive disease for most years, however from 2011, serotype 12B/12F, which is not included in either PCV7 or PCV13 became a prominent cause of IPD. It was the commonest cause of IPD in 2011 and 2013 and the second and third commonest cause of IPD in 2014 and 2012 respectively. Serotype 1, which has been reported to be more common in disease than carriage [29] was seen 60 times in carriage and 132 times in disease making it the commonest cause of IPD. The second biggest contributor to IPD was serotype 5, which was isolated only once in carriage and 84 times in disease. One serotype 5 isolate was from an unknown source.

## 3.2 MLST and BAPS clustering

MLST was performed on all the isolates and BAPS clustering on a subset of these as part of a larger global collection. The BAPS cluster for the rest of the strains was inferred from their MLST results. 43 BAPS clusters, representing distinct lineages were observed in this dataset. Although it was seen here that serotype 1 has two major Sequence Types (STs), ST3081 and ST618 belonging to BAPS clusters 21 and 31 respectively, there were other less frequent STs such as ST303, 217, 10649, 3575 which belong to BAPS 21 as well as STs 2084, 3581, 3579, and 618 which belong to BAPS 31. ST618 was the most common ST until 2005 with most of the isolates appearing in disease. However, ST3081 first appeared in 2004 and although not seen in 2005 and 2006, it overtook ST618 as the most common serotype 1 ST in 2007. This trend continued until 2014 with ST618 last isolated in 2011 (Table 3.1). Also, serotype

1 was seen only 58 times in carriage, ST3081 was responsible for approximately 72% (42/58) of those with ST618 isolated only 14 times (24%) in carriage. The other STs that contributed to carriage were ST217 and ST303, contributing ~3% each. When serotype 1 isolates were divided into two groups based on their area of isolation with those isolated in the Western region and Fajara forming one group and those isolated from either Central river region or Upper river region forming another group, not a single ST618 was isolated in either Central or Upper river region of The Gambia from 2008-2014. Conversely, ST618 was last seen in the Western river region in 2011.

**Table 3.1 The distribution of serotype 1 lineages between 1996-2014.**
The columns represent the lineages and the rows represent the year of isolation.

| | ST3575 | 618 | 612 | 3579 | 3581 | 2084 | 217 | 3081 | 10649 | 303 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1996 | 1 | 6 | | | | | | | | |
| 1997 | | 2 | | | | | | | | |
| 1998 | | 1 | 1 | 1 | | | | | | |
| 1999 | | 1 | 1 | 1 | | | | | | |
| 2000 | | 1 | | | | | | | | |
| 2001 | | 1 | | | | | | | | |
| 2002 | | 7 | | | 2 | 2 | | | | |
| 2003 | | 5 | | | | 1 | 5 | | | |
| 2004 | | 1 | | | | | | 1 | | |
| 2005 | | 2 | | | | | | | | |
| 2006 | | | | | | | | | | |
| 2007 | | 14 | | | | | 2 | 22 | | 2 |
| 2008 | | 1 | | | | | | 23 | 1 | |
| 2009 | | | | | | | 2 | 16 | | |
| 2010 | | 6 | | | | | | 13 | | 1 |
| 2011 | | 3 | | | | | | 8 | | |
| 2012 | | | | | | | | 13 | | |
| 2013 | | | | | | | 1 | 7 | 1 | |
| 2014 | | | | | | | | 13 | | |

Furthermore, the only serotype 5 isolated in carriage was an ST289, BAPS20 lineage. This same ST was also responsible for about 28% (24/84) of serotype 5 IPD. The other serotype 5 STs that contributed to IPD were 3398, 3404 and 9935, responsible for approximately 18%, 52% and 1% respectively.

Interestingly, all the serotype 12B/12F isolates belong to ST989. The first appearance of this strain was in disease, in 2002. It was later isolated in carriage in 2007 and reappeared in disease in 2008. Although it increased in carriage in 2009 and 2010, it was from 2011 that it began to contribute significantly to invasive disease.

## 3.3 Conserved lipoprotein genes

The focus of the study is to identify pneumococcal lipoproteins, but most importantly, lipoproteins that are highly conserved across all serotypes. The lipoprotein pattern searches with the G+LPP, G+LPPv2 and the Prosite patterns produced 127, 136 and 167 results respectively, which together converged into 169 predicted lipoproteins. However, looking at their prevalence and choosing only those present in at least 90% of genomes, a total of 40 genes were selected for further analysis. These genes and their prevalence in the genomes screened are summarised in Figure 3.1. Additionally, those genes predicted to be lipoproteins by the Roary output were also included in the downstream screening tests. Together 55 putative lipoproteins were tested using the four tools mentioned above (SignalP, LipoP, Phobius and DOLOP) and only the proteins predicted to be lipoproteins in at least 3 of the four tools were selected for further analyses. These proteins are 30 in total as shown in Table 3.2.

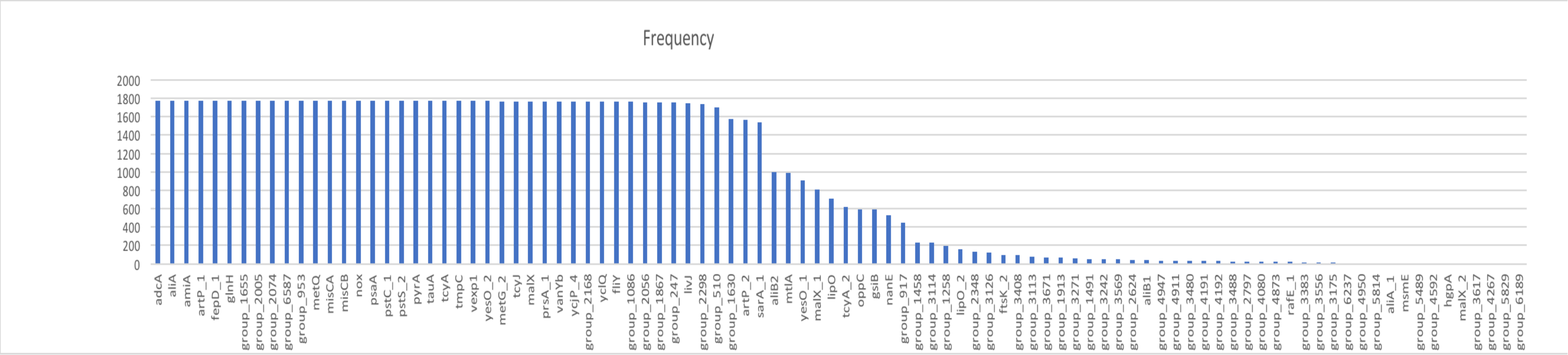


**Figure 3.1 This figure shows a subset of the lipoproteins from the pattern searches and their prevalence.**

These lipoproteins were arranged from left to right in order of decreasing prevalence. The x-axis has the gene names while the y-axis is their prevalence in the genomes screened.

## 3.4 Whole Genome Phylogeny

The whole genome phylogenetic tree was typical in that serotypes clustered together but also revealed potential serotype switching events, where a serotype is observed in multiple lineages on the tree such as serotype 6B(6E) (Fig. 3.2). The gene presence/absence information for the 30 candidate proteins were over-laid on this tree to reveal several interesting observations. Absence here means completely absent from the genome or truncated (less than 80% sequence length). First, iron transporter *pitA* was present in 1666 (94%) of genomes. It was mostly absent in two lineages that produce the same serotype, serotype 23B and genotype 23B+. It was also absent in 7/16 (~43%) of serotype 17F isolates, two of which were recovered in disease. Further, *pitA* was absent in 5/34 (~15%) of serotype 16F isolates and again 2 of these isolates were disease isolates recovered from adult patients. A serotype 19A disease isolate from a child <2years old also lacked *pitA* as well as 2 serotype-4 disease isolates from adults.

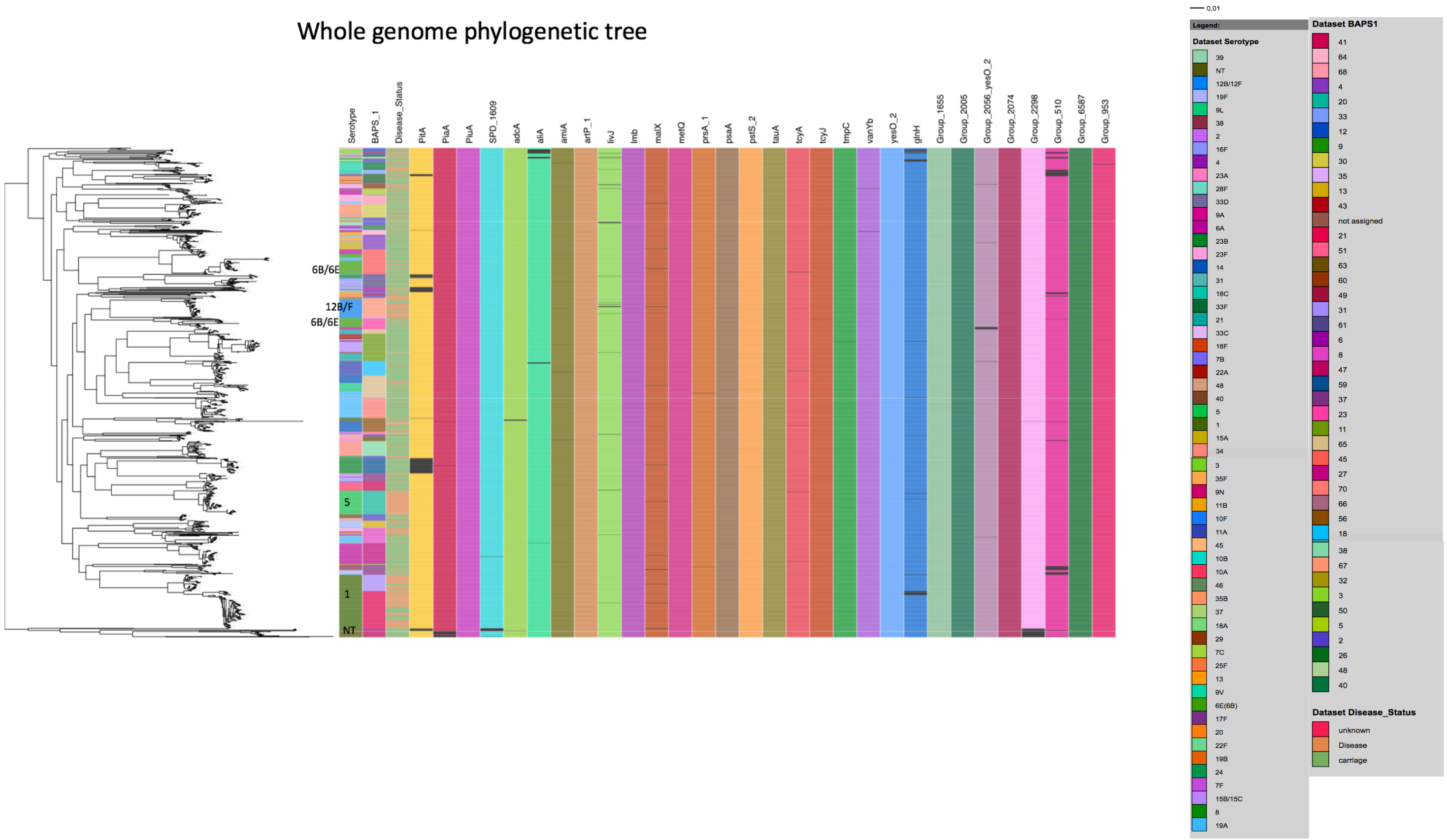


**Figure 3.2 Whole genome phylogenetic tree with candidate gene presence/absence profiles.**

The first three columns represent Serotypes, BAPS clusters and Disease status respectively. Black in the rest of the columns represents absence and any other colour represents presence.

The *piaA* gene, which encodes another iron transporter was found in ~98% of genomes covering all serotypes including some non-typables (NT). However, it was also missing in a subset of the NTs. All the genomes it was missing in belonged to BAPS cluster 47 (~95%) except 1 isolate, which belonged to BAPS cluster 56. This was also true for the recently identified iron transporter gene, *SPD_1609,* which was absent in 11 strains, all NTs within BAPS cluster 47. The overall prevalence of *SPD_1609* in the genomes was approximately 99%.

The zinc transporter lipoprotein encoding gene, *adcA* is also highly prevalent in the screened genomes with almost 100% of genomes possessing it. It is absent in only 5 NTs, 4 of which belonged to BAPS 56 and the remaining one to BAPS 47.

Further, *aliA* was present in approximately 99% of genomes and absent (truncated in this case) in 22 genomes, however, all the absences occurred in serotype 3. They occurred in serotype 3 within BAPS clusters 8, 12, 48, and 49. About 40% of the serotype 3 isolates were isolated from disease and all the BAPS clusters had representatives in this group.

Another gene absent in more than 20 genomes was the *livJ*. This gene is absent in both disease and carriage strains covering several serotypes (12B/12F, 1, 3, 6A, 9V, 23F, 23B1, 7F, 39, 10A and 22A). 10 of the 23 strains it was absent in were disease isolates. The disease isolates include 4/5 of serotype 12B/12F, which was the most prevalent serotype in the disease isolates, one serotype 9V, one of 6A, one 23F, one 7F and a serotype 3 strain.

The *amiA* gene was absent in only 3 strains, 1 serotype 9V belonged to BAPS 40 and isolated from disease and serotypes 11A and 23A both carriage strains and belonged to BAPS 18 and 63 respectively. *malX* was truncated or absent in 14 samples including both carriage and disease strains. The strains it was absent in include serotypes 9L (1), 12B/12F (1), 6B(6E) (1), 38 (2), 1 (2), 5 (1), 23F (1), 6A (1), 15A (1), 23B (1) and NT (2). The *tcyA* gene had an overall prevalence of approximately 100%. It was absent in only 5 genomes and these genomes belonged to serotypes, 14, 5, 6A, 11A, and 6B(6E). However, only the serotype 5 strain was a disease isolate.

*glnH* was found in all the isolates but was found to be truncated in as many as 38 isolates of which 20 were recovered from disease. These diseased strains include serotypes 1 (60%), 3 (25%), 5 (10%) and 19A (5%).

Further, *Group_2056* genes were absent in a total of 13 strains and all these strains were carriage strains. Serotype 6A strains accounted for about 70% of absences. One each of serotypes 23F, 11A, 19A, and 15A were also lacking this gene. *Group_2298* was absent in 31 samples, all of which were NTs. BAPS cluster 47 was represented ~94% and one strain each of BAPS 56 and 2 also lacked the gene.
*Group_510* was one of the less prevalent genes in this data set as it was absent in 71 genomes. 21 (~30%) of these genomes were found in disease. The disease isolates include several serotypes including 3, 38, 25F, 22F, 18C, 18A, 17F, 23A, and 6B. Group_953 is absent in only one strain belonging to serotype 9V and recovered in carriage.

*prsA_1* was also highly prevalent and was found to be absent in only 2 strains belonging to serotype 38 BAPS 37 and serotype 19A BAPS 65. Both strains were carriage strains. Gene *tcyJ* was almost 100% prevalent but it was absent in a single serotype 6B, BAPS 23 strain which was recovered from disease. Similarly, *tmpC* was also absent in only one serotype 7F carriage isolate belonging to BAPS cluster 11.
Also, *vanYb* was absent in only 2 serotype 6A strains, both recovered from carriage.

Twelve of the genes in this study were found in all the isolates, these genes include *piuA, psaA, artP_1, lmb, metQ, pstS_2, tauA, yesO_2, Group_1655, Group_2005, Group_2074*, and *Group_6587*.

## 3.5 Gene Trees and annotation

Using MUSCLE aligned nucleotide sequences, the gene trees were built using RAxML [131]. The number of SNP sites used to build each tree is summarised in Table 3.2. The trees are not rooted as I am only interested in their relationship to each other. All the gene trees were subsequently annotated with serotype information, BAPS cluster, disease status as well as the protein alleles. Prior to assigning alleles, the nucleotide sequences were translated into protein sequences to exclude the effect of synonymous mutations. The amino acid length and number of alleles found for each protein is summarised in Table 3.3.

**Table 3.2. This table summarises the number of taxa and SNP sites used to reconstruct each gene tree.**

| Gene | No. of taxa | No. of SNP sites |
|---|---|---|
| *Group_1655* | 1769 | 55 |
| *Group_2005* | 1769 | 150 |
| *Group_2056* | 1756 | 75 |
| *Group_2074* | 1769 | 34 |
| *Group_2298* | 1738 | 32 |
| *Group_510* | 1697 | 41 |
| *Group_6587* | 1769 | 96 |
| *Group_953* | 1768 | 74 |
| *adcA* | 1764 | 190 |
| *aliA* | 1747 | 362 |
| *amiA* | 1766 | 60 |
| *artP_1* | 1769 | 48 |
| *glnH* | 1731 | 102 |
| *livJ* | 1746 | 134 |
| *lmb* | 1769 | 81 |
| *malX* | 1755 | 55 |
| *metQ* | 1769 | 68 |
| *piaA* | 1747 | 38 |
| *piuA* | 1769 | 60 |
| *pitA* | 1666 | 33 |
| *prsA* | 1767 | 51 |
| *psaA* | 1769 | 67 |
| *pstS_2* | 1769 | 45 |
| *SPD_1609* | 1758 | 199 |
| *tauA* | 1769 | 67 |
| *tcyA* | 1764 | 95 |
| *tcyJ* | 1768 | 145 |
| *tmpC* | 1768 | 84 |
| *vanYb* | 1767 | 256 |
| yesO_2 | 1769 | 68 |

In the phylogenetic maximum likelihood trees shown in Fig. 3.3 through to Fig. 3.32, specific serotypes, BAPS cluster and/or allele have been annotated only in special cases. The prefixes S-, B- and A- used in the annotations denote serotype, BAPS cluster and allele respectively.

**Table 3.3 Summary of protein length and number of allele.**

| Protein | AA length | No. of Alleles |
|---|---|---|
| Group_1655 | 165 | 36 |
| Group_2005 | 503 | 26 |
| Group_2056 | 445 | 38 |
| Group_2074 | 188 | 11 |
| Group_2298 | 185 | 26 |
| Group_510 | 164 | 39 |
| Group_6587 | 268 | 31 |
| Group_953 | 292 | 28 |
| AdcA | 501 | 82 |
| AliA | 662 | 126 |
| AmiA | 660 | 37 |
| artP_1 | 278 | 38 |
| GlnH | 275 | 67 |
| LivJ | 386 | 37 |
| Lmb | 305 | 23 |
| MalX | 423 | 51 |
| MetQ | 284 | 28 |
| PiaA | 342 | 31 |
| PiuA | 322 | 60 |
| PitA | 122 | 25 |
| PrsA | 316 | 24 |
| PsaA | 309 | 17 |
| PstS_2 | 291 | 23 |
| SPD_1609 | 357 | 101 |
| TauA | 335 | 15 |
| TcyA | 278 | 40 |
| TcyJ | 266 | 48 |
| TmpC | 350 | 21 |
| VanYb | 238 | 53 |
| YesO_2 | 442 | 19 |

There were 38 SNP sites used for building the phylogeny of the *piaA* gene Fig. 3.3. Briefly, very few clustering by serotype can be seen from this tree. Even though this protein has 31 alleles, it is clear from the figure that one allele (assigned number 12 here) is the dominant allele covering almost every lineage and serotype. However, serotype 19A, BAPS 70 strains seem to have a unique allele, 19. Serotype 19A, BAPS 8 strains have allele 6. Serotype 6A, BAPS 27 strains have both the dominant allele 12 and a few other strains possessed allele 11.

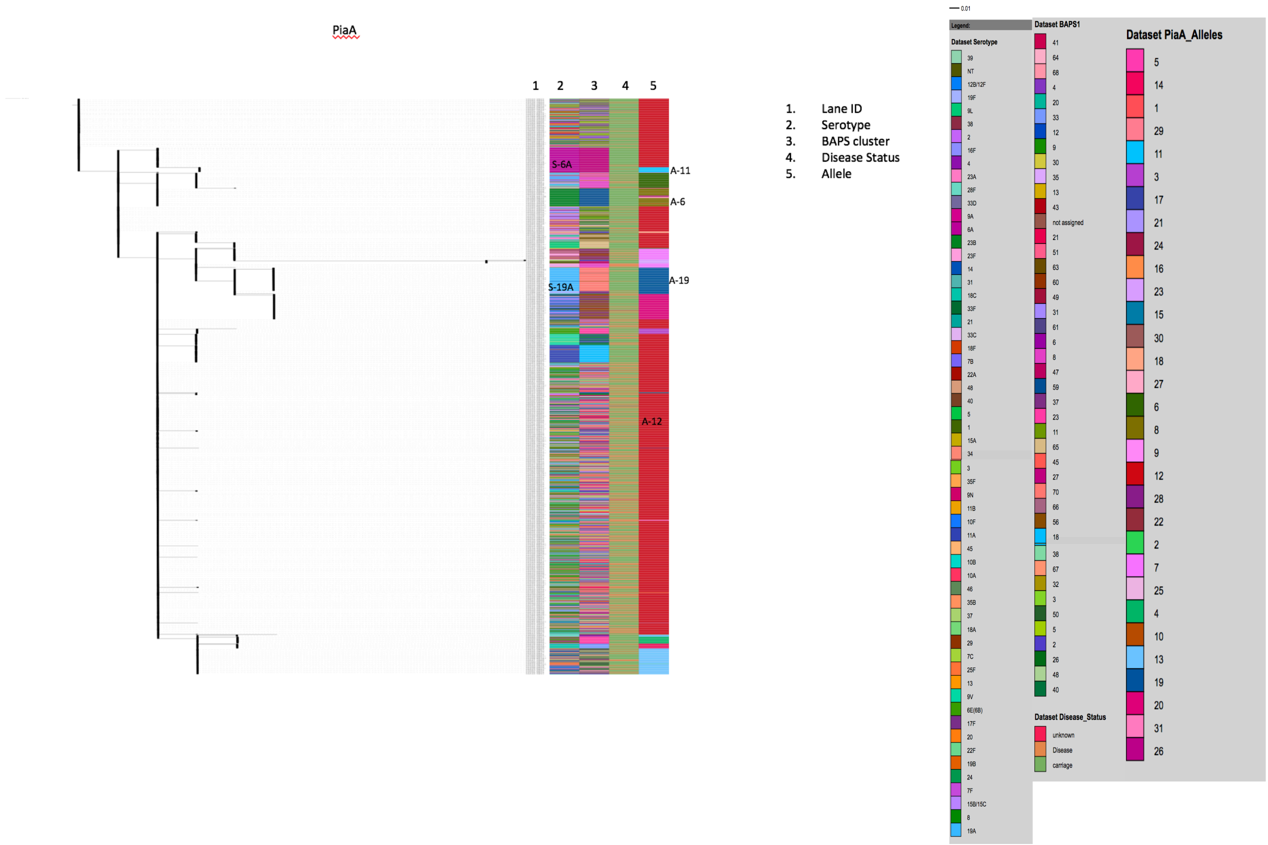


**Figure 3.3 Phylogenetic gene tree of *piaA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**38 SNP sites used to reconstruct the phylogeny of the *piaA* gene**

The *piuA* gene tree shows some clustering by lineage (Fig. 3.4). Allele 22 is the most prevalent, covering several serotypes and lineages including serotype 1, 13, 19A and 19F. Allele 19 has a strong association with disease, with almost 100% of isolates with this allele recovered in disease and these strains also belong to serotype 5, BAPS 20. Further, serotype 1 BAPS 31 strains also possess a unique allele, 52.

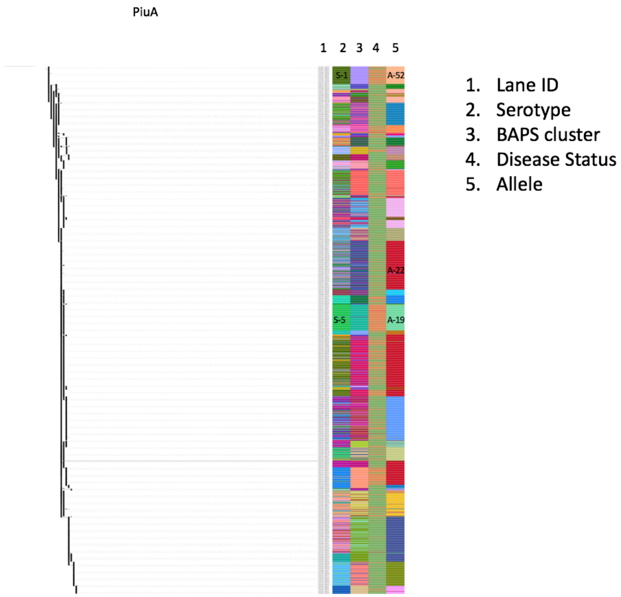
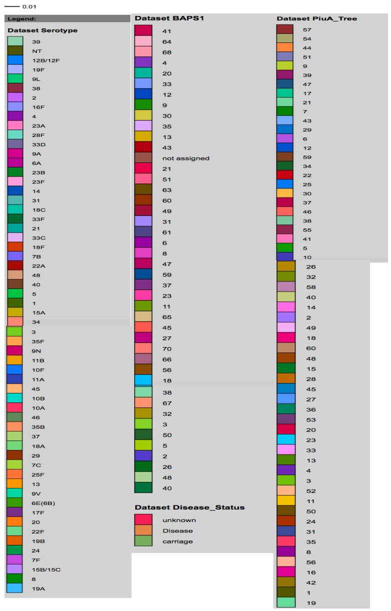


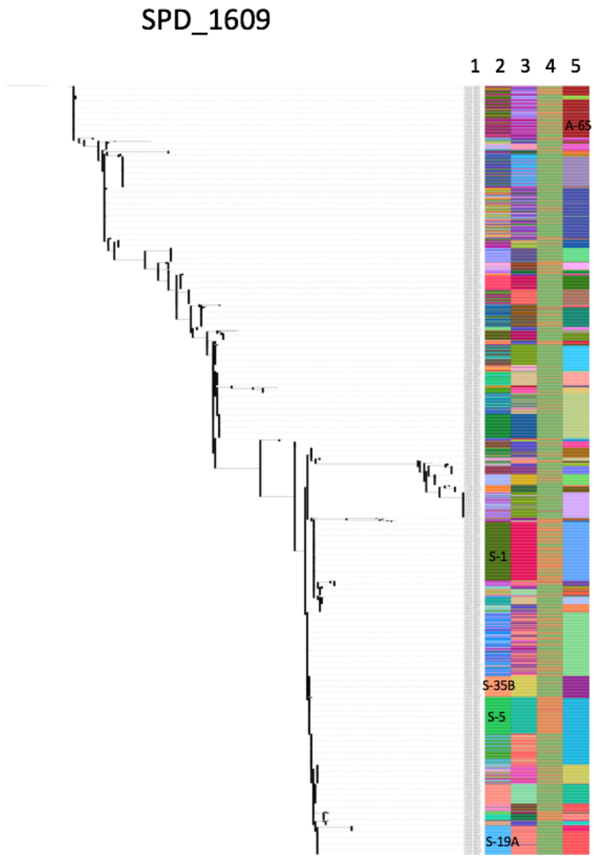
**Figure 3.4 Phylogenetic gene tree of *piuA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.



**60 SNP sites used for the construction of this phylogenetic tree.**

There is quite some clustering by lineage going on with the *SPD_1609* gene tree as illustrated in Fig. 3.5. Serotype 1 and BAPS 31 strains clustered with other serotypes including 6 and 11B and most of them had allele 65 although a few have allele 4. The rest of the serotype 1 strains belonging to BAPS 21 clustered together and had a unique allele (42) to them. Other clustering by serotype include serotypes 19A, 5, and 35B.



1. Lane ID
2. Serotype
3. BAPS cluster
4. Disease Status
5. Allele

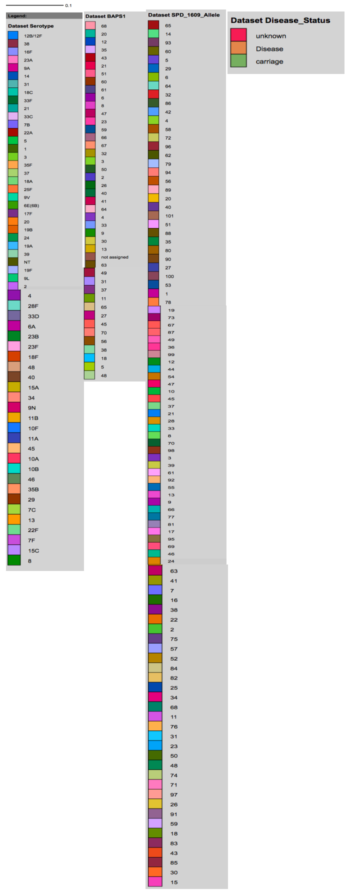


**Figure 3.5 Phylogenetic gene tree of *SPD_1609*.**

This shows the nucleotide relationship of the genes extracted from the genomes.

**199 SNP sites were used to reconstruct this gene tree**

*pitA* encodes an iron transporter lipoprotein, which had the shortest sequence (122AA) of the iron transporters. Consistently, it also had the smallest number of alleles with only 25 alleles. The phylogenetic tree has less clustering by serotype and it has few major alleles that cover most serotypes across all lineages (Fig 3.6). A few serotype 19A, BAPS 70 strains have a unique allele.

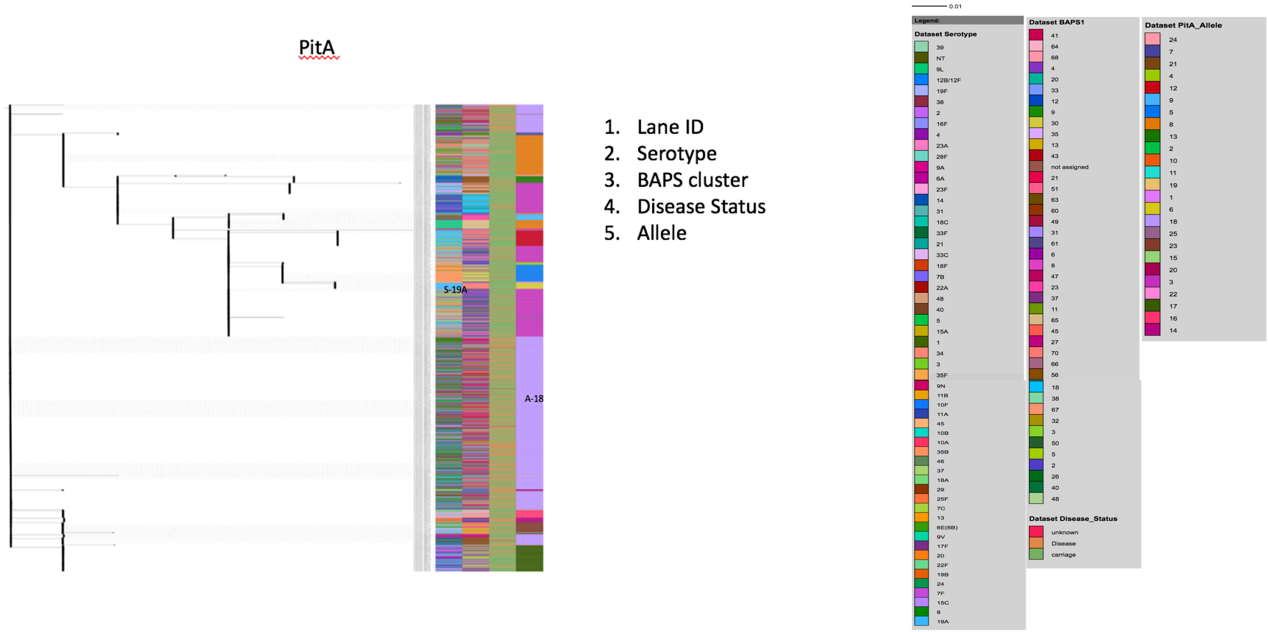


**Figure 3.6 Phylogenetic gene tree of *pitA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**33 SNP sites were used to reconstruct this gene tree.**

*psaA* encodes a manganese transporter lipoprotein, PsaA, which had one of the fewest number of alleles (17). Also, there was only one dominant allele, 1 (Fig. 3.7). This allele was present in >90% of the genomes and the only serotype that had a unique allele was serotype 35B, BAPS 30, which has allele 4.

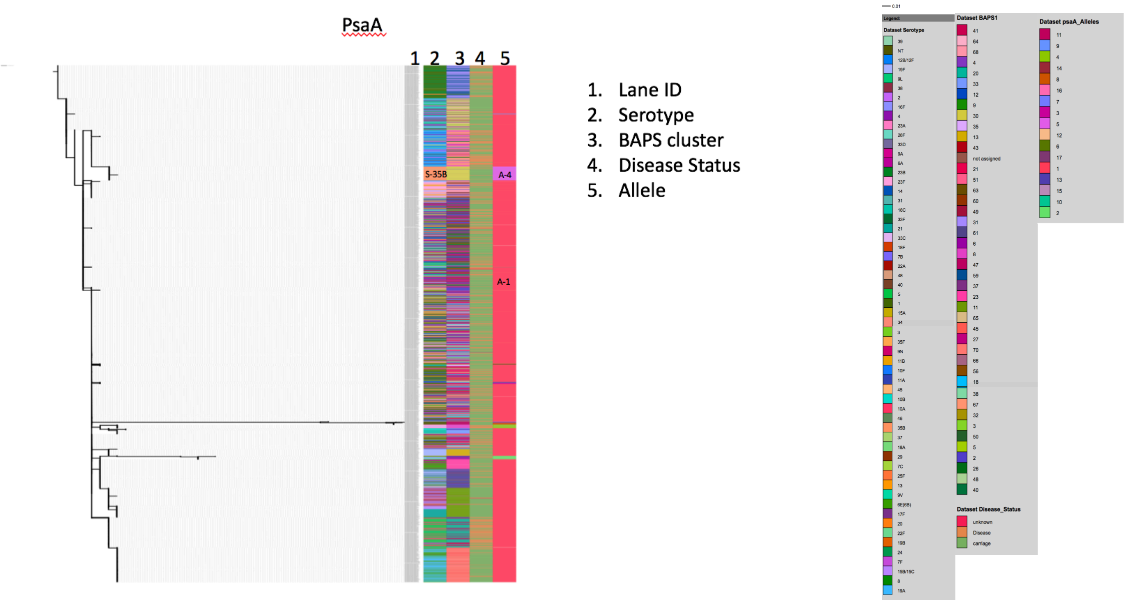


**Figure 3.7 Phylogenetic gene tree of psaA.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**67 SNP sites were used to reconstruct this gene tree.**

*adcA* encodes a zinc transporter lipoprotein. Some lineages had unique protein alleles to them (Fig. 3.8). The two serotype-1 lineages (21 and 31) clustered separately. Genes from lineage 31 clustered with other serotypes including 5, 14, 19A and 35B. Lineage 21 proteins clustered together and had the same allele with only serotype 25F proteins. Furthermore, a subset of serotype 5, BAPS 20 proteins which were all recovered from disease also clustered together and had a unique allele (2).

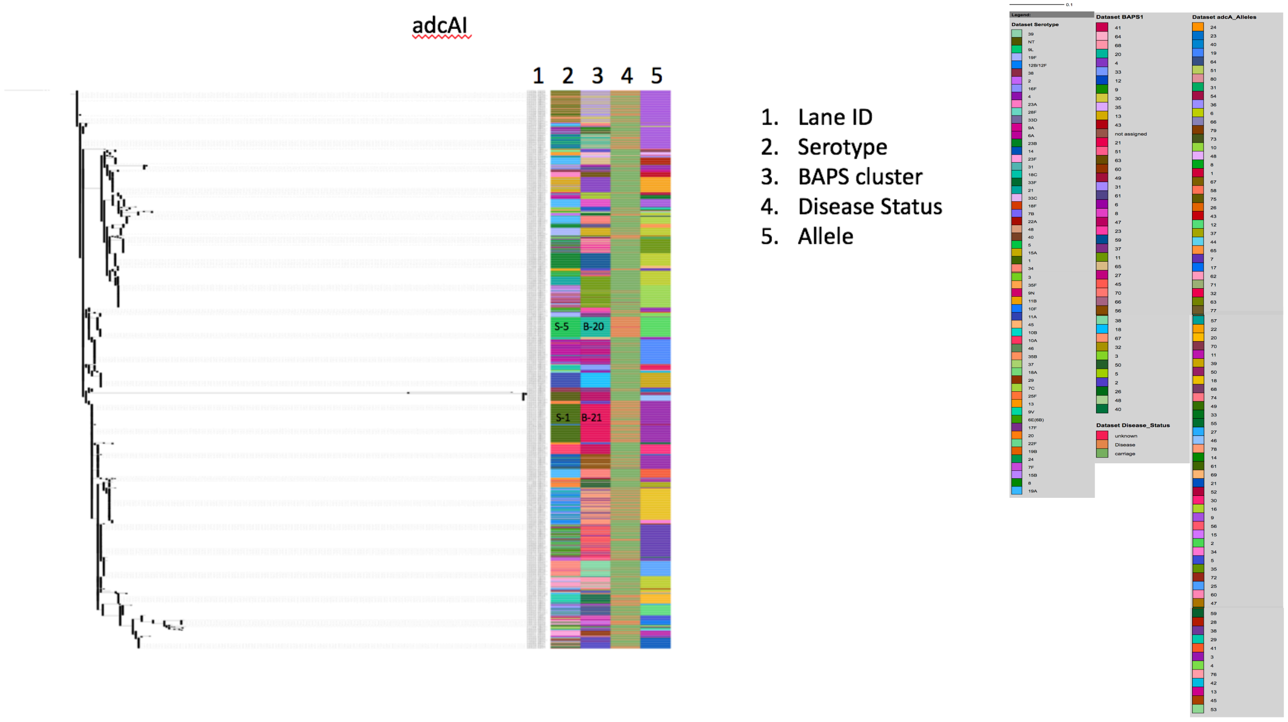
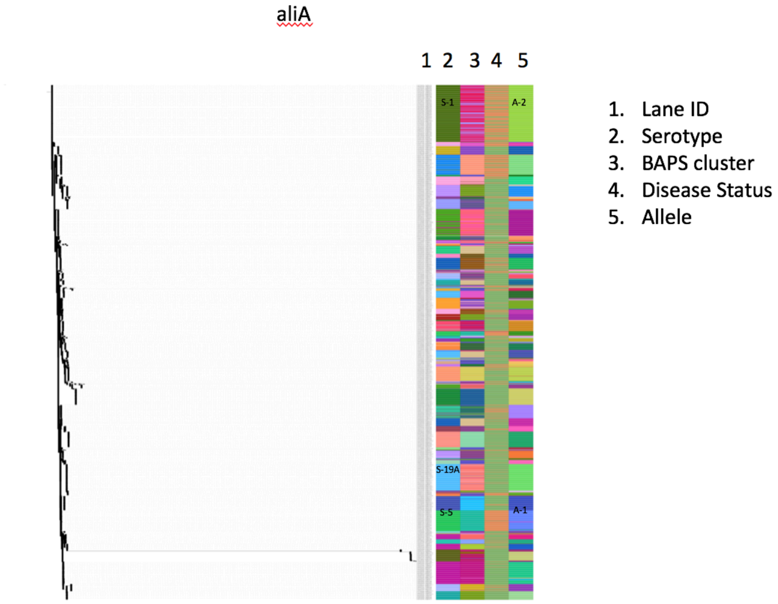


**Figure 3.8 Phylogenetic gene tree of *adcA*.**

**190 SNP sites used to reconstruct this gene tree.**

This gene tree shows the nucleotide relationship of the genes extracted from the genomes.

AliA is a big protein with a sequence length of 662AA. The gene tree shows a lot of clusters with many alleles unique to the serotype from which the protein was obtained from (Fig 3.9). Some of these clusters include serotype 1 protein genes (both lineages) having allele 2 unique to them, serotype 19A (BAPS 45, 70, 37, 68 and 65) having allele 4 and some serotype 5 BAPS 20 isolates having allele 1.



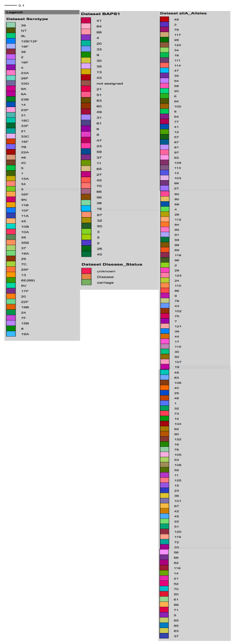


**Figure 3.9 Phylogenetic gene tree of *aliA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**362 SNP sites used to reconstruct this gene tree.**

AmiA, is a protein approximately the same size as AliA in terms of sequence length. It had two (2 & 3) main alleles which cover almost every serotype and several randomly occurring alleles across the tree. Only serotype 1 BAPS 31 and serotype 23F seem to have unique alleles (Fig. 3.10).

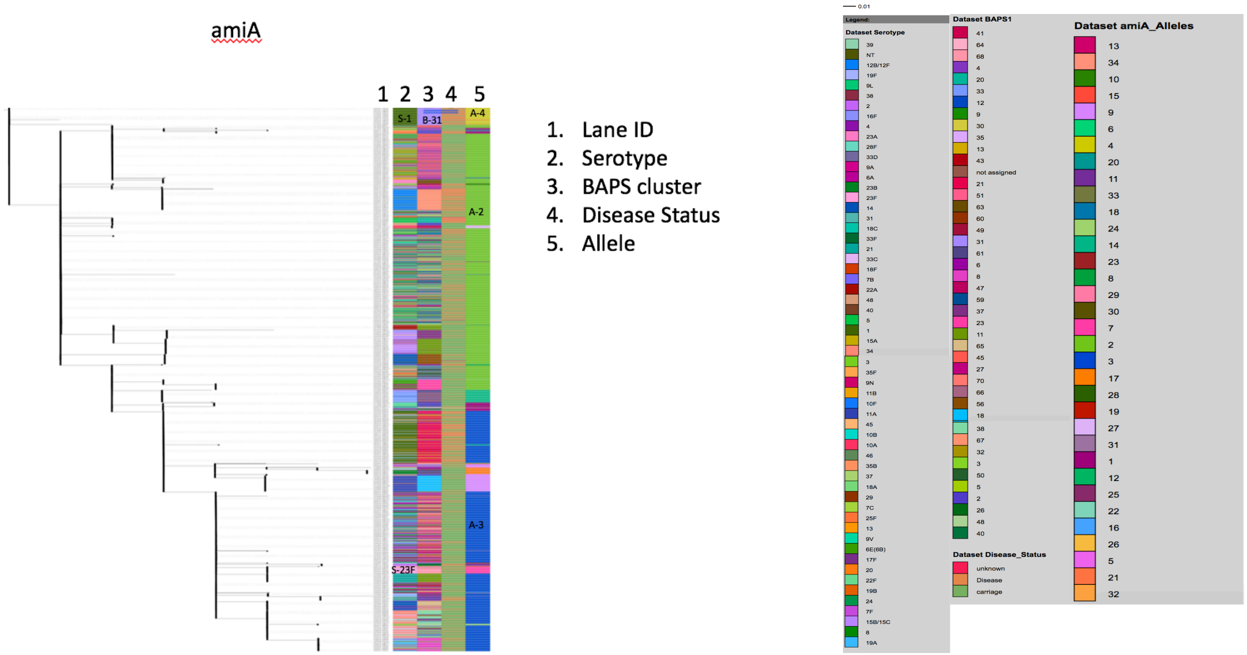
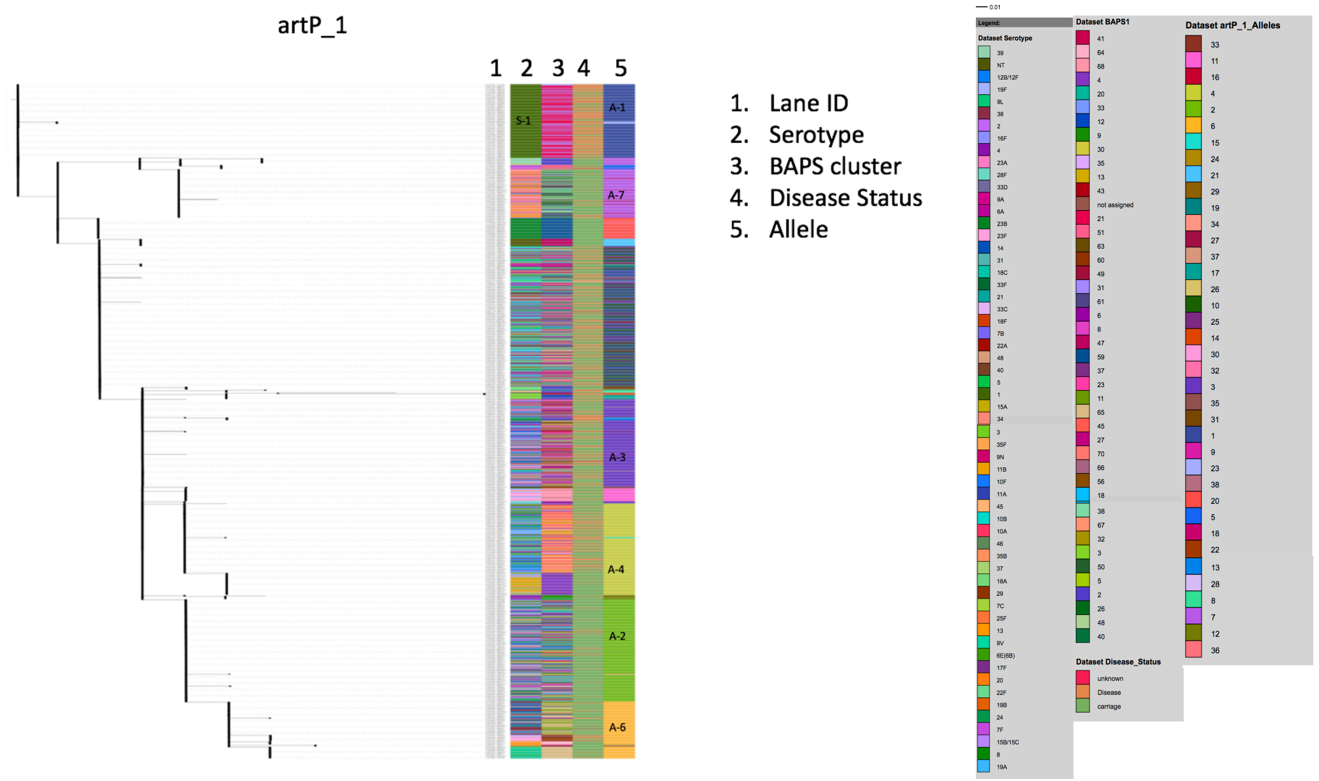


**Figure 3.10 Phylogenetic gene tree of *amiA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**60 SNP sites were used to reconstruct this gene tree.**

The *artP_1* phylogenetic tree shows less clustering by serotype except for serotype 1s. It had 38 alleles; however, a few alleles represent almost all the lineages (Fig 3.11). Allele 1 covers both lineages of serotype 1 as well as several other lineages representing many serotypes such as 19A, 5, 6A, 38, 35B, 25F etc. Alleles 2, 3, 4, 6 and 7 are also major alleles representing several lineages.



**48 SNP sites used to reconstruct this gene tree.**

**Figure 3.11 Phylogenetic gene tree of *artP_1*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

The GlnH protein is a 275AA with relatively many alleles (67) (Fig 3.12). Although there were alleles covering several lineages, there was also quite a few clustering by lineage in this protein. The two lineages of serotype 1 clustered separately, with each cluster having a unique allele. The same is true for serotype 5 BAPS 20 and serotype 19A BAPS 70 strains too. A similar observation is true for serotype 10A, 23B and some NTs.

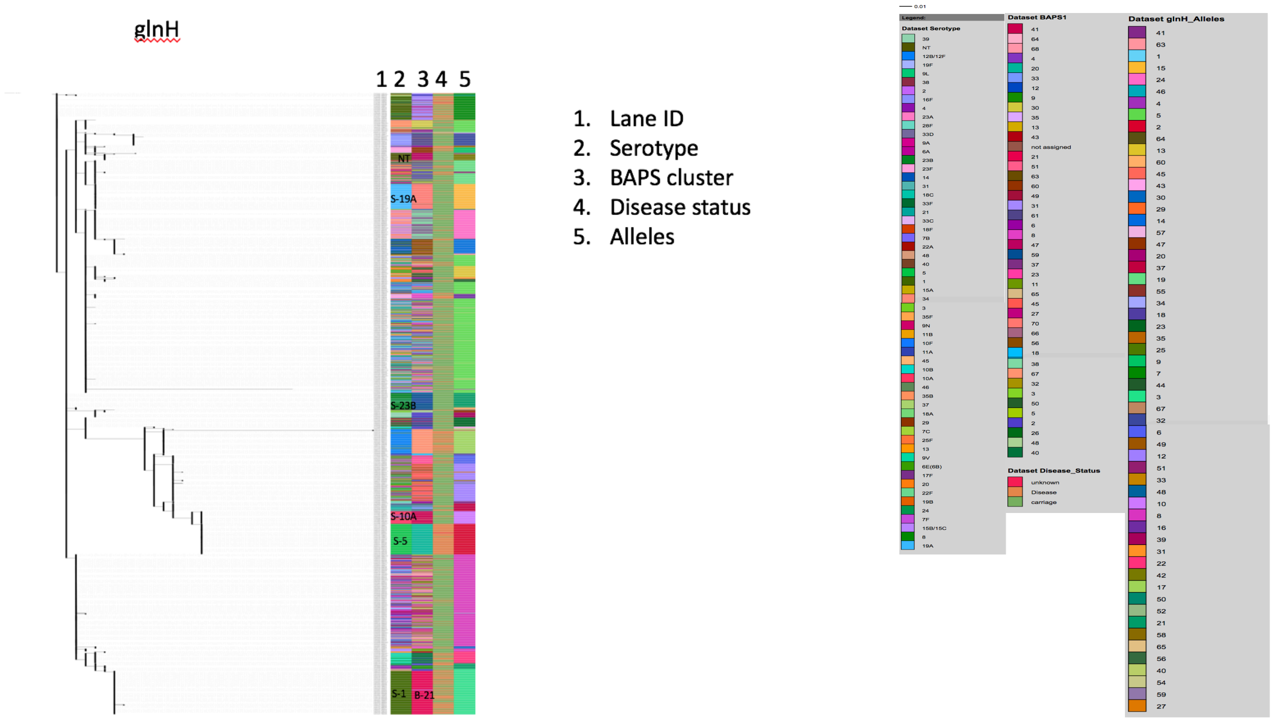


**Figure 3.12 Phylogenetic gene tree of *glnH*.**

This gene tree shows the nucleotide relationship of the genes extracted from the genomes.

**102 SNP sites were used to reconstruct this gene tree.**

*Group_510* encodes a short lipoprotein, which had 39 alleles. From the tree, it is clear that two alleles (1 & 2) are more dominant covering almost every lineage of every serotype (Fig. 3.13). However, BAPS 5 (serotype 6A and a few 15A), BAPS 63 (23A), and BAPS 37 (15B/C and 13) clustered together with a unique allele.

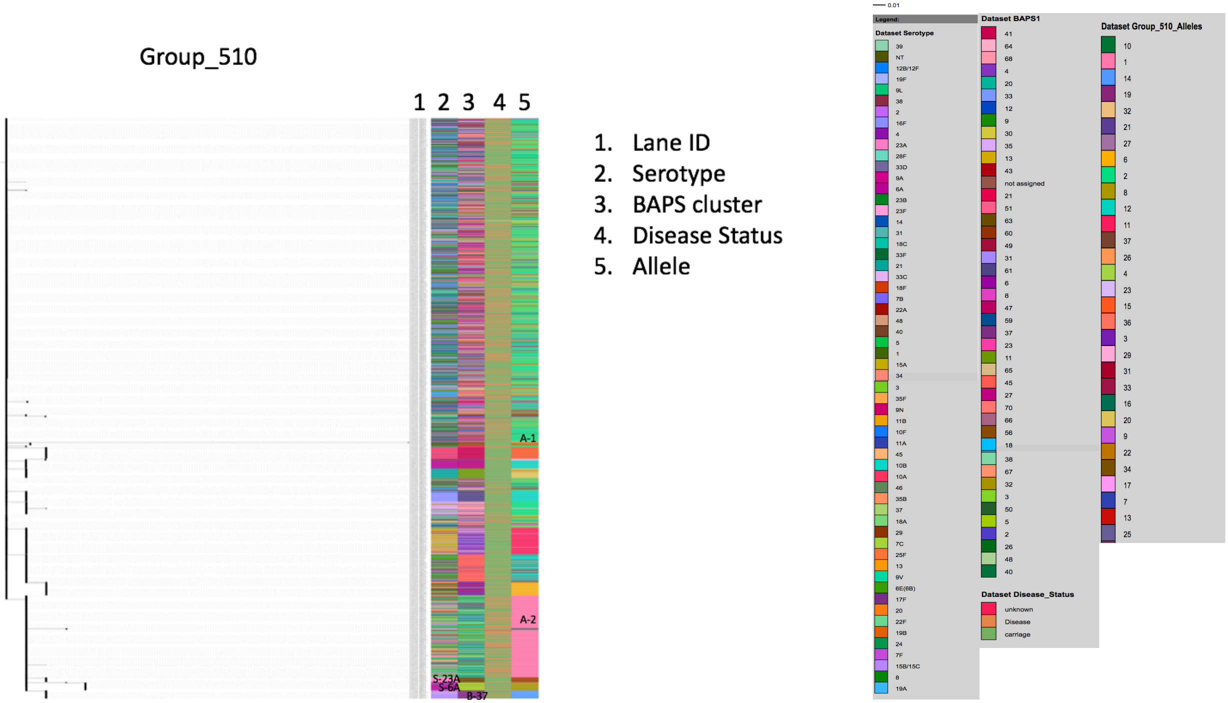


**Figure 3.13 Phylogenetic gene tree of *Group_510*.** This gene tree shows the nucleotide relationship of the genes extracted from the genomes. **41 SNP sites used to reconstruct this gene tree**.

Group_1655 lipoproteins also have a short (165AA) sequence length. The most prominent alleles from the gene tree are alleles 2 and 4. Additionally, allele 11, which covers BAPS 67 of serotype 12B/12F is also important as this group includes many disease strains (Fig. 3.14).

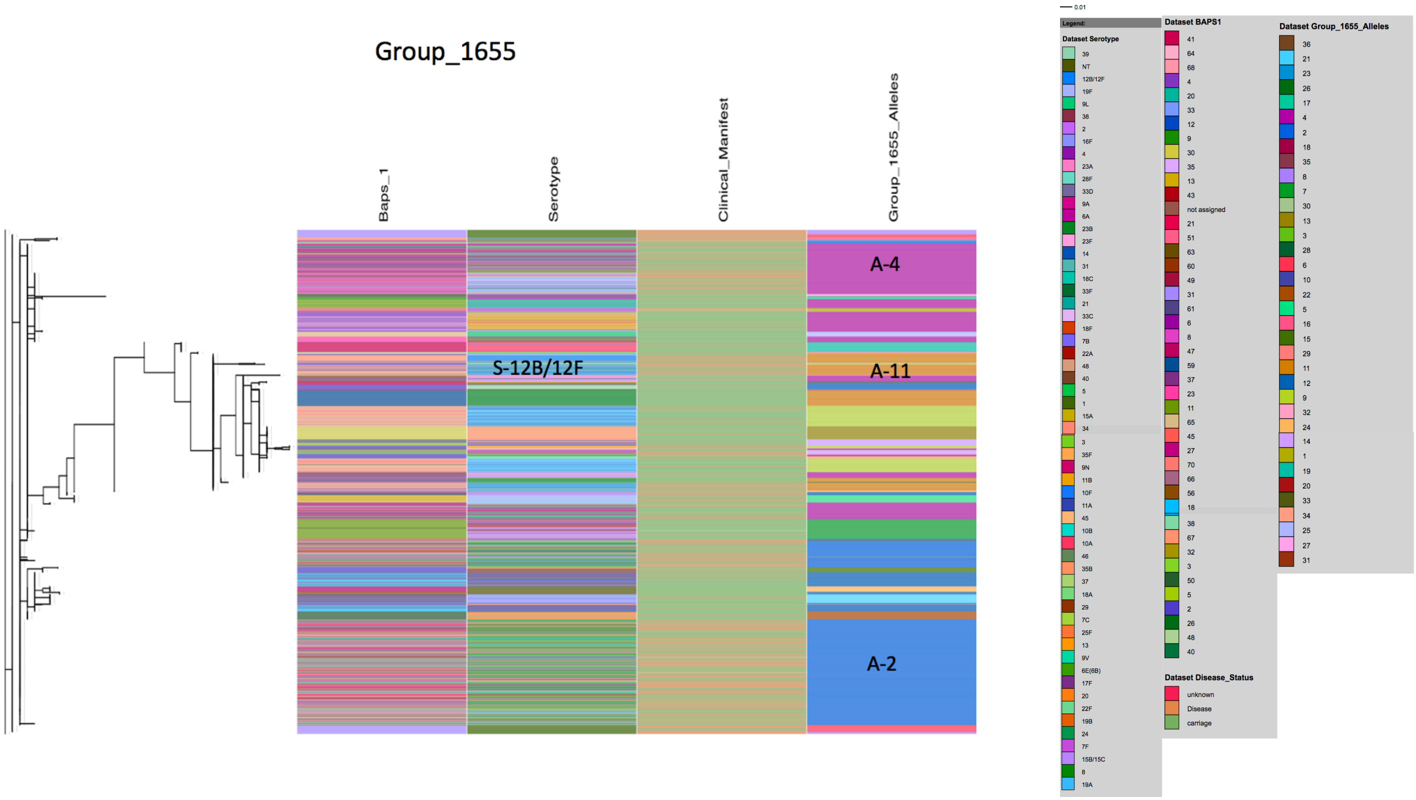


**Figure 3.14 Phylogenetic gene tree of *Group_1655*.**

This gene tree shows the nucleotide relationship of the genes extracted from the genomes.

**55 SNP sites used to reconstruct this gene tree**

Group_953 lipoproteins have longer amino acid sequences than both Group_510 and Group_1655 lipoproteins (Table 3.3) and had 28 alleles. More than 50% of the isolates belonged to allele 1, which includes almost all lineages. Lineages with unique alleles were seen only twice. BAPS 70 serotype 19A and BAPS 2 serotype 3 (Fig. 3.15). The latter group has only carriage strains.

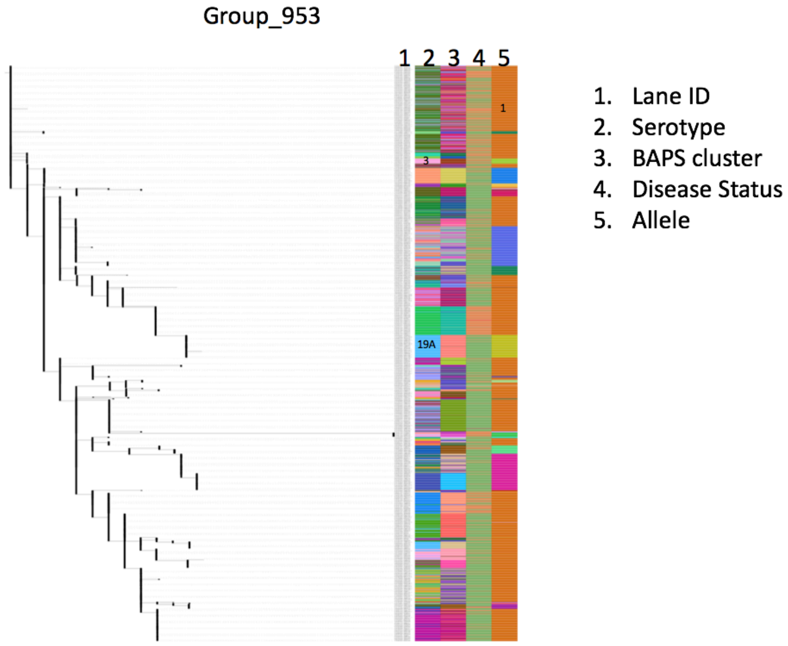
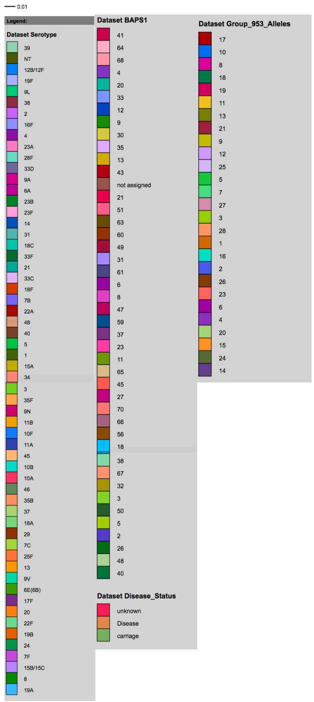


**Figure 3.15 Phylogenetic gene tree of Group_953.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.



**74 SNP sites used to reconstruct this gene tree.**

Analysis of Group_2005 lipoproteins revealed 26 alleles (Fig. 3.16). However, allele 1 represented approximately 90% of genomes including all the major lineages and serotypes. This allele also included almost all the disease isolates. The next allele with the most members was allele 11, which consists of BAPS 18 serotype 11A and BAPS 37 serotype 19F strains which were all carriage strains.

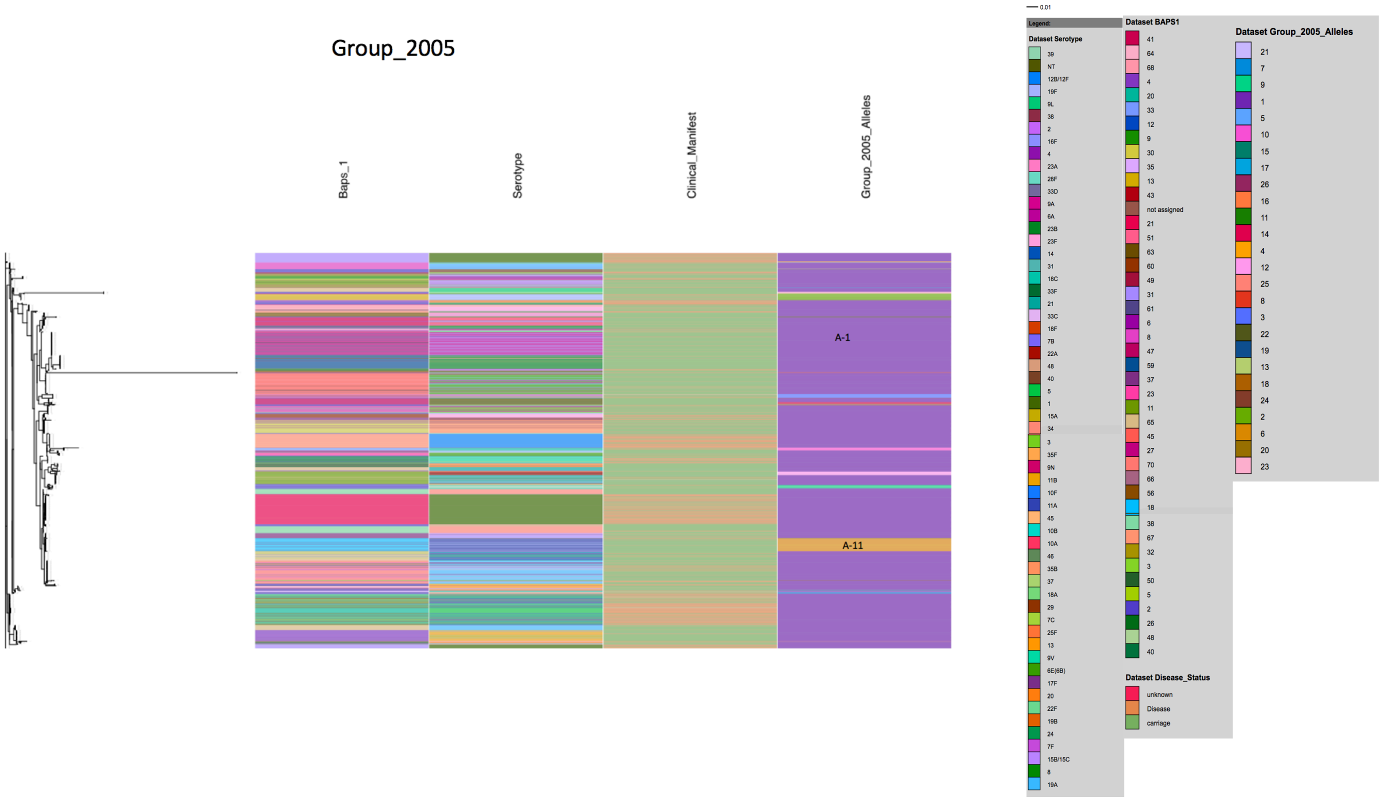


**Figure 3.16 Phylogenetic gene tree of Group_2005.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**150 SNP sites used to reconstruct this gene tree.**

*Group_2056* encodes a long 445AA lipoprotein, which had 38 alleles (Fig. 3.17). Similar to Group_2005, allele 1 represented all the major lineages and serotypes. Allele 2 also had a few important lineages including BAPS 8 serotypes 23F and 19F strains as well as BAPS 67 serotype 46 strains, some of which were disease strains.

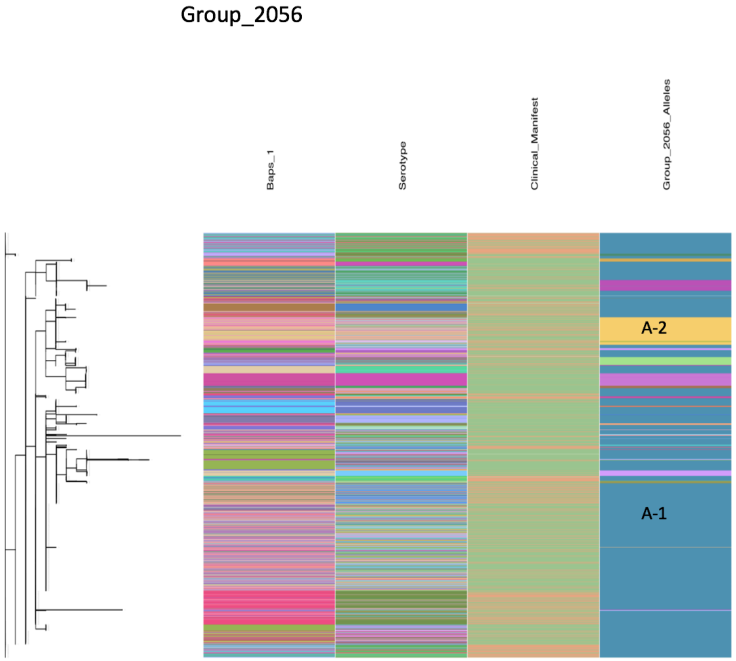
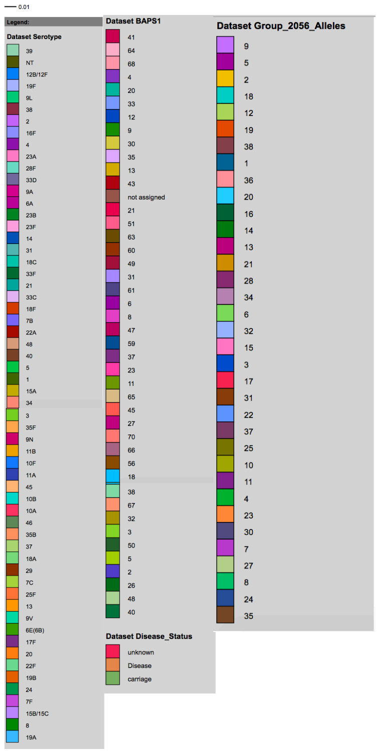


**Figure 3.17 Phylogenetic gene tree of *Group_2056*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.



**75 SNP sites used to reconstruct this gene tree**

Further, *Group_2074* encode lipoproteins with short sequence lengths (188AA) and hence a relatively small number of alleles (11). Alleles 1 and 2 were the only major alleles and together represented >90% of isolates. The NTs, mostly belonging to BAPS 47 and a few BAPS 57s were clustered together and had 2 alleles (allele 3 and 6) unique to them (Fig 3.18). Although in the minority, BAPS 2 of serotype 11B also had a unique allele and this group included some disease strains.

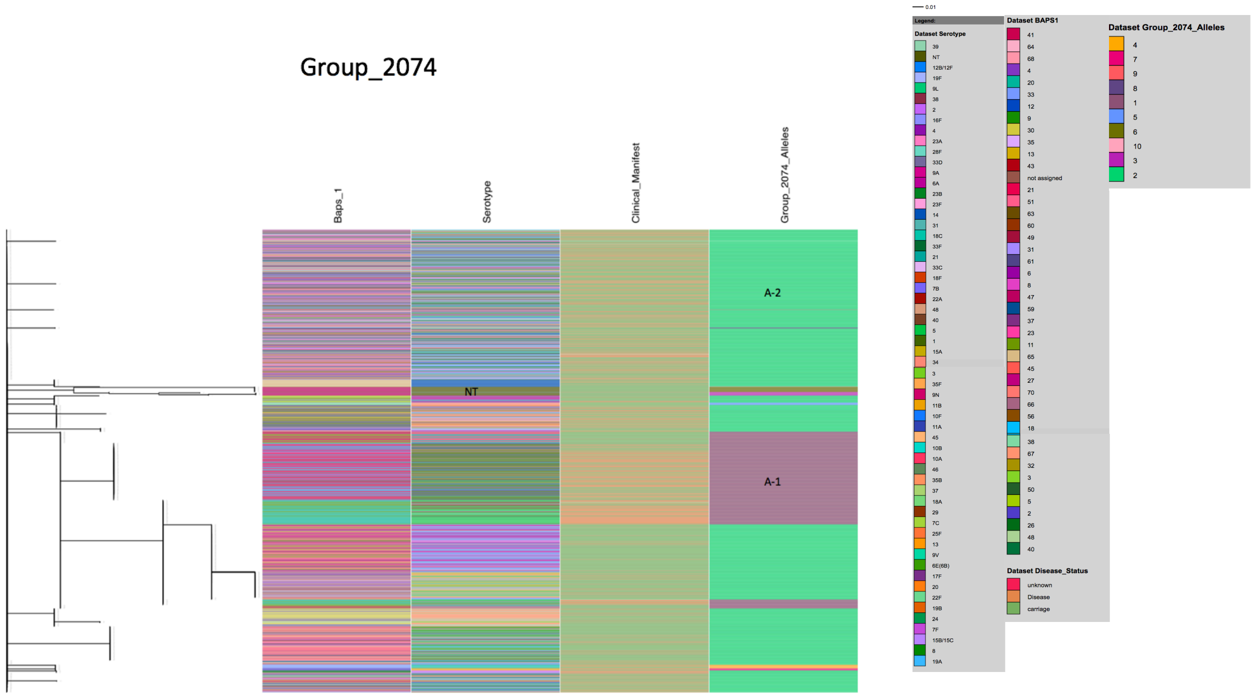
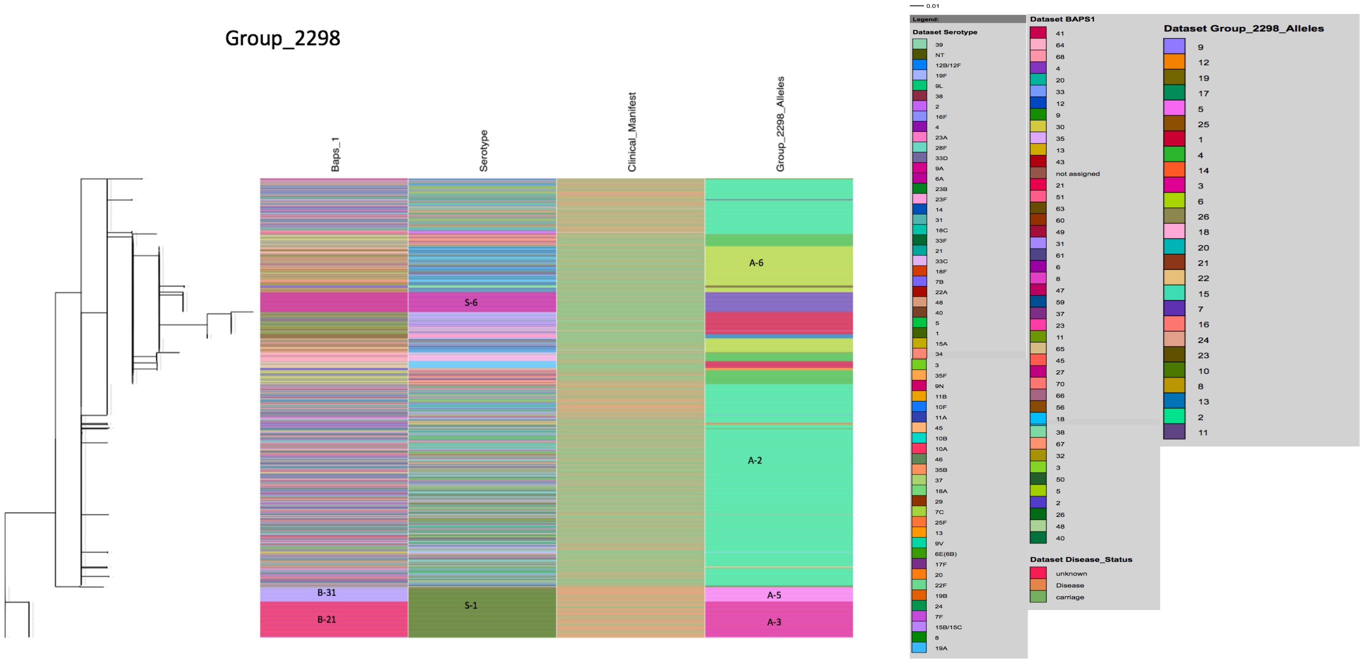


**Figure 3.18 Phylogenetic gene tree of Group_2074.**

The tree shows the nucleotide relationship of the genes extracted from the genomes.

**34 SNP sites used for the reconstruction of this gene tree.**

Although Group_2298 lipoproteins had a similar amino acid length to Group_2074 proteins (Table 3.3), their allele count of 26 was higher (Fig. 3.19). Most of the serotype 1 strains clustered together. Further, both lineage 31 and 21 had unique alleles (allele 5 and 3 respectively). Allele 2 was the most prevalent and it covered several lineages and allele 6 also had broad coverage. Other alleles that were specific to certain lineages include allele 7, 13, and 23 which were unique to 6A BAPS 27, 23A BAPS 63 and 18A BAPS 2 respectively.



**32 SNP sites used for the reconstruction of this gene tree**

**Figure 3.19 Phylogenetic gene tree of *Group_2298*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

31 alleles were found in Group_6587 proteins. Allele 1 and 4 were the dominant alleles covering many lineages including disease strains. Allele 21 was found in only serotype 23A BAPS 63 strains. Allele 3 represented both lineages of serotype 1 (21 &31) but was also present in serotype 19F BAPS 13 as well as BAPS 2 of serotype 40 and BAPS 5 of 6A, 9A and 9V (Fig. 3.20).

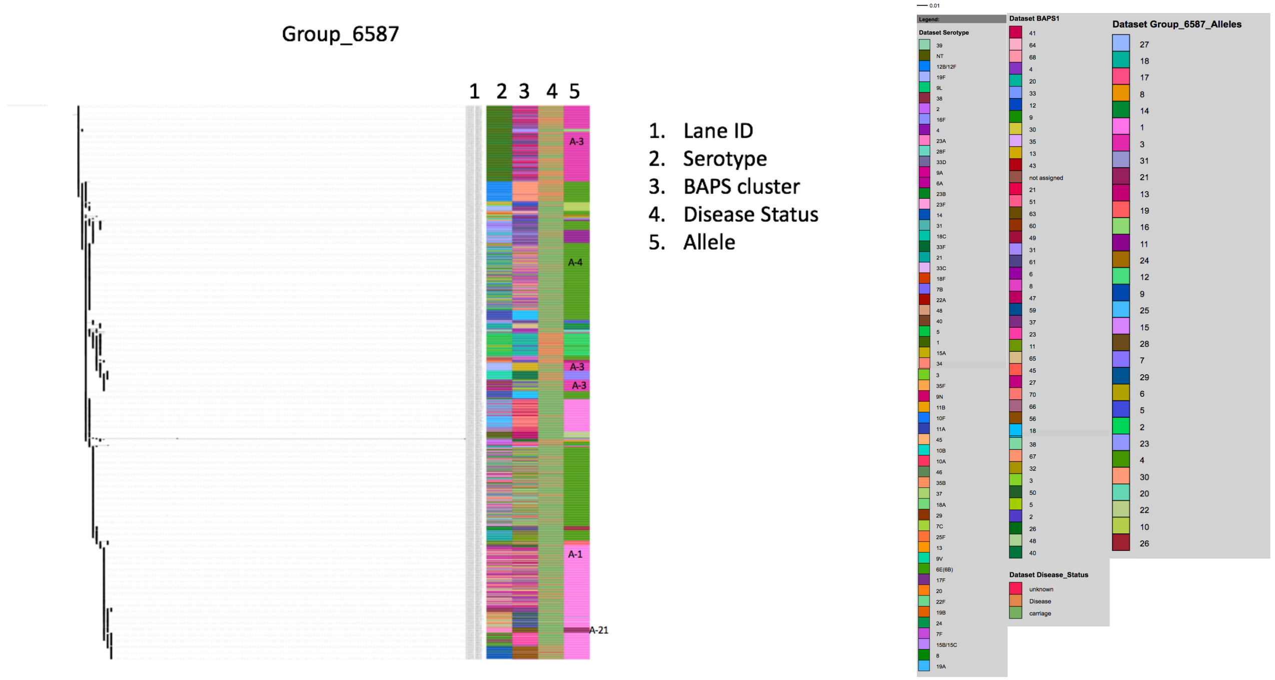


**Figure 3.20 Phylogenetic gene tree of *Group_6587*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**96 SNP sites used for the reconstruction of this gene tree.**

LivJ had 37 alleles in this study. Most of the major alleles including alleles 1, 2 and 4 covered several lineages and serotypes, however, allele 3 was confined to serotype 1 isolates, representing both lineages (BAPS 21 & 31). Allele 9 was also found in only BAPS 67 strains, which included serotypes 46 and 12B/12F strains (Fig. 3.21).

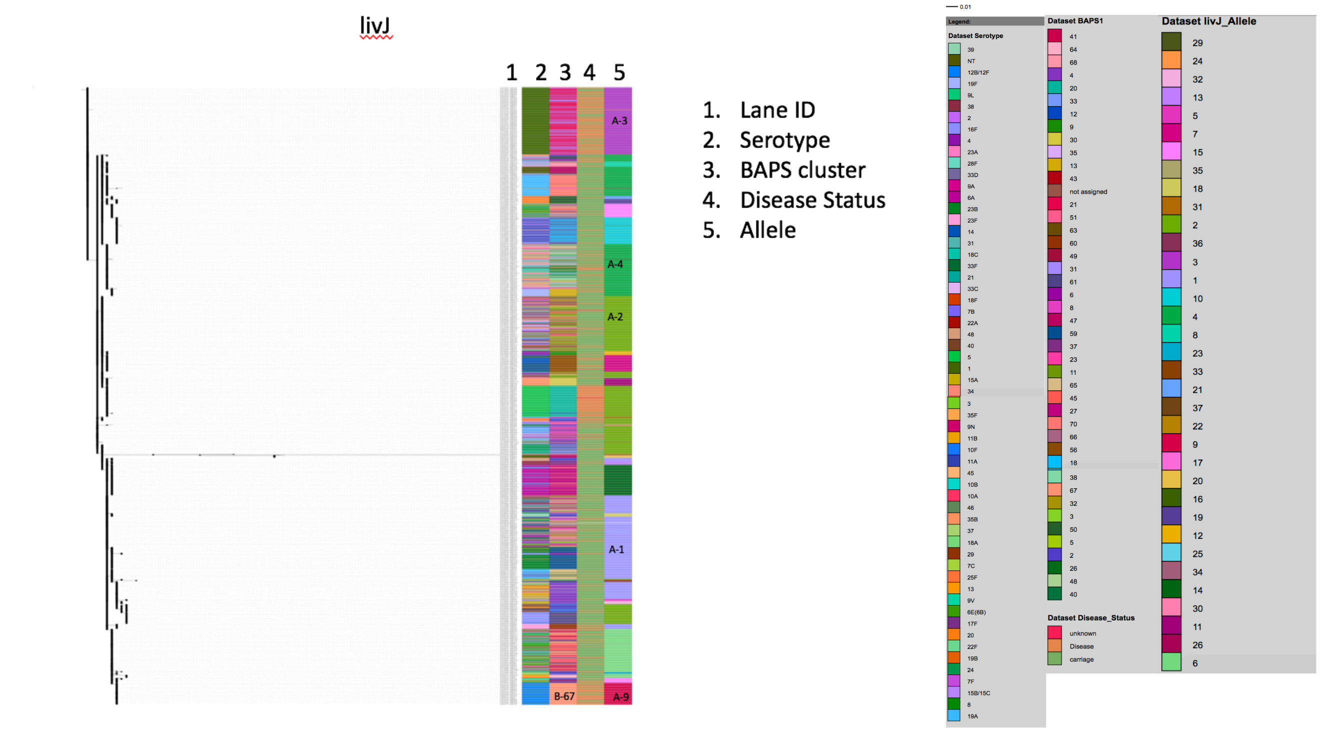


**Figure 3.21 Phylogenetic gene tree of *livJ*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**134 SNP sites used to reconstruct this gene tree.**

MalX is a 423AA lipoprotein and had 51 alleles. Despite the high number of alleles, only a few alleles were more prevalent. These alleles include 1, 7, 10 and 12. No evidence of an allele being present in only one lineage was apparent.

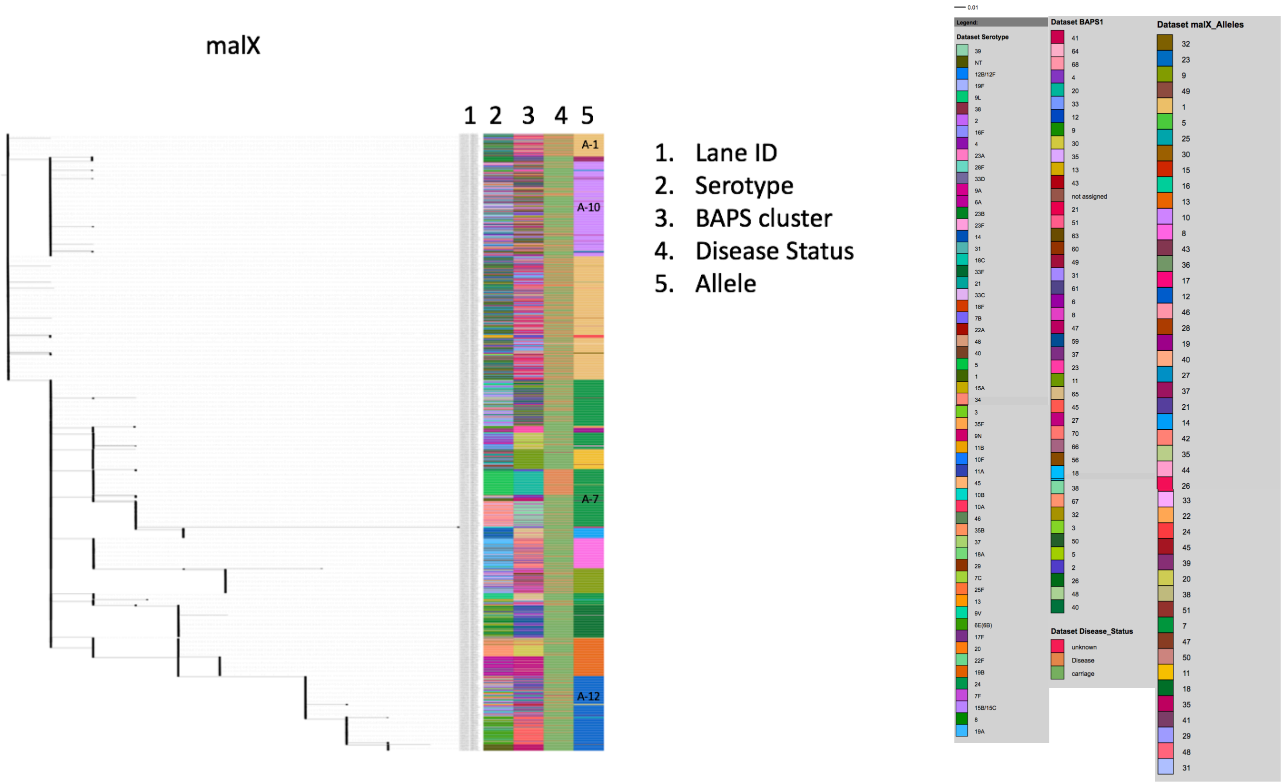


**Figure 3.22 Phylogenetic gene tree of *malX*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**55 SNP sites used to reconstruct this gene tree.**

The *lmb* gene encodes a 305AA long lipoprotein, which had 23 alleles. Phylogenetic analysis showed serotype clustering of only serotype 1s and 5s. Allele 1 was the most prevalent allele and represented many lineages including BAPS 67 of serotype 12B/12F, which consisted of many disease isolates. Further, alleles 2 and 3 were very important as they covered the highly virulent serotype 5 and 1 lineages respectively.

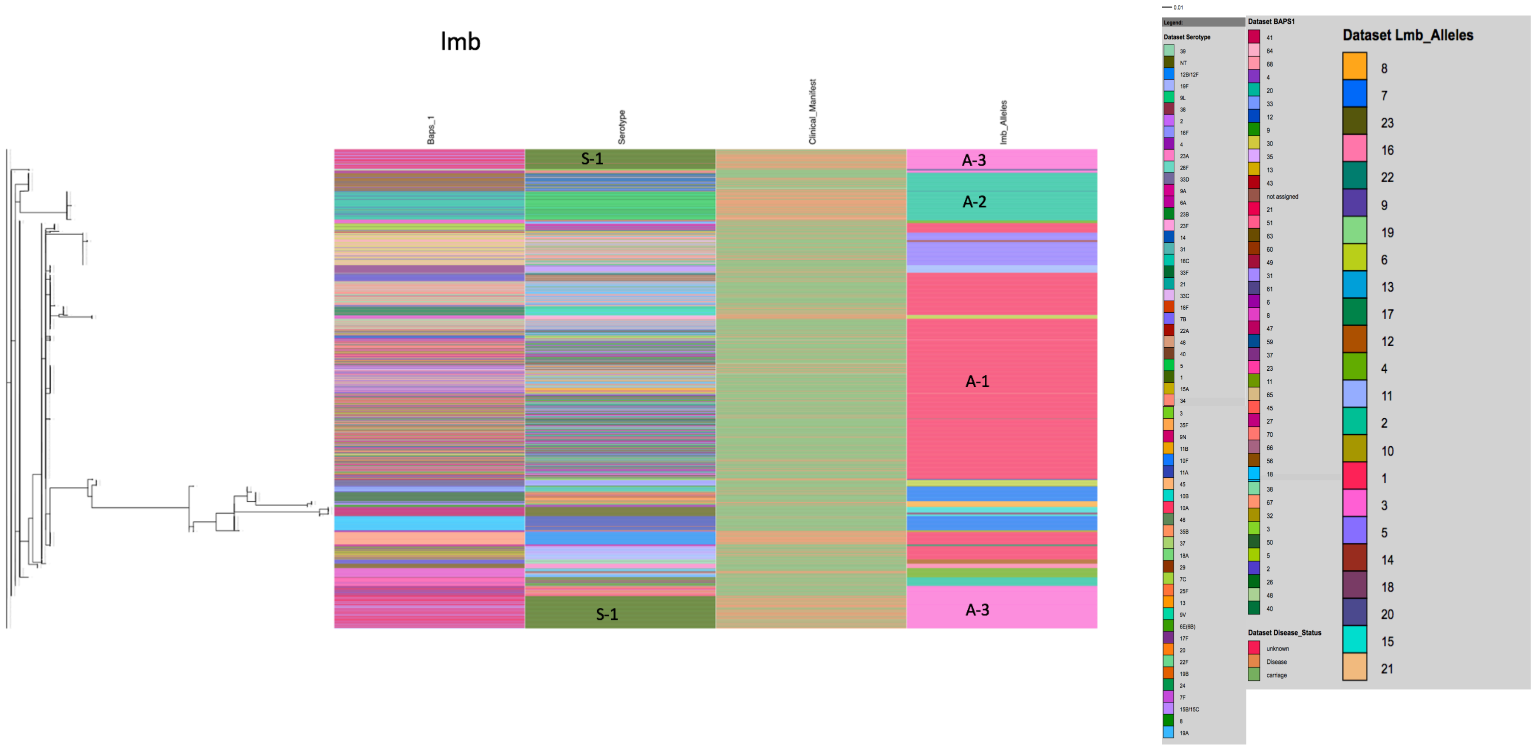


**Figure 3.23 Phylogenetic gene tree of *lmb*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**81 SNP sites used to reconstruct this gene tree.**

The MetQ lipoprotein had 28 alleles. Allele 2 was clearly the most prevalent allele, but there were other alleles covering important lineages including disease strains. These include allele 5 and 7 as well as allele 1, which covered both lineages of serotype 1 and a few serotype-4 BAPS 9 and serotype 23F BAPS 60 & 68 strains.

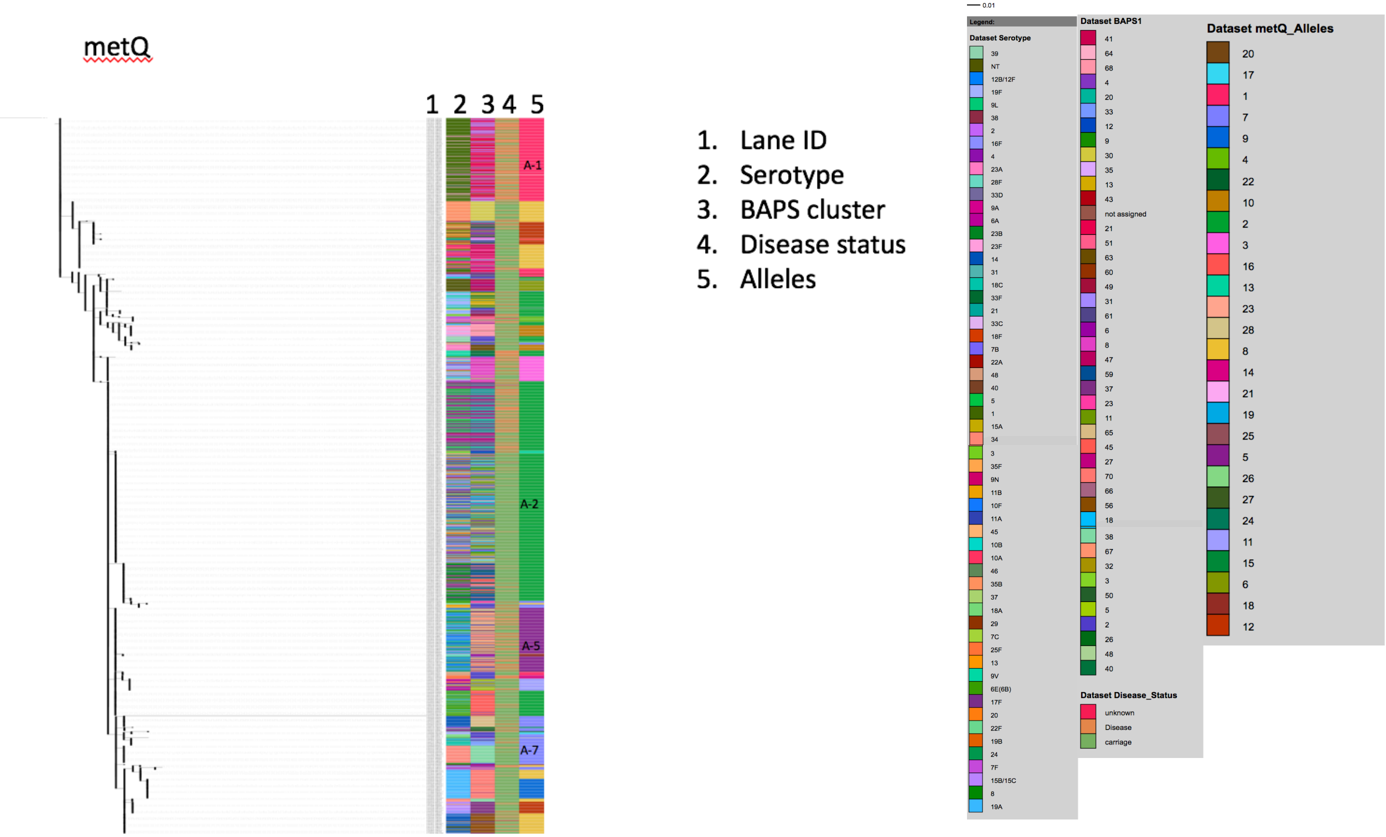


**Figure 3.24 Phylogenetic gene tree of *metQ*.**

**68 SNP sites used to reconstruct this gene tree.**

The PstS_2 lipoproteins are typically 291AA long and had 23 alleles here. From the phylogenetic gene tree analysis, it was clear that allele 1 was the predominant allele present in more than half the isolates. A few BAPS 21 isolates clustered with BAPS 31 strains as well as BAPS 18 (serotype 11A and 20), BAPS 2 (19F) and BAPS 56 (14 and NTs) strains, all having allele 4. However, most BAPS 21 strains clustered away from these and had a unique allele, 2.

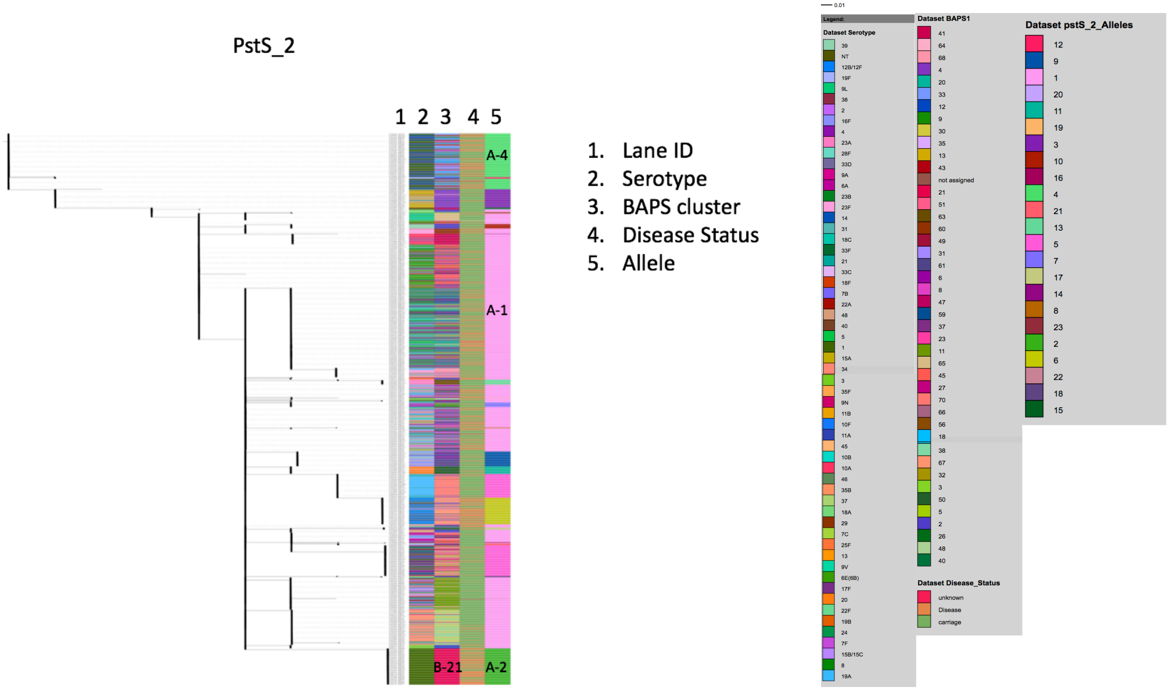
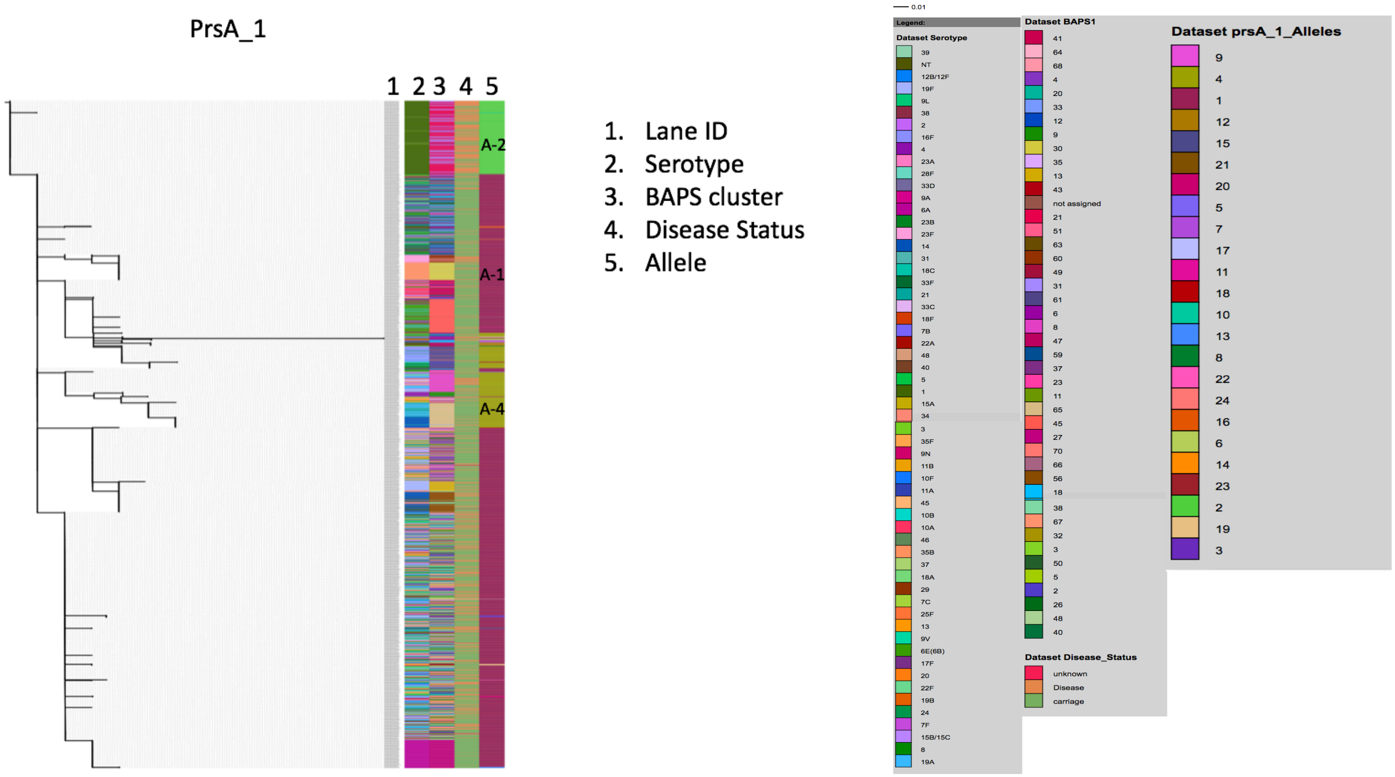


**Figure 3.25 Phylogenetic gene tree of *pstS_2*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**45 SNP sites used to reconstruct this gene tree.**

PrsA is a 316AA protein with 24 alleles seen. Similar to some of the proteins already seen, one allele (allele 1) was present in more than half the genomes. The next two alleles prevalent in this protein were alleles 4 and 2. Allele 4 was present in more lineages than allele 2 but allele 2 was the prevalent allele in both lineages of serotype 1 except a few strains which possessed allele 10.



**51 SNP sites used to reconstruct this gene tree**

**Figure 3.26 Phylogenetic gene tree *prsA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

The *tcyA* gene encodes a 278AA protein with 40 alleles here. Allele 1 was predominant, present in many lineages including the serotype 1 lineages. Another allele also prevalent in several lineages was allele 6. Allele 2 was present in almost all serotype 5 BAPS 20 strains. Furthermore, it was present in approximately all strains of serotype 14, some BAPS 37 serotype 19F strains and also BAPS 56 of NTs.

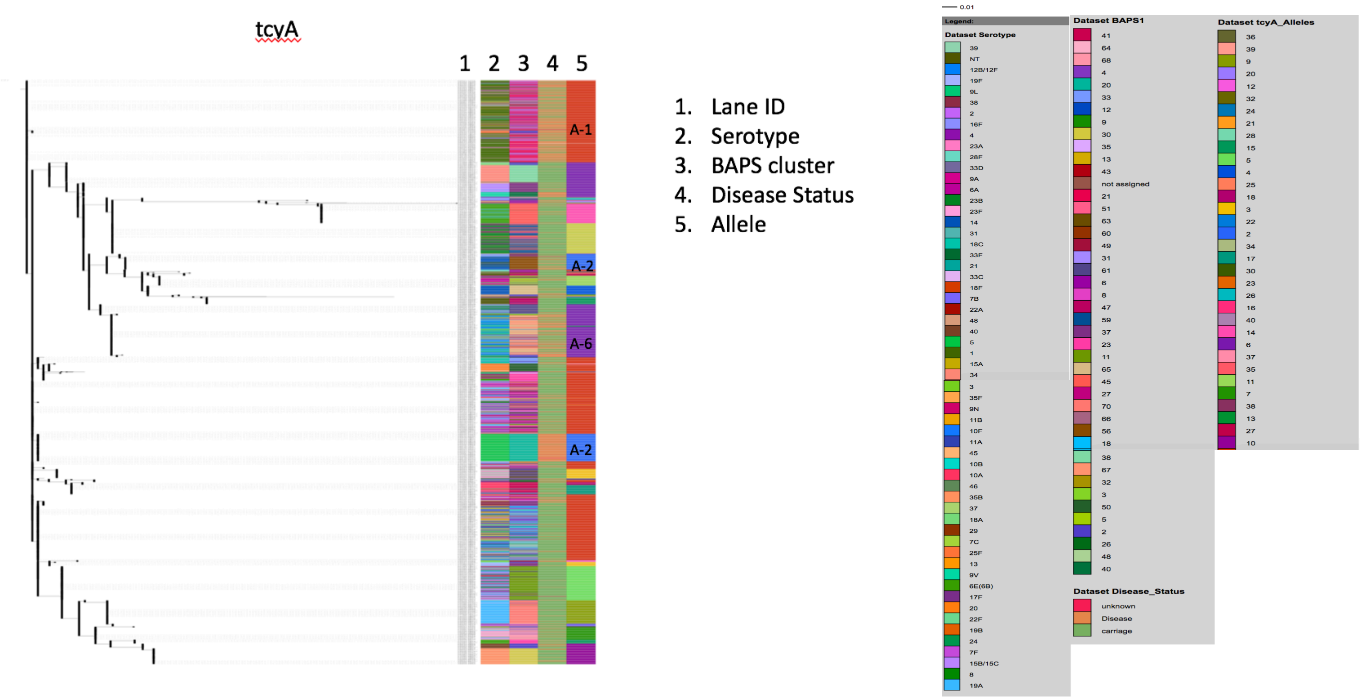


**Figure 3.27 Phylogenetic gene tree of *tcyA*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**95 SNP sites used to reconstruct this gene tree.**

Predicted lipoprotein TcyJ is 266AA long with relatively many alleles (48). It had several alleles that were prevalent in several lineages including alleles, 1, 3, 9 and 10. Allele 3 was the predominant allele in both lineages of serotype 1. Another allele prevalent in an important lineage (BAPS 20, serotype 5) was allele 2.

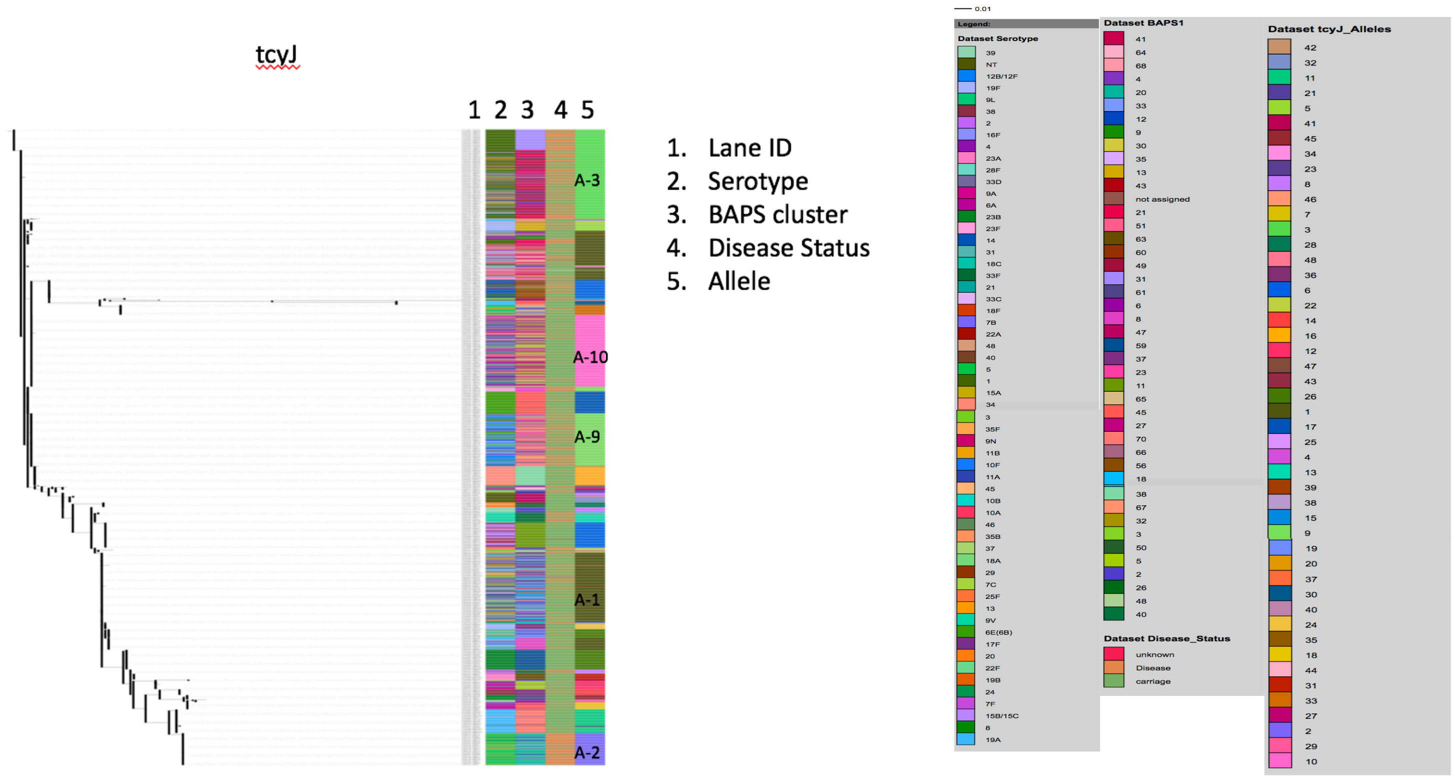


**Figure 3.28 Phylogenetic gene tree of *tcyJ*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**145 SNP sites used for the reconstruction of this gene tree.**

TmpC had only 21 alleles even though it is 350AA long. Allele 1 was the most abundant allele in this protein. Together with alleles 2, 3, 5 and 7, they represented more than 90% of the genomes. All of these alleles covered several lineages but allele 7 had a higher prevalence in BAPS 67 serotype 12B/12F strains with a few other lineages including BAPS 12 & 2 (serotype 3), BAPS 13 (19F) and BAPS 47 (NTs).

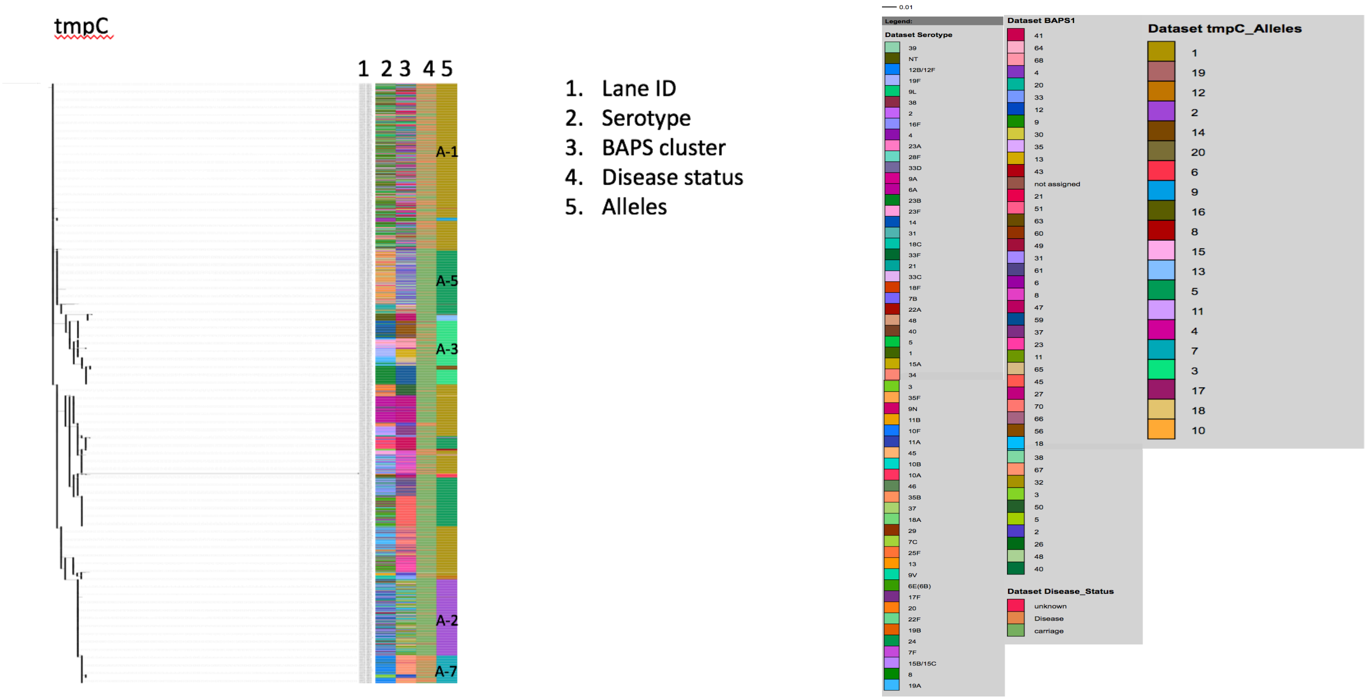


**Figure 3.29 Phylogenetic gene tree of *tmpC*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**84 SNP sites used to reconstruct this gene tree.**

VanYb is 238AA long with 53 alleles. More than 30% of genomes had allele 4, which was the most prevalent allele. Other major alleles included allele 1, 5, 14, 16 and 3. The latter was the predominant allele in serotype 5, BAPS 20 strains.

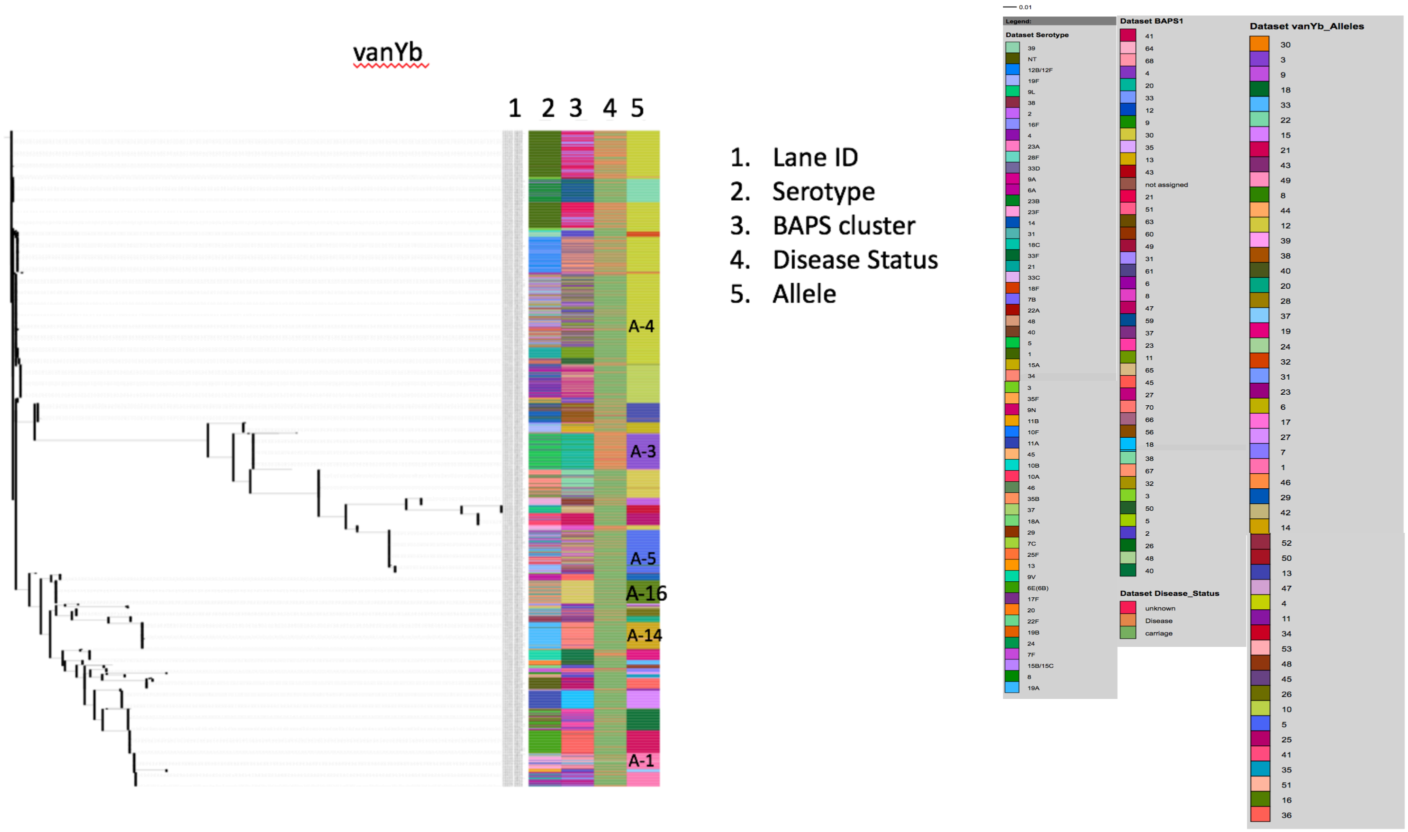


**Figure 3.30 Phylogenetic gene tree of *vanYb*.**

This tree shows the nucleotide relationship of the genes extracted from the genomes.

**256 SNP sites used to reconstruct this gene tree.**

The YesO_2 protein was a large protein with 442AA. It had only 19 alleles here. More than 90% of the genomes had allele 1, which covered approximately all lineages. The next most prevalent allele was allele 4 seen in some 19A, 9V and serotype 13 isolates.

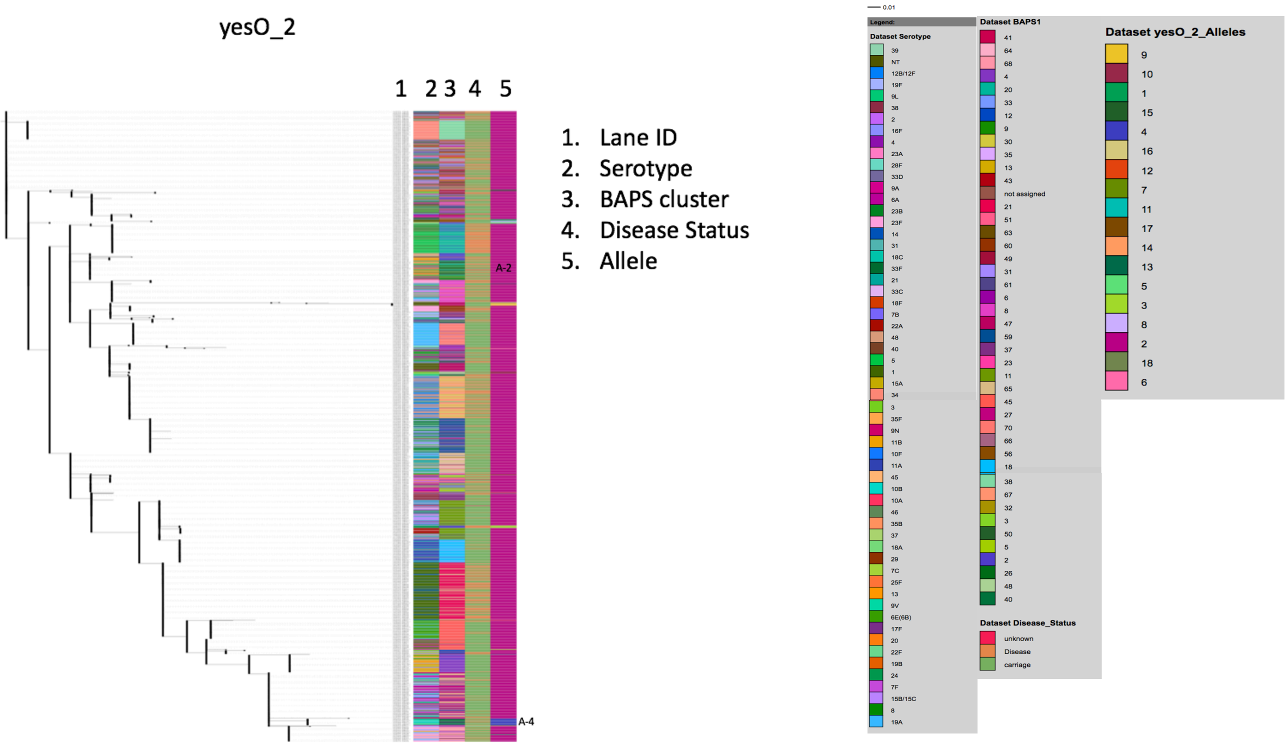


**Figure 3.31 Phylogenetic gene tree of *yesO_2*.**

This tree stars the nucleotide relationship of the genes extracted from the genomes.

**68 SNP sites used to reconstruct this gene tree.**

TauA lipoprotein is an interesting protein which was 335AA long and has only 15 alleles. Like YesO_2, more than 90% of the genomes possessed allele 1. Some BAPS 47 NTs had a unique allele, 8.

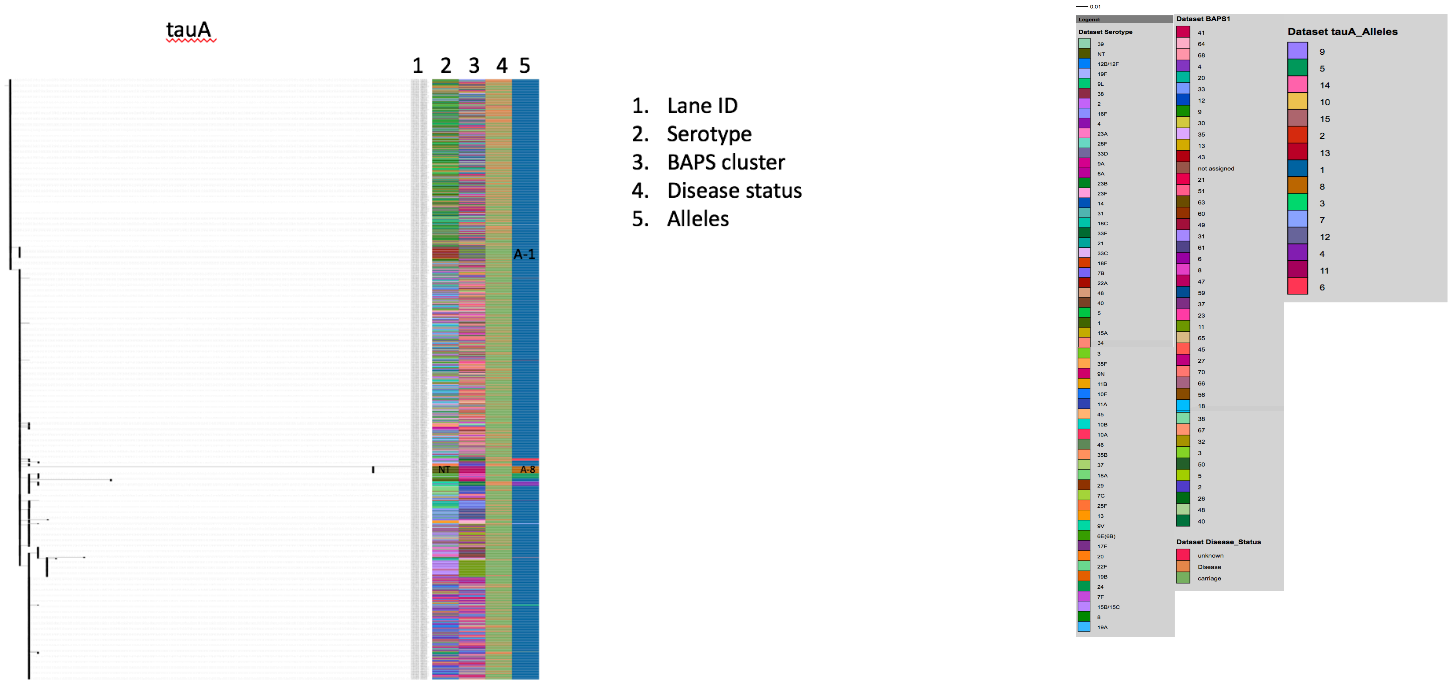


**Figure 3.32 Phylogenetic gene tree of *tauA*.**

This gene tree shows the nucleotide relationship of the genes extracted from the genomes.

**67 SNP sites used to reconstruct this gene tree.**

## 3.6 Protein Epitope Prediction

The linear epitope prediction of all the proteins was performed using the four prediction methods mentioned earlier. Subsequently, their discontinuous epitopes were predicted by both DiscoTope2 and ElliPro. The proportion of the mature proteins predicted by each of the linear prediction as epitope is depicted in Figure 3.33. Consistent with being the most sensitive of all the linear prediction tools, the Bepipred method had predicted more regions as epitopes than any other method overall. This next method with the highest percentage of protein predicted as epitope is the Karplus and Schulz method followed by the Parker method. The Parker method is only slightly more sensitive than the Chou and Fasman method. The schematic representation of the predicted epitopes from the all four methods are also presented below. For all the linear prediction methods, the curves above the threshold (red line) are the predicted epitopes coloured yellow.

Furthermore, the discontinuous epitope prediction by ElliPro also gives the actual number of both predicted linear and discontinuous epitopes and these are presented in Table 3.5. Both ElliPro and DiscoTope2 predicted discontinuous epitopes will be presented using Jmol [152]. For both prediction methods, yellow represents predicted sites. Proteins will be grouped according to function and only a few examples will be presented here and the rest would be in the appendix. Additionally, the proportion of the proteins predicted to be part of an epitope are presented in Fig. 3.32. Based on this analysis, the most immunogenic proteins include TauA, GlnH, Group_2074, Group_2005 and VanYb. All of these proteins except GlnH are ranked amongst the top 10 proteins in my ranking scheme (Table 3.6).

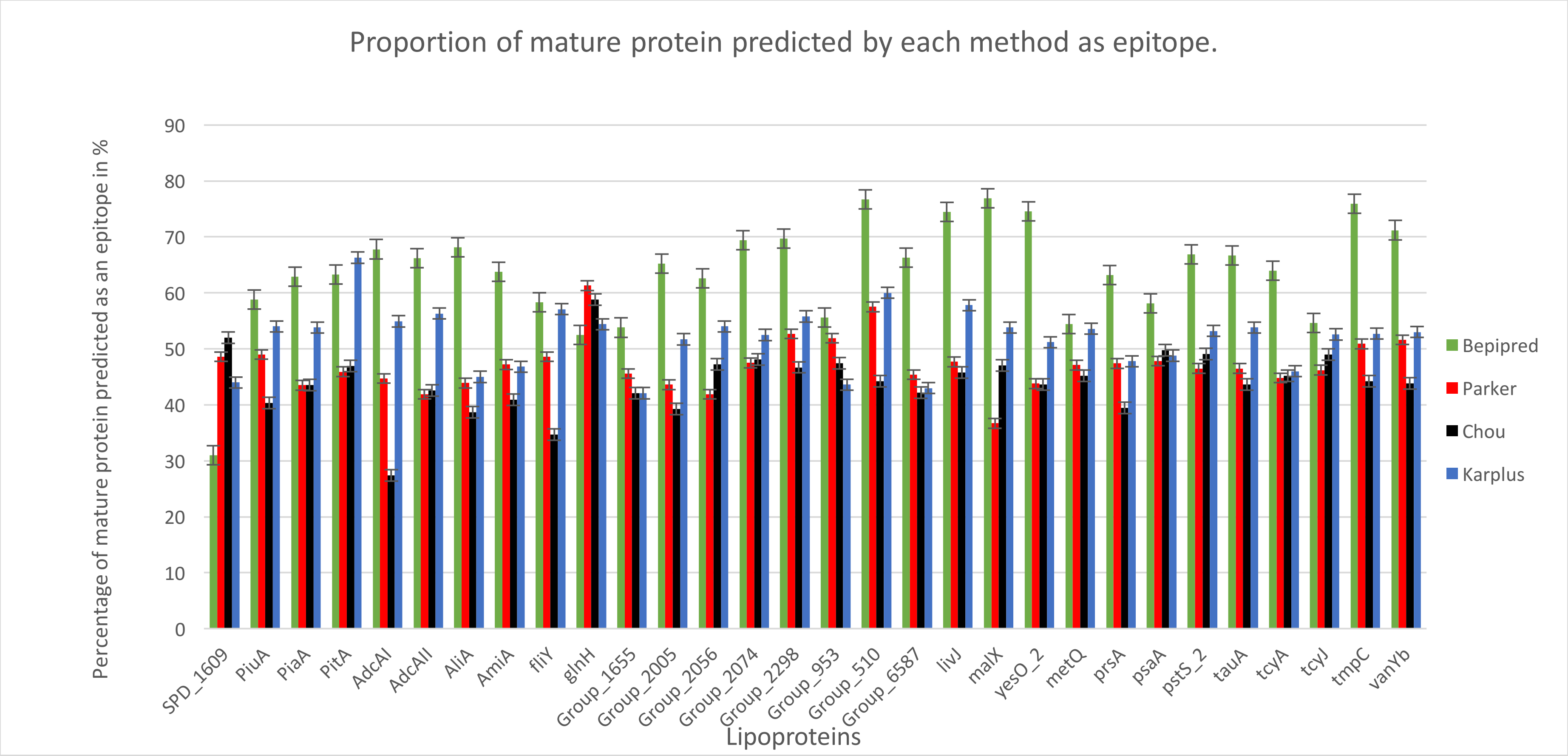


**Figure 3.33 Percentage of mature protein predicted as an Epitope**
The percentage of the mature protein predicted by each linear epitope prediction method is presented here. Green represents the Bepipred method, Red represents the Parker method, Black is the Chou and Fasman method and Blue is the Karplus and Schulz method. The height of the bars represents the percentage of the mature protein predicted as an epitope. Error bars are also placed on top of each bar.   The Bepipred method, which has shown to be the more sensitive of all the methods have been consistent in predicting a greater percentage of the proteins as epitopes except in two proteins. The Karplus and Schulz method has the second highest sensitivity overall followed by the Parker method, which is slightly more sensitive than the Chou and Fasman method.

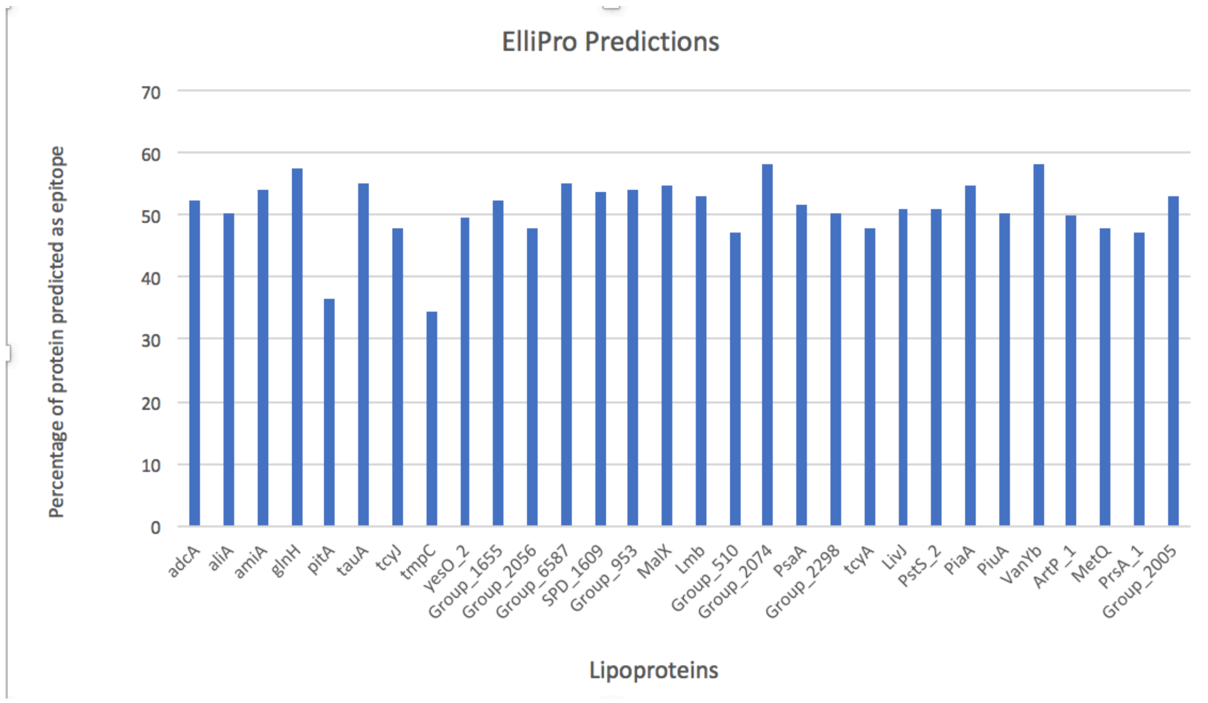


**Figure 3.34 Percentage of proteins predicted as epitope by ElliPro**
The horizontal axis is labelled with the protein names and the vertical axis has the percentages. The height of the bars for each protein represents the percentage of the protein that is predicted by ElliPro to be a part of an Epitope.

**Table 3.5 Epitope prediction results of ElliPro.**

This table includes the protein model used and the source of the model.

| Lipoprotein | Protein Model and/or (Source) | Chains | No. of Linear Epitopes (ElliPro) | No. Discont epitopes (ElliPro) |
|---|---|---|---|---|
| **Group_1655** | (Phyre2) | 1 | 7 | 7 |
| **Group_2005** | 5MLT (PDB) | 2 | 12 | 9 |
| **Group_2056** | (Phyre2) | 1 | 13 | 5 |
| **Group_2074** | 4EVM (PDB) | 1 | 7 | 2 |
| **Group_2298** | 4HQZ (PDB) | 2 | 10 | 3 |
| **Group_510** | 3GE2 (PDB) | 1 | 4 | 4 |
| **Group_6587** | (Phyre2) | 1 | 7 | 3 |
| **Group_953** | (I-TASSER) | 1 | 7 | 7 |
| **AdcA** | (Phyre) | 1 | 12 | 5 |
| **AliA** | (I-TASSER) | 1 | 17 | 4 |
| **AmiA** | (Phyre2) | 1 | 13 | 11 |
| **ArtP_1** | 4OHN (PDB) | 1 | 9 | 4 |
| **GlnH** | (I-TASSER) | 1 | 11 | 7 |
| **LivJ** | 4GNR(PDB) | 1 | 11 | 4 |
| **Lmb** | 3CX3 (PDB) | 2 | 10 | 7 |
| **MalX** | 2XD2 (PDB) | 2 | 12 | 3 |
| **MetQ** | 4Q5T (PDB) | 1 | 8 | 6 |
| **PiaA** | 4HMO (PDB) | 1 | 10 | 3 |
| **PiuA** | 4JCC (PDB) | 1 | 7 | 5 |
| **PitA** | (Phyre2) | 1 | 3 | 4 |
| **PrsA** | 5TVL (PDB) | 4 | 14 | 5 |
| **PsaA** | 4UTP (PDB) | 2 | 14 | 4 |
| **PstS_2** | 4H1X (PDB) | 1 | 11 | 6 |
| **SPD_1609** | (I-TASSER) | 1 | 11 | 4 |
| **TauA** | (Phyre2) | 1 | 11 | 6 |
| **TcyA** | 4EQ9 (PDB) | 1 | 7 | 6 |

| | | | | |
|---|---|---|---|---|
| **TcyJ** | 5COR (PDB) | 2 | 15 | 4 |
| **TmpC** | (I-TASSER) | 1 | 11 | 4 |
| **VanYb** | 4NT9 (PDB) | 3 | 16 | 3 |
| **YesO_2** | (Phyre2) | 1 | 18 | 5 |

PiaA, PiuA, PitA and SPD_1609 are iron transporter proteins. With the exception of PitA, the rest have similar sizes and the number of epitopes predicted for these proteins are presented in Table 3.4. Most of the major epitopes (higher peaks) of PiaA are predicted corrected by all four methods (Figure 3.35). The areas with the highest peaks predicted by most or all methods includes areas approximately between amino acid 20-40, 80-100, 180-200 and 240-260. For the discontinuous prediction, protein model 4HMO from PDB, which was 99% identical to my protein sequence was used. The ElliPro method (Fig. 3.36) predicted more sites as discontinuous epitopes than DiscoTope2 for PiaA (Fig. 3.37). This was also true for PiuA (Fig. A2 and A3) and SPD_1609 (Fig. A5 and A6) but DiscoTope2 predicted more sites as epitopes than ElliPro for PitA (Fig. A8 and A9). PiuA had 11 linear epitopes predicted by Bepipred. The four methods had their highest prediction peaks approximately between 20-40, 80-100, 130-145 and 160-190. Protein model 4JCC, which had 94% identity to PiuA was used for the discontinuous prediction and ElliPro predicted 5 discontinuous epitopes. Also, SPD_1609 and PitA had 16 and 8 linear epitopes predicted by Bepipred respectively. PDB did not have a suitable model for either protein and Phyre2 could not model SPD_1609 with a high level of confidence, so I-TASSER was used for the modelling. However, Phyre2 was used for modelling PitA. Using these models, ElliPro predicted 4 discontinuous epitopes for both proteins. DiscoTope2 had no sites predicted as epitope for SPD_1609 (Fig. A6).

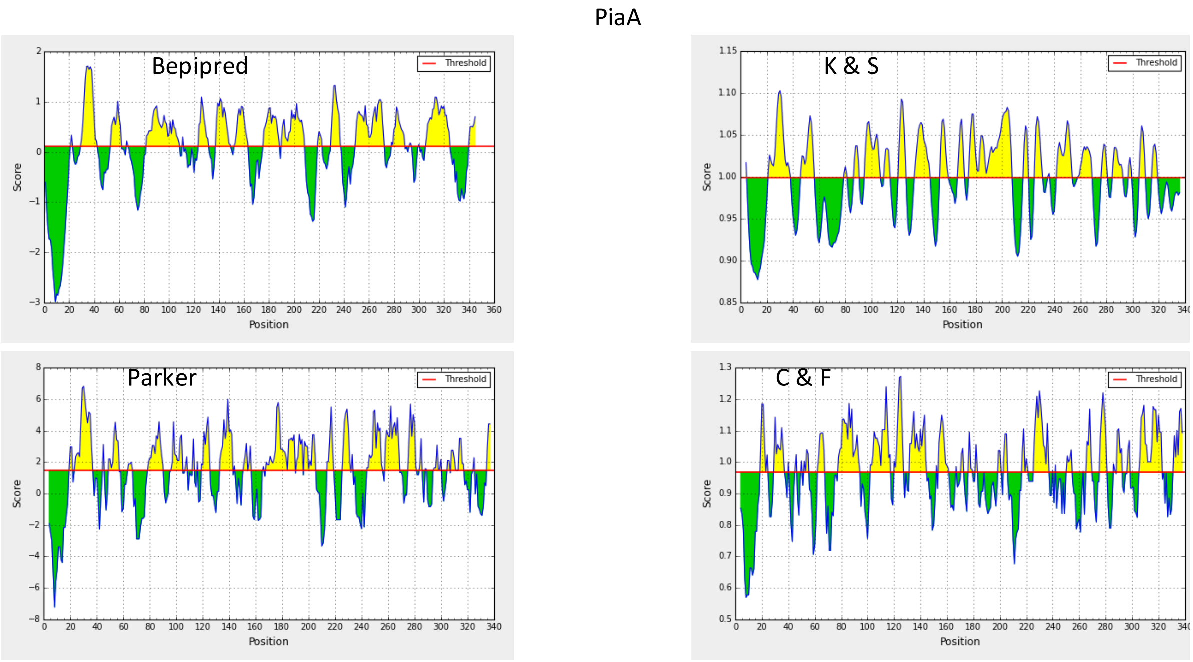


**Figure 3.35 Linear epitope predictions of the PiaA protein.**

This was predicted using Bepipred, Parker, Karplus and Schulz (K&S) and Chou and Fasman (C&F) methods.

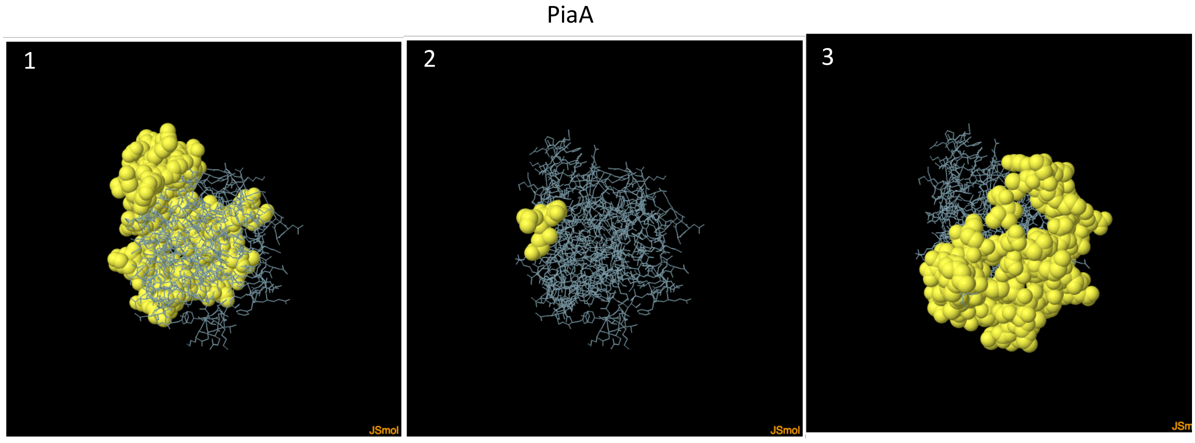


**Figure 3.36 ElliPro predicted discontinuous epitopes for PiaA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with one having the highest score. The yellow spheres represent residues part of the predicted epitope.

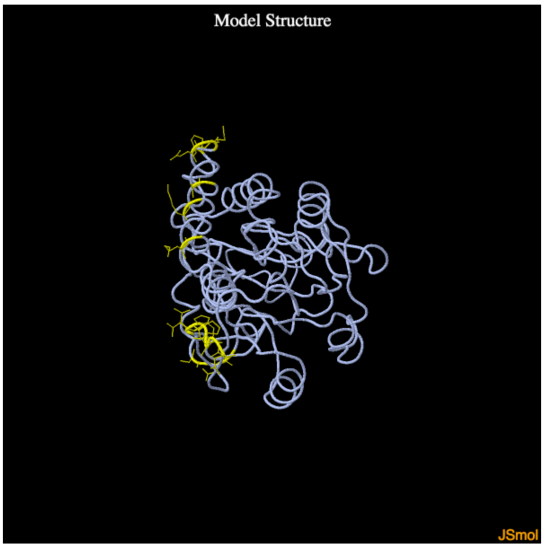


**Figure 3.37 DiscoTope2 predicted discontinuous epitopes for PiaA.**
The parts coloured yellow are the predicted epitopes.

AdcA and Lmb are both involved in zinc transport. However, these two lipoproteins are different in both their size and structure. Bepipred predicted 22 and 13 linear epitopes for AdcA and Lmb respectively. The epitopes for AdcA are spread evenly across the whole sequence for all the methods (Fig A10) while epitopes with the highest predictions in Lmb fall at the beginning, middle and end of the sequence. PDB model, 3CX3 was used for the discontinuous predictions of Lmb but AdcA was modelled using the Phyre2 server. ElliPro predicted 5 discontinuous epitopes for AdcA and the first two predictions cover the predicted sites by the DiscoTope2 method (Fig. A11-A12). Furthermore, both ElliPro and DiscoTope2 predictions for Lmb were in good concordance.

AliA and AmiA are both involved in oligopeptide transport. Despite their structural difference, they have approximately the same sequence length. Both proteins have predicted epitopes spread evenly across their sequences by all the methods. With 35 and 32 predicted linear epitopes by Bepipred for AliA and AmiA respectively, they had the first and second highest number of predicted linear epitopes in this dataset. Conversely, ElliPro predicted more discontinuous epitopes (11) for AmiA than AliA (4). The protein model used for the discontinuous predictions was modelled using the I-TASSER server. For both proteins ElliPro predicted more discontinuous epitopes than

DiscoTope2, however, there was good concordance with the areas predicted by both methods (Fig. A17-A18 and Fig. A20-A21).

PsaA had 14 linear epitopes predicted by Bepipred. A similar prediction pattern was observed with the other linear prediction methods. The 4UTP model was used for the discontinuous predictions. The two discontinuous prediction methods showed high concordance for both chains in this model as illustrated below.

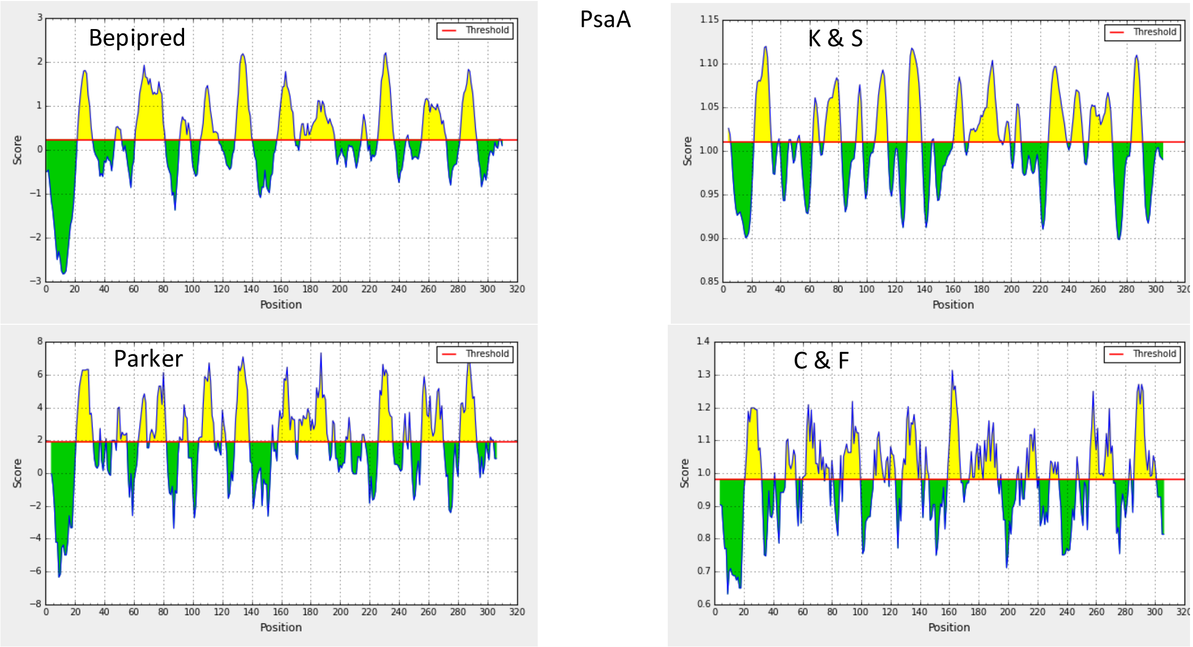


**Figure 3.38 Linear epitope predictions of the PsaA protein.**

This was predicted using Bepipred, Parker, Karplus and Schulz (K&S) and Chou and Fasman (C&F) methods.

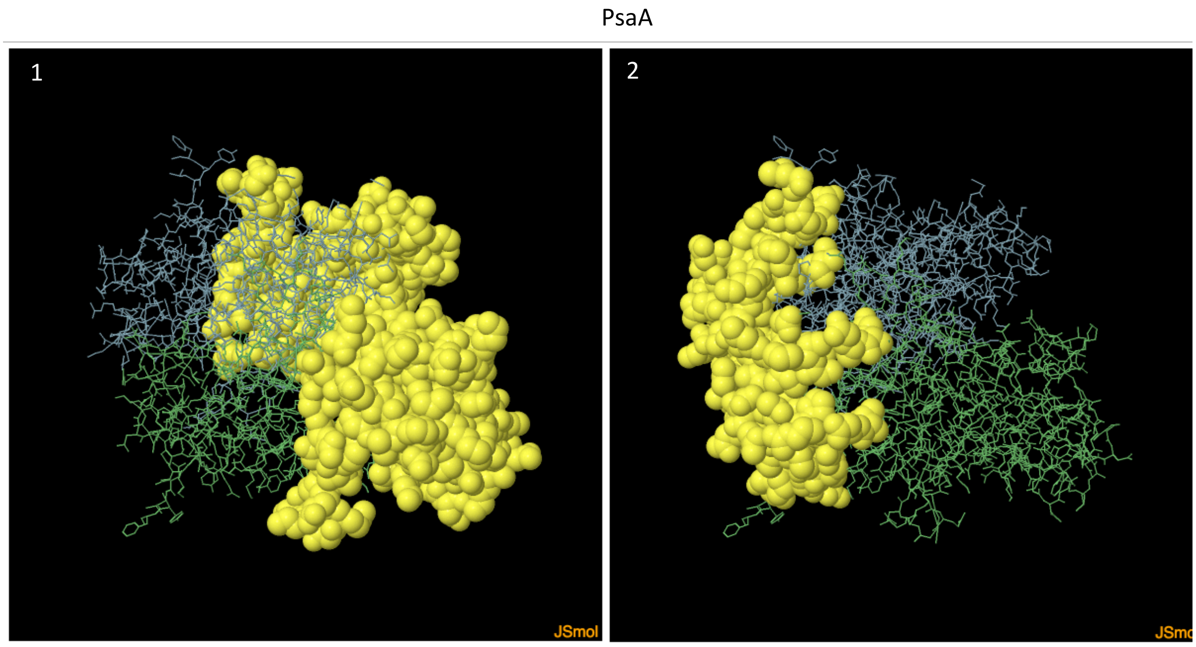


**Figure 3.39 ElliPro predicted discontinuous epitopes for PsaA.**

The numbers represent the different epitopes predicted in the order of decreasing overall score with one having the highest score. The yellow spheres represent residues part of the predicted epitope.

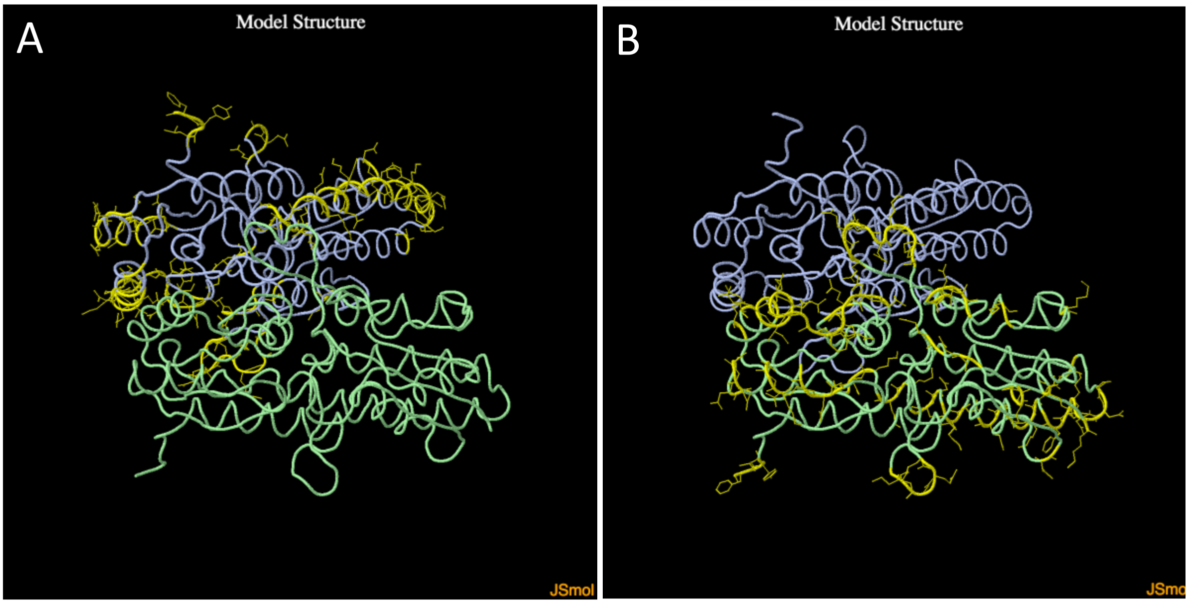


**Figure 3.40 DiscoTope2 predicted discontinuous epitopes for PsaA.**

The parts coloured yellow are the predicted epitopes. A is the prediction for chain A and B for chain B.

MalX had 15 linear epitopes predicted by Bepipred. The epitopes are evenly spread across the protein sequence. Protein model 2XD2 was used for the discontinuous predictions. This model has two chains and the ElliPro prediction predicted 3 large epitopes. In contrast, DiscoTope2 predicted only a few small regions. However, most of these regions were also predicted by ElliPro.

YesO_2 had 18 predicted linear epitopes by Bepipred. These are spread evenly across the sequence with several high peaks (Fig. A25). The protein was modelled using Phyre2 for the discontinuous predictions. The 5 epitopes predicted by ElliPro covers more sites than the DiscoTope2 prediction, however, the two methods concurred in almost all the DiscoTope2 predicted sites.

PrsA_1 had 12 Bepipred predicted linear epitopes. The highest peaks predicted by all the methods are around the middle of the sequence and at the far end (Fig. 3.41). The 5TVL model, which has 4 chains was used for the discontinuous predictions. ElliPro predicted 5 discontinuous epitopes (Fig. 3.42) and these covered the entire structure of this protein. DiscoTope2 also predicted many regions of this protein as epitopes (Figure 3.43).

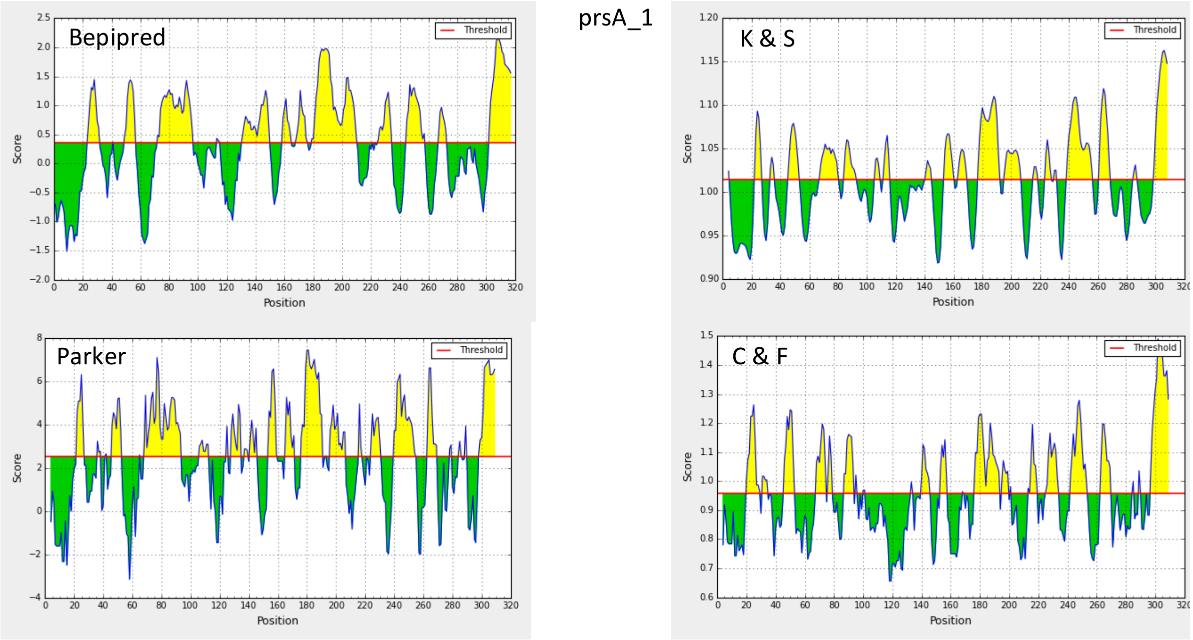


**Figure 3.41 Linear epitope predictions of the PrsA_1 protein.**

This was predicted using Bepipred, Parker, Karplus and Schulz (K&S) and Chou and Fasman (C&F) methods.

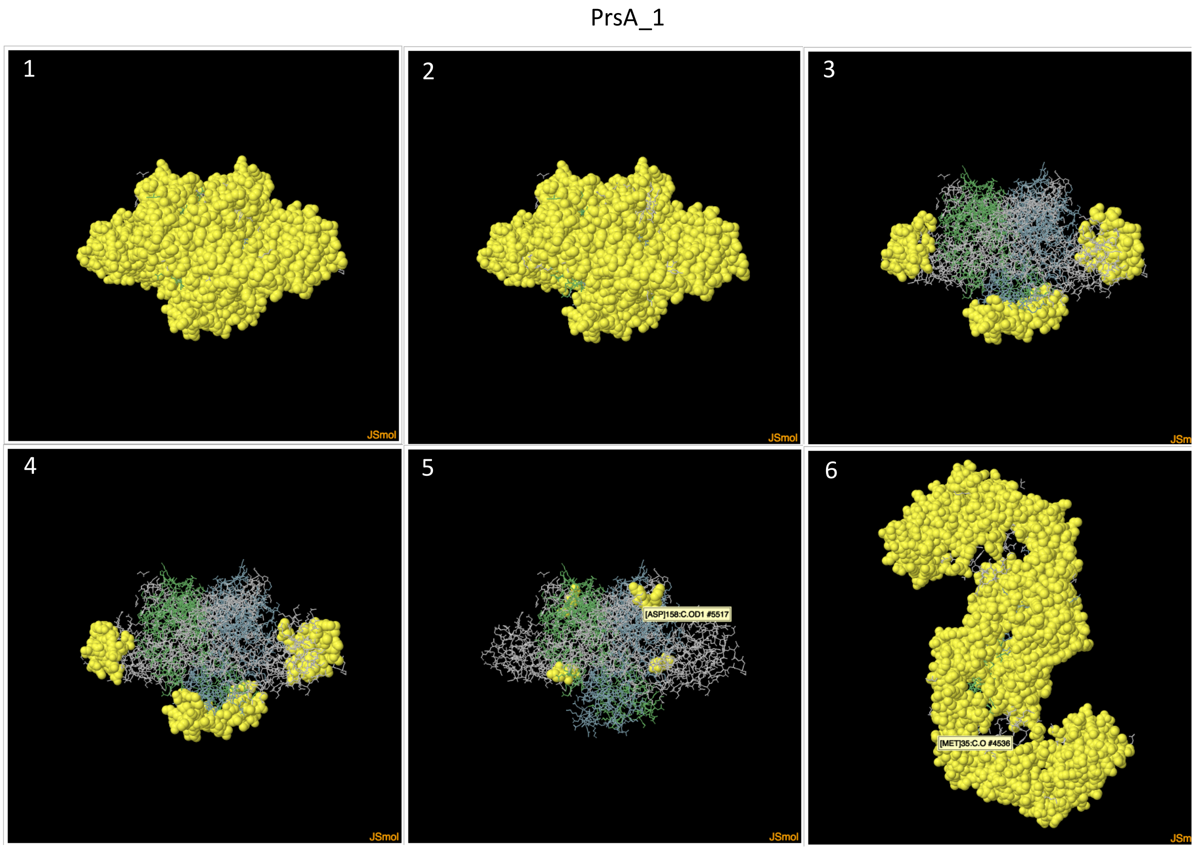


**Figure 3.42 ElliPro predicted discontinuous epitopes for PrsA_1.**

The numbers represent the different epitopes predicted in order of decreasing overall score with one having the highest score. The number 6 in this figure is just the first prediction reoriented to illustrate the size of the protein and all the regions predicted as epitopes. The yellow spheres represent residues part of the predicted epitope.

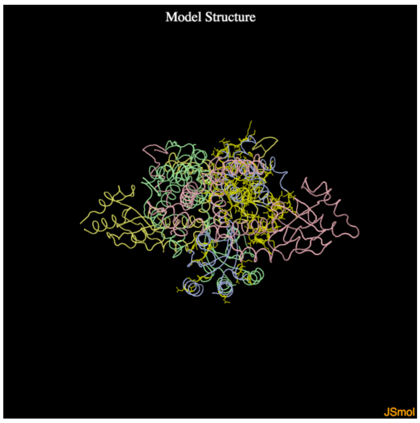


**Figure 3.43 DiscoTope2 predicted discontinuous epitopes for prsA_1.**
The parts coloured yellow are the predicted epitopes.

Group_2005 and Group_2056 lipoproteins are carbohydrate transporters. Group_2005 are bigger than Group_2056 lipoproteins and they are structurally different. Bepipred predicted 31 linear epitopes for Group_2005 and 21 for Group_2056 proteins. All the methods predicted numerous epitopes over short sequence stretches. Group_2005 had a sufficient PDB protein model (5MLT) but Group_2056 had to be modelled *de novo* using the Phyre2 server. ElliPro predicted more discontinuous epitopes for Group_2005 (9) than Group_2056 (5) lipoproteins (Fig. A29 and A32). The ElliPro predictions for both proteins concurred with the epitopes predicted by DiscoTope2 (Fig. A30 and Fig. A33).

TauA had 14 predicted linear epitopes by Bepipred. The linear prediction methods agreed mostly and the highest peaks can be seen around the start and middle regions of the sequence (Figure A34). Phyre2 was used to model the protein for the discontinuous predictions. The two discontinuous prediction methods agreed mostly but again ElliPro had more regions predicted.

Bepipred predicted 14 linear epitopes for MetQ. All the linear prediction methods agreed especially in areas with the highest peaks (Fig. A37). Protein model 4Q5T was used to predict discontinuous epitopes. The 6 epitopes predicted by ElliPro are highly concordant with the DiscoTope2 predicted regions (Fig. A38 and A39).

Bepipred predicted 10 linear epitopes for PstS_2. The epitopes with the highest scores can be seen around the start and end regions of the sequence but overall, the predicted epitopes are evenly spread across the sequence (Fig. A40). The regions predicted as discontinuous epitopes by DiscoTope2 (Fig. A42) are also predicted by ElliPro (Fig. A41) but ElliPro had more regions predicted.

The pneumococcus has many amino acid transporters and quite a few are amongst the proteins in this dataset. GlnH, LivJ, ArtP_1, TcyA and TcyJ are all lipoproteins involved in this process. Although all of them have unique structures suggesting affinity for different amino acids, they are all relatively the same size. TcyJ is the smallest with 2666AA and the largest is ArtP_1 with 278AA. The Bepipred predictions of these proteins are summarised in Table 3.4. Also, the model used and the number of predicted epitopes by ElliPro for each protein is summarised in Table 3.5. While ElliPro predicted 7 discontinuous sites for GlnH (Fig. A47) covering many regions, DiscoTope2 predicted only a few sites (Fig. A48). Similarly, DiscoTope2 had predicted less sites as epitopes for all the proteins. It predicted only a few small regions as epitopes for both ArtP_1 (Fig. A45) and LivJ (Fig. A51) compared to ElliPro, which had 4 predicted regions covering most of structure of each of these two proteins. Furthermore, TcyJ and TcyA had 2 and 6 predicted discontinuous epitopes by ElliPro respectively. These predictions were in high concordance with the DiscoTope2 predicted sites.

Group_2074 and Group_2298 are thioredoxin proteins. Their Bepipred predicted linear epitopes are summarised in Table 3.4. The protein models used for the discontinuous epitopes were 4EVM and 4HQZ for Group_2074 and Group_2298 respectively. ElliPro predicted 2 discontinuous epitopes for Group_2074 (Fig. A59), however, DiscoTope2 predicted no epitopes for Group_2074 (Fig. A60). The ElliPro prediction for Group_2298 (Fig. A62) also covered the predicted sites by DiscoTope2 (Fig. A63) for this protein.

Bepipred predicted only 6 linear epitopes for VanYb. Although some of the methods seem to predict more epitopes, all the methods seem to be in concordance with the regions with the highest peaks (Fig. A64). Protein model 4NT9 was used for the discontinuous predictions. This protein has three chains and the 3 predicted epitopes by ElliPro agreed well with the DiscoTope2 predictions for each of the chains (Fig. A65-A66).

TmpC had 14 linear epitopes predicted by Bepipred. These are spread evenly across the sequence (Fig. A67). The I-TASSER server was used to model this protein for the discontinuous predictions. ElliPro had four epitopes predicted and DiscoTope2 has only small segments of the protein predicted as epitopes (Fig. A69). These segments are also covered by the first two predictions of ElliPro (Fig. A68).

Group_510 had 6 epitopes predicted by Bepipred. The Bepipred prediction covered most parts between approximately residue 25 and 95, also between 105 and 120. The Parker and K&S methods are more in agreement, predicting most areas between approximately residue 25-75. The C&F method also predicted a similar area but starting around the 40th residue. PDB model 3GE2, which was 98% identical to my protein was used for the discontinuous epitope predictions. ElliPro predicted 4 discontinuous epitopes (Fig. A71). Most of the ElliPro predicted epitopes agreed with those predicted by DiscoTope2, however, it seems as DiscoTope2 predicted more sites as epitopes.

Bepipred predicted 8 epitopes for Group_6587. The predicted epitopes with the highest peaks for all methods are found approximately between residues 125 and 220. The sequence between 240-250 was also predicted as an epitope by all methods (Fig. A73). ElliPro predicted 3 discontinuous epitopes using a Phyre2 modelled protein (Fig. A74). Although it predicted more sites than the DiscoTope2 method, the regions predicted by DiscoTope2 (Fig. A75) were in concordance with those predicted by ElliPro.

The linear epitope count from the Bepipred method for Group_953 was 12. The beginning and middle part of the sequence seems to have more predicted epitopes (Fig. A76). The protein model used for the discontinuous predictions was modelled *de novo* using the I-TASSER server. ElliPro had 7 predicted discontinuous epitopes (Fig. A77) and these were consistent with the sites predicted by DiscoTope2 (Fig. A78).

The linear predictions for Group_1655 were similar for all the methods (Fig. A79). This protein had a good model from PDB (2MVB) with 99% identity, which has one chain.

The 7 predicted epitopes by ElliPro and the DiscoTope2 predictions are presented in Fig. A80 and Fig. A81 respectively. The second and third predictions of ElliPro seem to be in concordance with the DiscoTope2 predicted sites.

## 3.7 Protein Rank

After all these analyses, I set out to rank my proteins using a simple scoring algorithm as detailed in the methods section. The ElliPro prediction method was used for the proportion of each lipoprotein predicted as epitope. The number of chains for each protein was obtained from the PDB database were available or from the Phyre2 or I-TASSER servers when modelled *de novo*. The results of this ranking are summarised in table 3.6 below. Briefly, proteins in the top 10 based on total points starting from number one are Group_2005, TauA, Group_2074, AmiA, PsaA, Lmb, vanYb, YesO_2, Group_953 and PrsA.

## Table 3.6 Protein characteristics and point-based ranking.

| Lipoprotein | Aa length | Points(Pts) | No. of Alleles | Pts | Percentage of protein predicted as an Epitopes (ElliPro) | Pts | Chains | Pts | Prevalence | Pts | Total Points | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Group_1655 | 165 | 2 | 36 | 20 | 52.3% | 52.3 | 1 | 2 | 100% | 10 | 86.3 | 17 |
| Group_2005 | 503 | 9 | 26 | 22 | 53% | 53 | 2 | 4 | 100% | 10 | 98 | 1 |
| Group_2056 | 445 | 7 | 35 | 20 | 47.9% | 47.9 | 1 | 2 | 99.4% | 9.4 | 86.3 | 17 |
| Group_2074 | 188 | 2 | 11 | 24 | 58% | 58 | 1 | 2 | 100% | 10 | 96 | 3 |
| Group_2298 | 185 | 2 | 26 | 22 | 50.4% | 50.4 | 2 | 4 | 98.2% | 8.2 | 86.6 | 16 |
| Group_510 | 164 | 2 | 39 | 20 | 47.1% | 47.1 | 1 | 2 | 95.9% | 5.9 | 77 | 26 |
| Group_6587 | 268 | 4 | 31 | 20 | 55% | 55 | 1 | 2 | 100% | 10 | 91 | 12 |
| Group_953 | 292 | 4 | 28 | 22 | 54.1% | 54.1 | 1 | 2 | 99.9 | 9.9 | 92 | 9 |
| adcA | 501 | 9 | 82 | 10 | 52.3% | 52.3 | 1 | 2 | 99.7% | 9.7 | 83 | 23 |
| aliA | 662 | 12 | 126 | 2 | 50.2% | 50.2 | 1 | 2 | 98.5% | 8.5 | 74.7 | 28 |
| amiA | 660 | 10 | 37 | 20 | 54% | 54 | 1 | 2 | 99.8 | 9.8 | 95.8 | 4 |
| artP_1 | 278 | 4 | 38 | 20 | 49.8% | 49.8 | 1 | 2 | 100% | 10 | 85.8 | 19 |
| glnH | 275 | 4 | 67 | 14 | 57.4% | 57.4 | 1 | 2 | 97.9% | 7.9 | 85.3 | 21 |
| livJ | 386 | 6 | 37 | 20 | 51.1% | 51.1 | 1 | 2 | 98.8% | 8.8 | 87.9 | 15 |
| Lmb | 305 | 5 | 23 | 22 | 52.9% | 52.9 | 2 | 4 | 100% | 10 | 93.9 | 6 |
| malX | 423 | 7 | 51 | 16 | 54.6% | 54.6 | 2 | 4 | 99.6% | 9.6 | 91.1 | 11 |
| metQ | 284 | 4 | 28 | 22 | 47.8% | 47.8 | 1 | 2 | 100% | 10 | 85.8 | 19 |
| PiaA | 342 | 5 | 31 | 20 | 54.9% | 54.9 | 1 | 2 | 98% | 8 | 89.9 | 13 |
| PiuA | 322 | 5 | 60 | 14 | 50.4% | 50.4 | 1 | 2 | 100% | 10 | 81.4 | 25 |
| PitA | 122 | 1 | 25 | 22 | 36.4% | 36.4 | 1 | 2 | 94.1% | 4.1 | 65.5 | 30 |
| prsA | 316 | 5 | 24 | 22 | 47% | 47 | 4 | 8 | 99.8% | 9.8 | 91.8 | 10 |
| psaA | 309 | 5 | 17 | 24 | 51.6% | 51.6 | 2 | 4 | 100% | 10 | 94.6 | 5 |
| pstS_2 | 291 | 4 | 23 | 22 | 51.1% | 51.1 | 1 | 2 | 100% | 10 | 89.1 | 14 |
| SPD_1609 | 357 | 6 | 101 | 6 | 53.7% | 53.7 | 1 | 2 | 99% | 9 | 76.7 | 27 |
| tauA | 335 | 5 | 15 | 24 | 55.1% | 55.1 | 1 | 2 | 100% | 10 | 96.1 | 2 |
| tcyA | 278 | 4 | 40 | 18 | 47.9% | 47.9 | 1 | 2 | 99.7% | 9.7 | 81.6 | 24 |
| tcyJ | 266 | 4 | 48 | 18 | 47.8% | 47.8 | 2 | 4 | 99.9% | 9.9 | 83.7 | 22 |
| tmpC | 350 | 5 | 21 | 22 | 34.3% | 34.3 | 1 | 2 | 99.9% | 9.9 | 73.2 | 29 |
| vanYb | 238 | 3 | 53 | 16 | 58.2% | 58.2 | 3 | 6 | 99.8% | 9.8 | 93 | 7 |
| YesO_2 | 442 | 7 | 19 | 24 | 49.5% | 49.5 | 1 | 2 | 100% | 10 | 92.5 | 8 |

## 3.8 Presence in other streptococcal species

Using a protein blast search against both NCBI non-redundant protein database and UniProt, all these candidate proteins were found in at least one streptococcus species other than *S. pneumoniae* with very high nucleotide identity across the entire length with the exception of PiaA. The results have been summarised in Table 3.7.

**Table 3.7 Presence of candidate proteins in non-pneumococcal Streptococcus.**

The table also has the level of identity of these proteins in each non-pneumococcal streptococci. Only the top results for each species are presented.

| Gene | Top Hit Non-pneumococcal Streptococcus | Percent ID |
|---|---|---|
| Group_1655 | *S. mitis/ S. pseudopneumoniae* | 97%/97.4 |
| Group_2005 | *S. pseudopneumoniae/ S. mitis/S. oralis* | 98.8%/98.0%/96.1% |
| Group_2056 | *S. mitis/ S. pseudopneumoniae* | 99.3%/99.3% |
| Group_2074 | *S. pseudopneumoniae* | 97% |
| Group_2298 | *S. mitis/ S. pseudopneumoniae* | 97%/96% |
| Group_510 | *S. mitis* | 77.3% |
| Group_6587 | *S. mitis/ S. oralis/ S. pseudopneumoniae* | 97%/96.3%/97% |
| Group_953 | *S. mitis/ S. pseudopneumoniae* | 96%/ 96% |
| AdcA | *S. mitis/ S. oralis/ S. pseudopneumoniae* | 99%/ 96.0%/ 99.2% |
| AliA | *S. mitis/ S. psuedopneumoniae* | 93.8%/ 92.6% |
| AmiA | *S. pseudopneumoniae* | 98.3% |
| ArtP_1 | *S. mitis/ S. pseudopneumoniae* | 98.6%/ 97.8% |
| GlnH | *S. mitis/ S. pseudopneumoniae* | 98.2%/97.8% |
| LivJ | *S. pseudopneumoniae/ S. mitis/ S. oralis* | 99%/ 98.4%/96.1% |
| Lmb | *S. mitis/ S. pseudopneumoniae* | 99%/99.3% |
| MalX | *S. mitis/ S. oralis* | 96.5%/94.6% |
| MetQ | *S. pseudopneumoniae/ S. mitis* | 99%/ 99% |
| PiaA | *None* | None |
| PiuA | *S. oralis/S. mitis* | 85%/86% |
| PitA | *S. mitis/ S. oralis/ S. pseudopneumoniae* | 100%/ 100%/ 100% |
| PrsA | *S. mitis/ S. oralis S. pseudopneumoniae /* | 96%/ 88.2%/ 99% |

| | | |
|---|---|---|
| PsaA | *S. pseudopneumoniae / S. mitis/ S. oralis* | 98%/ 97%/ 100% |
| PstS_2 | *S. pseudopneumoniae / S. dysgalactiae* | 99.3%/ 89.3% |
| SPD_1609 | *S. pseudopneumoniae* | 90% |
| TauA | *S. pseudopneumoniae / S. mitis* | 96%/96% |
| TcyA | *S. pseudopneumoniae / S. mitis/ S. oralis* | 96% /96% / 91% |
| TcyJ | *S. mitis/ S. pseudopneumoniae* | 97.7%/ 98.1% |
| TmpC | *S. mitis/ S. oralis S. pseudopneumoniae* | 99%/ 95.7%/ 97.7% |
| VanYb | *S. pseudopneumoniae / S. mitis / S. oralis* | 99%/ 98%/ 97.1% |
| YesO_2 | *S. mitis/ S. oralis S. pseudopneumoniae* | 99%/ 95.2%/ 98%/ |

## 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 $\alpha$-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 $\alpha$-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.

# References:

1. Kadioglu, A., et al., *The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease.* Nat Rev Micro, 2008. **6**(4): p. 288-301.
2. Argondizzo, A.P.C., et al., *Pneumococcal Predictive Proteins Selected by Microbial Genomic Approach Are Serotype Cross-Reactive and Bind to Host Extracellular Matrix Proteins.* Applied Biochemistry and Biotechnology, 2017: p. 1-22.
3. Lanie, J.A., et al., *Genome Sequence of Avery's Virulent Serotype 2 Strain D39 of Streptococcus pneumoniae and Comparison with That of Unencapsulated Laboratory Strain R6.* Journal of Bacteriology, 2007. **189**(1): p. 38-51.
4. Bogaert, D., R. de Groot, and P.W.M. Hermans, *Streptococcus pneumoniae colonisation: the key to pneumococcal disease.* The Lancet Infectious Diseases, 2004. **4**(3): p. 144-154.
5. McCool, T.L. and J.N. Weiser, *Limited Role of Antibody in Clearance of Streptococcus pneumoniae in a Murine Model of Colonization.* Infection and Immunity, 2004. **72**(10): p. 5807-5813.
6. Hausdorff, W.P., et al., *Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I.* Clin Infect Dis, 2000. **30**(1): p. 100-21.
7. Robinson, K.A., et al., *Epidemiology of invasive streptococcus pneumoniae infections in the united states, 1995-1998: Opportunities for prevention in the conjugate vaccine era.* JAMA, 2001. **285**(13): p. 1729-1735.
8. Weinberger, D.M., R. Malley, and M. Lipsitch, *Serotype replacement in disease after pneumococcal vaccination.* Lancet, 2011. **378**(9807): p. 1962-73.
9. Spratt, B.G. and B.M. Greenwood, *Prevention of pneumococcal disease by vaccination: does serotype replacement matter?* Lancet, 2000. **356**(9237): p. 1210-1.
10. Diawara, I., et al., *Invasive pneumococcal disease among children younger than 5 years of age before and after introduction of pneumococcal conjugate vaccine in Casablanca, Morocco.* Int J Infect Dis, 2015. **40**: p. 95-101.
11. Lehmann, D., et al., *The changing epidemiology of invasive pneumococcal disease in aboriginal and non-aboriginal western Australians from 1997 through 2007 and emergence of nonvaccine serotypes.* Clin Infect Dis, 2010. **50**(11): p. 1477-86.
12. Wenger, J.D., et al., *Invasive pneumococcal disease in Alaskan children: impact of the seven-valent pneumococcal conjugate vaccine and the role of water supply.* Pediatr Infect Dis J, 2010. **29**(3): p. 251-6.
13. Obaro, S.K., et al., *Carriage of pneumococci after pneumococcal vaccination.* Lancet, 1996. **348**(9022): p. 271-2.
14. Cheung, Y.B., et al., *Nasopharyngeal carriage of Streptococcus pneumoniae in Gambian children who participated in a 9-valent pneumococcal conjugate vaccine trial and in their younger siblings.* Pediatr Infect Dis J, 2009. **28**(11): p. 990-5.
15. Levy, C., et al., *PneumococcaL meningitis in french children before and after the introduction of pneumococcal conjugate vaccine.* Pediatr Infect Dis J, 2011. **30**(2): p. 168-70.
16. Brueggemann, A.B., et al., *Vaccine escape recombinants emerge after pneumococcal vaccination in the United States.* PLoS Pathog, 2007. **3**(11): p. e168.

17. Demczuk, W.H.B., et al., *Phylogenetic analysis of emergent Streptococcus pneumoniae serotype 22F causing invasive pneumococcal disease using whole genome sequencing.* PLoS One, 2017. **12**(5): p. e0178040.
18. Croucher, N.J., et al., *Rapid Pneumococcal Evolution in Response to Clinical Interventions.* Science, 2011. **331**(6016): p. 430-434.
19. Wyres, K.L., et al., *Pneumococcal Capsular Switching: A Historical Perspective.* The Journal of Infectious Diseases, 2013. **207**(3): p. 439-449.
20. Croucher, N.J., et al., *Population genomics of post-vaccine changes in pneumococcal epidemiology.* Nature genetics, 2013. **45**(6): p. 656-663.
21. Tettelin, H., et al., *Complete genome sequence of a virulent isolate of Streptococcus pneumoniae.* Science, 2001. **293**(5529): p. 498-506.
22. Hoskins, J., et al., *Genome of the Bacterium Streptococcus pneumoniae Strain R6.* Journal of Bacteriology, 2001. **183**(19): p. 5709-5717.
23. Brown, J.S., et al., *Immunization with components of two iron uptake ABC transporters protects mice against systemic Streptococcus pneumoniae infection.* Infect Immun, 2001. **69**(11): p. 6702-6.
24. Jomaa, M., et al., *Immunization with the iron uptake ABC transporter proteins PiaA and PiuA prevents respiratory infection with Streptococcus pneumoniae.* Vaccine, 2006. **24**(24): p. 5133-9.
25. Bentley, S.D., et al., *Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes.* PLoS Genet, 2006. **2**(3): p. e31.
26. Lloyd-Evans, N., et al., *Nasopharyngeal carriage of pneumococci in Gambian children and in their families.* Pediatr Infect Dis J, 1996. **15**(10): p. 866-71.
27. Hill, P.C., et al., *Nasopharyngeal carriage of Streptococcus pneumoniae in Gambian infants: a longitudinal study.* Clin Infect Dis, 2008. **46**(6): p. 807-14.
28. Chaguza, C., et al., *Understanding pneumococcal serotype 1 biology through population genomic analysis.* BMC Infect Dis, 2016. **16**(1): p. 649.
29. Brueggemann, A.B. and B.G. Spratt, *Geographic Distribution and Clonal Diversity of Streptococcus pneumoniae Serotype 1 Isolates.* Journal of Clinical Microbiology, 2003. **41**(11): p. 4966-4970.
30. Dawid, S., M.E. Sebert, and J.N. Weiser, *Bacteriocin Activity of Streptococcus pneumoniae Is Controlled by the Serine Protease HtrA via Posttranscriptional Regulation.* Journal of Bacteriology, 2009. **191**(5): p. 1509-1518.
31. Moore, M.R., et al., *Population Snapshot of Emergent Streptococcus pneumoniae Serotype 19A in the United States, 2005.* The Journal of Infectious Diseases, 2008. **197**(7): p. 1016-1027.
32. Paterson, G.K. and C.J. Orihuela, *Pneumococci: immunology of the innate host response.* Respirology (Carlton, Vic.), 2010. **15**(7): p. 1057-1063.
33. Tuomanen, E.I. and H.R. Masure, *Molecular and cellular biology of pneumococcal infection.* Microb Drug Resist, 1997. **3**(4): p. 297-308.
34. O'Brien, K.L., et al., *Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates.* The Lancet. **374**(9693): p. 893-902.
35. Malley, R. and P.W. Anderson, *Serotype-independent pneumococcal experimental vaccines that induce cellular as well as humoral immunity.* Proceedings of the National Academy of Sciences, 2012. **109**(10): p. 3623-3627.
36. Lu, Y.J., et al., *Interleukin-17A mediates acquired immunity to pneumococcal colonization.* PLoS Pathog, 2008. **4**(9): p. e1000159.

37. Zhang, Z., T.B. Clarke, and J.N. Weiser, *Cellular effectors mediating Th17-dependent clearance of pneumococcal colonization in mice.* The Journal of Clinical Investigation, 2009. **119**(7): p. 1899-1909.

38. Kadioglu, A., et al., *CD4-T-Lymphocyte Interactions with Pneumolysin and Pneumococci Suggest a Crucial Protective Role in the Host Response to Pneumococcal Infection.* Infection and Immunity, 2004. **72**(5): p. 2689-2697.

39. Malley, R., et al., *CD4+ T cells mediate antibody-independent acquired immunity to pneumococcal colonization.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(13): p. 4848-4853.

40. O'Brien, K.L., et al., *Burden of disease caused by <em>Streptococcus pneumoniae</em> in children younger than 5 years: global estimates.* The Lancet. **374**(9693): p. 893-902.

41. Inostroza, J., et al., *Capsular Serotype and Antibiotic Resistance of Streptococcus pneumoniae Isolates in Two Chilean Cities.* Clinical and Diagnostic Laboratory Immunology, 1998. **5**(2): p. 176-180.

42. Boken, D.J., et al., *Colonization with penicillin-nonsusceptible Streptococcus pneumoniae in urban and rural child-care centers.* Pediatr Infect Dis J, 1996. **15**(8): p. 667-72.

43. del Amo, E., et al., *Serotypes and Clonal Diversity of Streptococcus pneumoniae Causing Invasive Disease in the Era of PCV13 in Catalonia, Spain.* PLoS ONE, 2016. **11**(3): p. e0151125.

44. Forgie, I.M., et al., *Etiology of acute lower respiratory tract infections in Gambian children: II. Acute lower respiratory tract infection in children ages one to nine years presenting at the hospital.* Pediatr Infect Dis J, 1991. **10**(1): p. 42-7.

45. Lim, C., et al., *Epidemiology and burden of multidrug-resistant bacterial infection in a developing country.* eLife, 2016. **5**: p. e18082.

46. Chang, B., et al., *Capsule Switching and Antimicrobial Resistance Acquired during Repeated Streptococcus pneumoniae Pneumonia Episodes.* Journal of Clinical Microbiology, 2015. **53**(10): p. 3318-3324.

47. Welte, T., A. Torres, and D. Nathwani, *Clinical and economic burden of community-acquired pneumonia among adults in Europe.* Thorax, 2012. **67**(1): p. 71-9.

48. Ikeogu, M.O., *Acute pneumonia in Zimbabwe: bacterial isolates by lung aspiration.* Archives of Disease in Childhood, 1988. **63**(10): p. 1266-1267.

49. Brouwer, M.C., A.R. Tunkel, and D. van de Beek, *Epidemiology, Diagnosis, and Antimicrobial Treatment of Acute Bacterial Meningitis.* Clinical Microbiology Reviews, 2010. **23**(3): p. 467-492.

50. Nuoh, R.D., et al., *Review of meningitis surveillance data, upper West Region, Ghana 2009-2013.* The Pan African Medical Journal, 2016. **25**(Suppl 1): p. 9.

51. Leimkugel, J., et al., *An outbreak of serotype 1 Streptococcus pneumoniae meningitis in northern Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics.* J Infect Dis, 2005. **192**(2): p. 192-9.

52. Yaro, S., et al., *Epidemiological and Molecular Characteristics of a Highly Lethal Pneumococcal Meningitis Epidemic in Burkina Faso.* Clinical Infectious Diseases, 2006. **43**(6): p. 693-700.

53. Molyneux, E., F.A. Riordan, and A. Walsh, *Acute bacterial meningitis in children presenting to the Royal Liverpool Children's Hospital, Liverpool, UK and the Queen*

*Elizabeth Central Hospital in Blantyre, Malawi: a world of difference.* Ann Trop Paediatr, 2006. **26**(1): p. 29-37.

54. Kwambana-Adams, B.A., et al., *An outbreak of pneumococcal meningitis among older children (≥5 years) and adults after the implementation of an infant vaccination programme with the 13-valent pneumococcal conjugate vaccine in Ghana.* BMC Infectious Diseases, 2016. **16**(1): p. 575.

55. Traore, Y., et al., *Incidence, Seasonality, Age Distribution, and Mortality of Pneumococcal Meningitis in Burkina Faso and Togo.* Clinical Infectious Diseases, 2009. **48**(Supplement_2): p. S181-S189.

56. The Kenyan Bacteraemia Study, G., et al., *Polymorphism in a lincRNA Associates with a Doubled Risk of Pneumococcal Bacteremia in Kenyan Children.* American Journal of Human Genetics, 2016. **98**(6): p. 1092-1100.

57. Laupland, K.B., *Incidence of bloodstream infection: a review of population-based studies.* Clin Microbiol Infect, 2013. **19**(6): p. 492-500.

58. Berkley , J.A., et al., *Bacteremia among Children Admitted to a Rural Hospital in Kenya.* New England Journal of Medicine, 2005. **352**(1): p. 39-47.

59. Rosenblut, A., et al., *Etiology of acute otitis media and serotype distribution of Streptococcus pneumoniae and Haemophilus influenzae in Chilean children <5 years of age.* Medicine, 2017. **96**(6): p. e5974.

60. Paradise, J.L., et al., *Language, speech sound production, and cognition in three-year-old children in relation to otitis media in their first three years of life.* Pediatrics, 2000. **105**(5): p. 1119-30.

61. Valenzuela, M.T., et al., *The burden of pneumococcal disease among Latin American and Caribbean children: review of the evidence.* Rev Panam Salud Publica, 2009. **25**(3): p. 270-9.

62. Aguilar, L., et al., *Microbiology of the middle ear fluid in Costa Rican children between 2002 and 2007.* Int J Pediatr Otorhinolaryngol, 2009. **73**(10): p. 1407-11.

63. Beekmann, S.E., et al., *Antimicrobial resistance in Streptococcus pneumoniae, Haemophilus influenzae, Moraxella catarrhalis and group A β-haemolytic streptococci in 2002–2003: Results of the multinational GRASP Surveillance Program.* International Journal of Antimicrobial Agents, 2005. **25**(2): p. 148-156.

64. Friedland, I.R. and K.P. Klugman, *Antibiotic-resistant pneumococcal disease in South African children.* Am J Dis Child, 1992. **146**(8): p. 920-3.

65. Jedrzejas, M.J., *Pneumococcal Virulence Factors: Structure and Function.* Microbiology and Molecular Biology Reviews, 2001. **65**(2): p. 187-207.

66. Adegbola, R.A., et al., *Serotype and antimicrobial susceptibility patterns of isolates of Streptococcus pneumoniae causing invasive disease in The Gambia 1996-2003.* Trop Med Int Health, 2006. **11**(7): p. 1128-35.

67. *Pneumococcal vaccines WHO position paper - 2012 - recommendations.* Vaccine, 2012. **30**(32): p. 4717-8.

68. Tai, S.S., *Streptococcus pneumoniae protein vaccine candidates: properties, activities and animal studies.* Crit Rev Microbiol, 2006. **32**(3): p. 139-53.

69. Rigden, D.J., M.Y. Galperin, and M.J. Jedrzejas, *Analysis of structure and function of putative surface-exposed proteins encoded in the Streptococcus pneumoniae genome: a bioinformatics-based approach to vaccine and drug design.* Crit Rev Biochem Mol Biol, 2003. **38**(2): p. 143-68.

70. Pavia, M., et al., *Efficacy of pneumococcal vaccination in children younger than 24 months: a meta-analysis.* Pediatrics, 2009. **123**(6): p. e1103-10.
71. Picazo, J., et al., *Effect of the different 13-valent pneumococcal conjugate vaccination uptakes on the invasive pneumococcal disease in children: Analysis of a hospital-based and population-based surveillance study in Madrid, Spain, 2007-2015.* PLoS One, 2017. **12**(2): p. e0172222.
72. Silfverdal, S.A., et al., *Immunogenicity of a 2-dose priming and booster vaccination with the 10-valent pneumococcal nontypeable Haemophilus influenzae protein D conjugate vaccine.* Pediatr Infect Dis J, 2009. **28**(10): p. e276-82.
73. Dagan, R., et al., *Efficacy of 13-valent pneumococcal conjugate vaccine (PCV13) versus that of 7-valent PCV (PCV7) against nasopharyngeal colonization of antibiotic-nonsusceptible Streptococcus pneumoniae.* J Infect Dis, 2015. **211**(7): p. 1144-53.
74. Kwambana-Adams, B.A., et al., *An outbreak of pneumococcal meningitis among older children (>/=5 years) and adults after the implementation of an infant vaccination programme with the 13-valent pneumococcal conjugate vaccine in Ghana.* BMC Infect Dis, 2016. **16**(1): p. 575.
75. Mackenzie, G.A., et al., *Effect of the introduction of pneumococcal conjugate vaccination on invasive pneumococcal disease in The Gambia: a population-based surveillance study.* Lancet Infect Dis, 2016. **16**(6): p. 703-11.
76. Huang, S.S., et al., *Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004.* Pediatrics, 2005. **116**(3): p. e408-13.
77. Pirez, M.C., et al., *Impact of universal pneumococcal vaccination on hospitalizations for pneumonia and meningitis in children in Montevideo, Uruguay.* Pediatr Infect Dis J, 2011. **30**(8): p. 669-74.
78. Weatherholtz, R., et al., *Invasive Pneumococcal Disease a Decade after Pneumococcal Conjugate Vaccine Use in an American Indian Population at High Risk for Disease.* Clinical Infectious Diseases, 2010. **50**(9): p. 1238-1246.
79. Richter, S.S., et al., *Changes in Pneumococcal Serotypes and Antimicrobial Resistance after Introduction of the 13-Valent Conjugate Vaccine in the United States.* Antimicrobial Agents and Chemotherapy, 2014. **58**(11): p. 6484-6489.
80. Wizemann, T.M., et al., *Use of a whole genome approach to identify vaccine molecules affording protection against Streptococcus pneumoniae infection.* Infect Immun, 2001. **69**(3): p. 1593-8.
81. Yamaguchi, K., F. Yu, and M. Inouye, *A single amino acid determinant of the membrane localization of lipoproteins in E. coli.* Cell, 1988. **53**(3): p. 423-432.
82. Juncker, A.S., et al., *Prediction of lipoprotein signal peptides in Gram-negative bacteria.* Protein Science, 2003. **12**(8): p. 1652-1662.
83. Jomaa, M., et al., *Antibodies to the iron uptake ABC transporter lipoproteins PiaA and PiuA promote opsonophagocytosis of Streptococcus pneumoniae.* Infect Immun, 2005. **73**(10): p. 6852-9.
84. Odutola, A., et al., *Efficacy of a novel, protein-based pneumococcal vaccine against nasopharyngeal carriage of Streptococcus pneumoniae in infants: A phase 2, randomized, controlled, observer-blind study.* Vaccine, 2017. **35**(19): p. 2531-2542.
85. Kaur, R., et al., *Human Antibodies to PhtD, PcpA, and Ply Reduce Adherence to Human Lung Epithelial Cells and Murine Nasopharyngeal Colonization by Streptococcus pneumoniae.* Infection and Immunity, 2014. **82**(12): p. 5069-5075.

86. Verhoeven, D., Q. Xu, and M.E. Pichichero, *Vaccination with a Streptococcus pneumoniae trivalent recombinant PcpA, PhtD and PlyD1 protein vaccine candidate protects against lethal pneumonia in an infant murine model.* Vaccine, 2014. **32**(26): p. 3205-3210.
87. Goulart, C., et al., *Recombinant BCG expressing a PspA-PdT fusion protein protects mice against pneumococcal lethal challenge in a prime-boost strategy.* Vaccine, 2017. **35**(13): p. 1683-1691.
88. Ren, B., et al., *The Absence of PspA or Presence of Antibody to PspA Facilitates the Complement-Dependent Phagocytosis of Pneumococci In Vitro.* Clinical and Vaccine Immunology, 2012. **19**(10): p. 1574-1582.
89. *Current status and perspectives on protein-based pneumococcal vaccines.* Critical Reviews in Microbiology, 2015. **41**(2): p. 190-200.
90. Khan, M.N., et al., *PcpA of Streptococcus pneumoniae mediates adherence to nasopharyngeal and lung epithelial cells and elicits functional antibodies in humans.* Microbes and Infection, 2012. **14**(12): p. 1102-1110.
91. Bologa, M., et al., *Safety and immunogenicity of pneumococcal protein vaccine candidates: Monovalent choline-binding protein A (PcpA) vaccine and bivalent PcpA–pneumococcal histidine triad protein D vaccine.* Vaccine, 2012. **30**(52): p. 7461-7468.
92. Pimenta, F.C., et al., *Intranasal Immunization with the Cholera Toxin B Subunit-Pneumococcal Surface Antigen A Fusion Protein Induces Protection against Colonization with Streptococcus pneumoniae and Has Negligible Impact on the Nasopharyngeal and Oral Microbiota of Mice.* Infection and Immunity, 2006. **74**(8): p. 4939-4944.
93. Glover, D.T., S.K. Hollingshead, and D.E. Briles, *Streptococcus pneumoniae Surface Protein PcpA Elicits Protection against Lung Infection and Fatal Sepsis.* Infection and Immunity, 2008. **76**(6): p. 2767-2776.
94. Feil, E.J. and B.G. Spratt, *Recombination and the Population Structures of Bacterial Pathogens.* Annual Review of Microbiology, 2001. **55**(1): p. 561-590.
95. Croucher, N.J., et al., *Bacterial genomes in epidemiology—present and future.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2013. **368**(1614): p. 20120202.
96. Croucher, N.J., et al., *Rapid pneumococcal evolution in response to clinical interventions.* Science, 2011. **331**(6016): p. 430-4.
97. Dowson, C.G., et al., *Horizontal gene transfer and the evolution of resistance and virulence determinants in Streptococcus.* Soc Appl Bacteriol Symp Ser, 1997. **26**: p. 42s-51s.
98. Argondizzo, A.P.C., et al., *Identification of Proteins in Streptococcus pneumoniae by Reverse Vaccinology and Genetic Diversity of These Proteins in Clinical Isolates.* Applied Biochemistry and Biotechnology, 2015. **175**(4): p. 2124-2165.
99. Reglier-Poupet, H., et al., *Maturation of lipoproteins by type II signal peptidase is required for phagosomal escape of Listeria monocytogenes.* J Biol Chem, 2003. **278**(49): p. 49469-77.
100. Hammerschmidt, S., et al., *Identification of Pneumococcal Surface Protein A as a Lactoferrin-Binding Protein of Streptococcus pneumoniae.* Infection and Immunity, 1999. **67**(4): p. 1683-1687.

101. Plumptre, C.D., et al., *AdcA and AdcAII employ distinct zinc acquisition mechanisms and contribute additively to zinc homeostasis in Streptococcus pneumoniae.* Molecular Microbiology, 2014. **91**(4): p. 834-851.

102. Maeda, Y., et al., *Novel 33-kilodalton lipoprotein from Mycobacterium leprae.* Infect Immun, 2002. **70**(8): p. 4106-11.

103. Sutcliffe, I.C. and R.R. Russell, *Lipoproteins of gram-positive bacteria.* Journal of Bacteriology, 1995. **177**(5): p. 1123-1128.

104. Countrymeters, *Gambia Population*
. 2017.

105. Ceesay, S.J., et al., *Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis.* The Lancet. **372**(9649): p. 1545-1554.

106. O'Brien, K.L. and H. Nohynek, *Report from a WHO Working Group: standard method for detecting upper respiratory carriage of Streptococcus pneumoniae.* Pediatr Infect Dis J, 2003. **22**(2): p. e1-11.

107. Page, A.J., et al., *Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data.* Microb Genom, 2016. **2**(8): p. e000083.

108. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.

109. Boetzer, M., et al., *Scaffolding pre-assembled contigs using SSPACE.* Bioinformatics, 2011. **27**(4): p. 578-9.

110. Boetzer, M. and W. Pirovano, *Toward almost closed genomes with GapFiller.* Genome Biology, 2012. **13**(6): p. R56.

111. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-9.

112. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biology, 2014. **15**(3): p. R46.

113. Page, A.J., Taylor B., Keane J.A., *Multilocus sequence typing by blast from de novo assemblies against PubMLST.* The Jounal of Open Source Software, 2016.

114. Kapatai, G., et al., *Whole genome sequencing of Streptococcus pneumoniae: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline.* PeerJ, 2016. **4**: p. e2477.

115. Cheng, L., et al., *Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software.* Molecular Biology and Evolution, 2013. **30**(5): p. 1224-1228.

116. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.

117. Page, A.J., et al., *SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments.* Microbial Genomics, 2016. **2**(4).

118. Page, A.J., et al., *Roary: rapid large-scale prokaryote pan genome analysis.* Bioinformatics, 2015.

119. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.

120. de Castro, E., et al., *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.* Nucleic Acids Research, 2006. **34**(suppl_2): p. W362-W365.

121. Sutcliffe, I.C. and D.J. Harrington, *Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes.* Microbiology, 2002. **148**(Pt 7): p. 2065-77.

122. Rahman, O., et al., *Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria.* World Journal of Microbiology and Biotechnology, 2008. **24**(11): p. 2377.
123. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**(1): p. 421.
124. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
125. Käll, L., A. Krogh, and E.L.L. Sonnhammer, *A Combined Transmembrane Topology and Signal Peptide Prediction Method.* Journal of Molecular Biology, 2004. **338**(5): p. 1027-1036.
126. Madan Babu, M. and K. Sankaran, *DOLOP--database of bacterial lipoproteins.* Bioinformatics, 2002. **18**(4): p. 641-3.
127. Babu, M.M., et al., *A Database of Bacterial Lipoproteins (DOLOP) with Functional Assignments to Predicted Lipoproteins.* Journal of Bacteriology, 2006. **188**(8): p. 2761-2773.
128. Terada, M., et al., *Lipoprotein sorting signals evaluated as the LolA-dependent release of lipoproteins from the cytoplasmic membrane of Escherichia coli.* J Biol Chem, 2001. **276**(50): p. 47690-4.
129. Gouy, M., S. Guindon, and O. Gascuel, *SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building.* Molecular Biology and Evolution, 2010. **27**(2): p. 221-224.
130. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
131. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313.
132. Hadfield, J., et al., *Phandango: an interactive viewer for bacterial population genomics.* bioRxiv, 2017.
133. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.* Nucleic Acids Research, 2016. **44**(Web Server issue): p. W242-W245.
134. Haste Andersen, P., M. Nielsen, and O. Lund, *Prediction of residues in discontinuous B-cell epitopes using protein 3D structures.* Protein Sci, 2006. **15**(11): p. 2558-67.
135. Larsen, J.E.P., O. Lund, and M. Nielsen, *Improved method for predicting linear B-cell epitopes.* Immunome Research, 2006. **2**: p. 2-2.
136. Parker, J.M., D. Guo, and R.S. Hodges, *New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites.* Biochemistry, 1986. **25**(19): p. 5425-32.
137. Hopp, T.P. and K.R. Woods, *Prediction of protein antigenic determinants from amino acid sequences.* Proceedings of the National Academy of Sciences of the United States of America, 1981. **78**(6): p. 3824-3828.
138. Welling, G.W., et al., *Prediction of sequential antigenic regions in proteins.* FEBS Lett, 1985. **188**(2): p. 215-8.
139. Levitt, M., *Conformational preferences of amino acids in globular proteins.* Biochemistry, 1978. **17**(20): p. 4277-85.

140. Pellequer, J.L., E. Westhof, and M.H. Van Regenmortel, *Correlation between the location of antigenic sites and the prediction of turns in proteins.* Immunol Lett, 1993. **36**(1): p. 83-99.

141. Chou, P.Y. and G.D. Fasman, *Prediction of the secondary structure of proteins from their amino acid sequence.* Adv Enzymol Relat Areas Mol Biol, 1978. **47**: p. 45-148.

142. Karplus, P.A. and G.E. Schulz, *Prediction of chain flexibility in proteins.* Naturwissenschaften, 1985. **72**(4): p. 212-213.

143. Ponomarenko, J., et al., *ElliPro: a new structure-based tool for the prediction of antibody epitopes.* BMC Bioinformatics, 2008. **9**(1): p. 514.

144. Thornton, J.M., et al., *Location of 'continuous' antigenic determinants in the protruding regions of proteins.* The EMBO Journal, 1986. **5**(2): p. 409-413.

145. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.

146. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nat Methods, 2015. **12**(1): p. 7-8.

147. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nature protocols, 2010. **5**(4): p. 725-738.

148. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.

149. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nat. Protocols, 2015. **10**(6): p. 845-858.

150. Fiser, A. and A. Šali, *Modeller: Generation and Refinement of Homology-Based Protein Structure Models*, in *Methods in Enzymology*. 2003, Academic Press. p. 461-491.

151. Geno, K.A., J.S. Saad, and M.H. Nahm, *Discovery of novel pneumococcal serotype, 35D: a natural WciG-deficient variant of serotype 35B.* Journal of Clinical Microbiology, 2017.

152. *Jmol: an open-source Java viewerfor chemical structures in 3D.*

153. Romero-Steiner, S., et al., *Inhibition of Pneumococcal Adherence to Human Nasopharyngeal Epithelial Cells by Anti-PsaA Antibodies.* Clinical and Diagnostic Laboratory Immunology, 2003. **10**(2): p. 246-251.

154. Claverys, J.P., *A new family of high-affinity ABC manganese and zinc permeases.* Res Microbiol, 2001. **152**(3-4): p. 231-43.

155. Moore, C.M. and J.D. Helmann, *Metal ion homeostasis in Bacillus subtilis.* Curr Opin Microbiol, 2005. **8**(2): p. 188-95.

156. Loisel, E., et al., *AdcAII, A New Pneumococcal Zn-Binding Protein Homologous with ABC Transporters: Biochemical and Structural Analysis.* Journal of Molecular Biology, 2008. **381**(3): p. 594-606.

157. Dintilhac, A., et al., *Competence and virulence of Streptococcus pneumoniae: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases.* Mol Microbiol, 1997. **25**(4): p. 727-39.

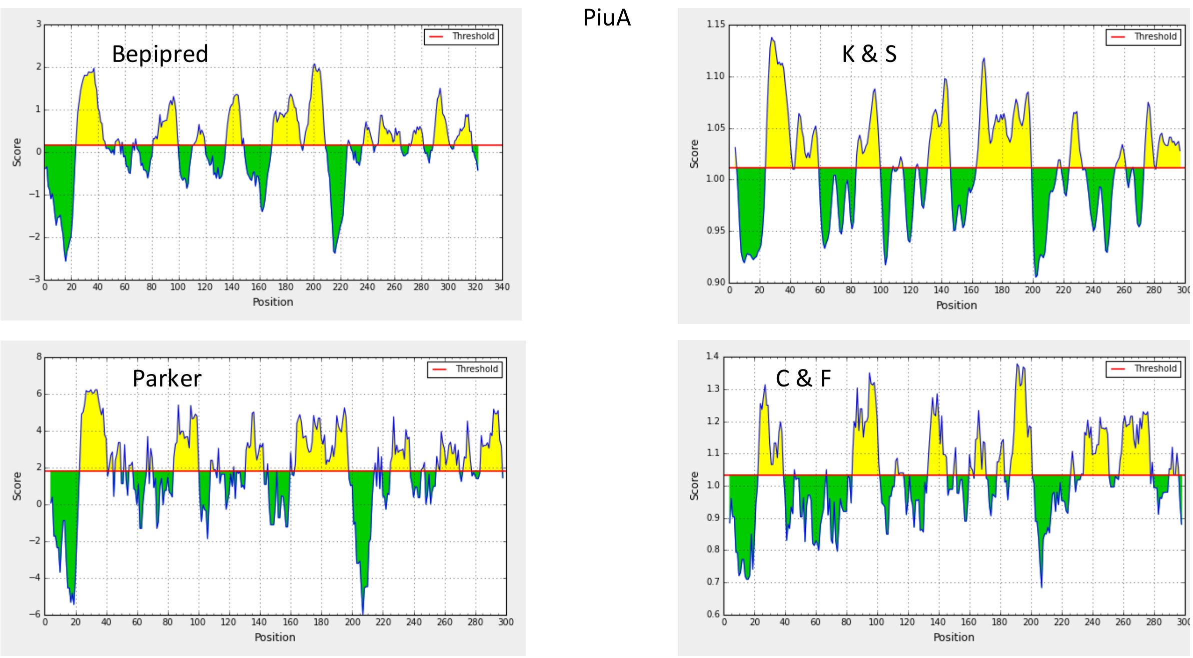
158. Bayle, L., et al., *Zinc uptake by Streptococcus pneumoniae depends on both AdcA and AdcAII and is essential for normal bacterial morphology and virulence.* Mol Microbiol, 2011. **82**(4): p. 904-16.

159. Brown, L.R., et al., *AdcAII of Streptococcus pneumoniae Affects Pneumococcal Invasiveness.* PLOS ONE, 2016. **11**(1): p. e0146785.

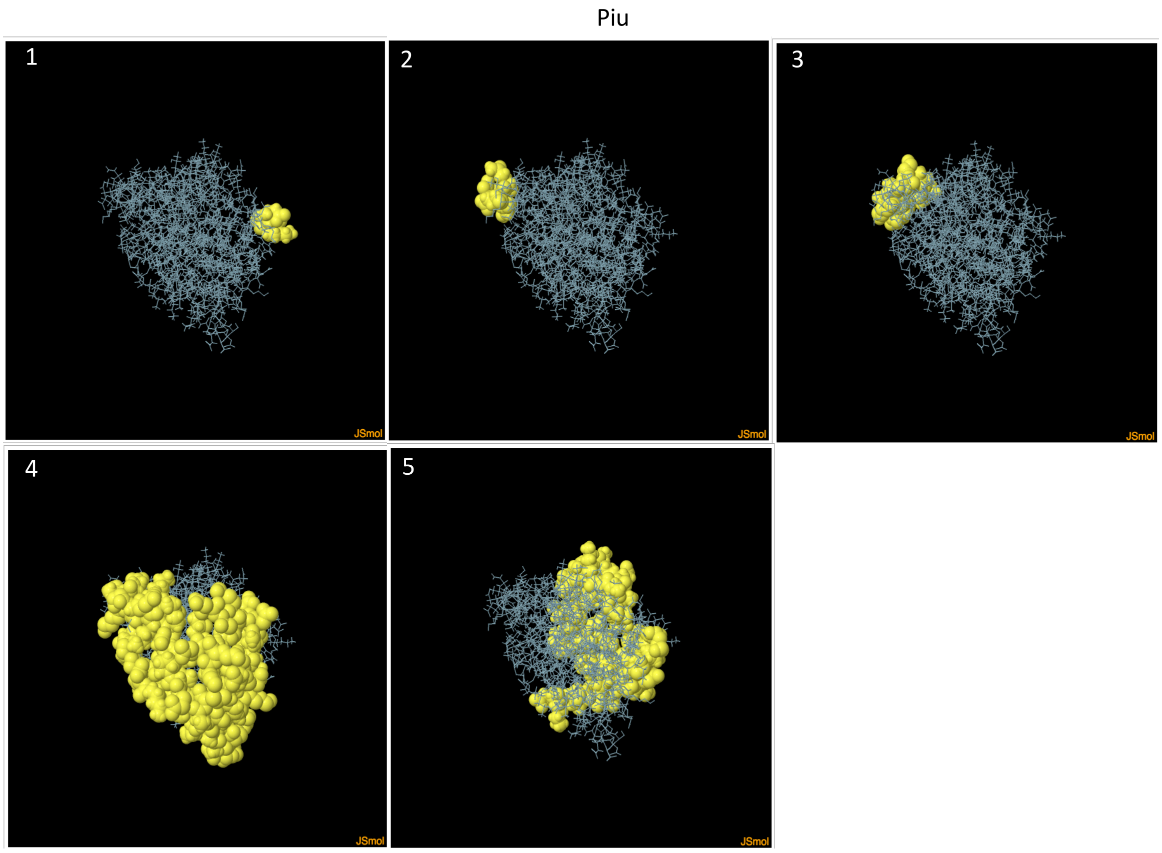
160. Berry, A.M. and J.C. Paton, *Sequence heterogeneity of PsaA, a 37-kilodalton putative adhesin essential for virulence of Streptococcus pneumoniae.* Infect Immun, 1996. **64**(12): p. 5255-62.
161. Wilson, R., et al., *Protection against Streptococcus pneumoniae lung infection after nasopharyngeal colonization requires both humoral and cellular immune responses.* Mucosal Immunology, 2015. **8**(3): p. 627-639.
162. Ogunniyi, A.D., et al., *Immunization of mice with combinations of pneumococcal virulence proteins elicits enhanced protection against challenge with Streptococcus pneumoniae.* Infect Immun, 2000. **68**(5): p. 3028-33.
163. Gor, D.O., et al., *Enhanced immunogenicity of pneumococcal surface adhesin A by genetic fusion to cytokines and evaluation of protective immunity in mice.* Infect Immun, 2002. **70**(10): p. 5589-95.
164. Sampson, J.S., et al., *Limited diversity of Streptococcus pneumoniae psaA among pneumococcal vaccine serotypes.* Infect Immun, 1997. **65**(5): p. 1967-71.
165. Goodell, E.W. and C.F. Higgins, *Uptake of cell wall peptides by Salmonella typhimurium and Escherichia coli.* Journal of Bacteriology, 1987. **169**(8): p. 3861-3865.
166. Claverys, J.-P., B. Grossiord, and G. Alloing, *Is the Ami-AliA/B oligopeptide permease of Streptococcus pneumoniae involved in sensing environmental conditions?* Research in Microbiology, 2000. **151**(6): p. 457-463.
167. Alloing, G., M.C. Trombe, and J.P. Claverys, *The ami locus of the Gram-positive bacterium Streptococcus pneumoniae is similar to binding protein-dependent transport operons of Gram-negative bacteria.* Molecular Microbiology, 1990. **4**(4): p. 633-644.
168. Alloing, G., P. de Philip, and J.P. Claverys, *Three highly homologous membrane-bound lipoproteins participate in oligopeptide transport by the Ami system of the gram-positive Streptococcus pneumoniae.* J Mol Biol, 1994. **241**(1): p. 44-58.
169. Gilson, E., et al., *Evidence for high affinity binding-protein dependent transport systems in gram-positive bacteria and in Mycoplasma.* Embo j, 1988. **7**(12): p. 3971-4.
170. Abbott, D.W., et al., *The molecular basis of glycogen breakdown and transport in Streptococcus pneumoniae.* Molecular Microbiology, 2010. **77**(1): p. 183-199.
171. Orihuela, C.J., et al., *Microarray Analysis of Pneumococcal Gene Expression during Invasive Disease.* Infection and Immunity, 2004. **72**(10): p. 5582-5596.
172. Yang, X.Y., et al., *Integrated Translatomics with Proteomics to Identify Novel Iron-Transporting Proteins in Streptococcus pneumoniae.* Front Microbiol, 2016. **7**: p. 78.
173. Brown, J.S., et al., *Characterization of Pit, a Streptococcus pneumoniae Iron Uptake ABC Transporter.* Infection and Immunity, 2002. **70**(8): p. 4389-4398.
174. Brown, J.S., S.M. Gilliland, and D.W. Holden, *A Streptococcus pneumoniae pathogenicity island encoding an ABC transporter involved in iron uptake and virulence.* Mol Microbiol, 2001. **40**(3): p. 572-85.
175. Whalan, R.H., et al., *Distribution and genetic diversity of the ABC transporter lipoproteins PiuA and PiaA within Streptococcus pneumoniae and related streptococci.* J Bacteriol, 2006. **188**(3): p. 1031-8.
176. Croucher, N.J., et al., *Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening.* Proc Natl Acad Sci U S A, 2017. **114**(3): p. E357-e366.

177. *UniProt: the universal protein knowledgebase.* Nucleic Acids Research, 2017. **45**(D1): p. D158-D169.
178. Scientific, T., *Amino Acid Physical Properties.* 2017.
179. Culurgioni, S., M. Tang, and M.A. Walsh, *Structural characterization of the Streptococcus pneumoniae carbohydrate substrate-binding protein SP0092.* Acta crystallographica. Section F, Structural biology communications, 2017. **73**(Pt 1): p. 54-61.
180. Bidossi, A., et al., *A functional genomics approach to establish the complement of carbohydrate transporters in Streptococcus pneumoniae.* PLoS One, 2012. **7**(3): p. e33320.
181. I. Sillitoe, N.D., T. Lewis, D. Lee, J. Lees, C. Orengo, *CATH: Protein Structure Classification Database*. 2017.
182. Novak, R., et al., *Identification of a Streptococcus pneumoniae Gene Locus Encoding Proteins of an ABC Phosphate Transporter and a Two-Component Regulatory System.* Journal of Bacteriology, 1999. **181**(4): p. 1126-1133.
183. Soualhine, H., et al., *A proteomic analysis of penicillin resistance in Streptococcus pneumoniae reveals a novel role for PstS, a subunit of the phosphate ABC transporter.* Mol Microbiol, 2005. **58**(5): p. 1430-40.
184. Abdullah, M.R., et al., *Structure of the pneumococcal l,d-carboxypeptidase DacB and pathophysiological effects of disabled cell wall hydrolases DacA and DacB.* Mol Microbiol, 2014. **93**(6): p. 1183-206.
185. Barendt, S.M., L.-T. Sham, and M.E. Winkler, *Characterization of Mutants Deficient in the l,d-Carboxypeptidase (DacB) and WalRK (VicRK) Regulon, Involved in Peptidoglycan Maturation of Streptococcus pneumoniae Serotype 2 Strain D39.* Journal of Bacteriology, 2011. **193**(9): p. 2290-2300.
186. Deka, R.K., et al., *The PnrA (Tp0319; TmpC) lipoprotein represents a new family of bacterial purine nucleoside receptor encoded within an ATP-binding cassette (ABC)-like operon in Treponema pallidum.* J Biol Chem, 2006. **281**(12): p. 8072-81.
187. Saleh, M., et al., *Molecular architecture of &lt;em&gt;Streptococcus pneumoniae&lt;/em&gt; surface thioredoxin-fold lipoproteins crucial for extracellular oxidative stress resistance and maintenance of virulence.* EMBO Molecular Medicine, 2013. **5**(12): p. 1852.
188. Whalan, R.H., et al., *PiuA and PiaA, iron uptake lipoproteins of Streptococcus pneumoniae, elicit serotype independent antibody responses following human pneumococcal septicaemia.* FEMS Immunol Med Microbiol, 2005. **43**(1): p. 73-80.
189. Giefing, C., et al., *Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies.* The Journal of Experimental Medicine, 2008. **205**(1): p. 117-131.

# Appendix



## Figure A1 Linear epitope predictions for PiuA.



## Figure A2 ElliPro predicted discontinuous epitopes for PiuA.

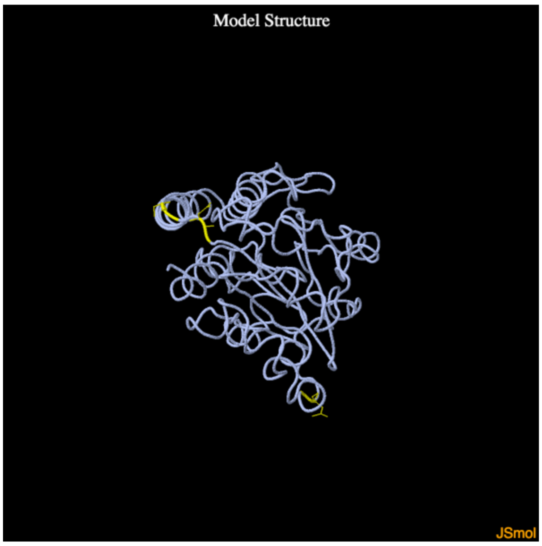


**Figure A3 DiscoTope2 predicted discontinuous epitopes for PiuA.**

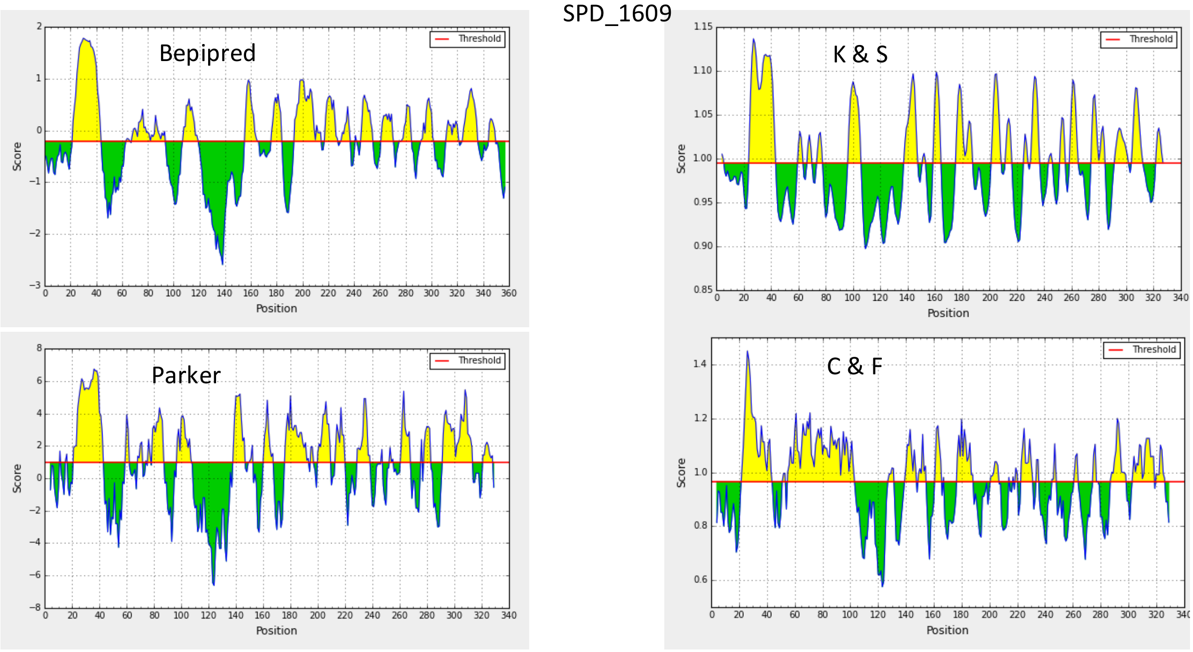


**Figure A4 Linear epitope prediction of the SPD_1609 protein.**

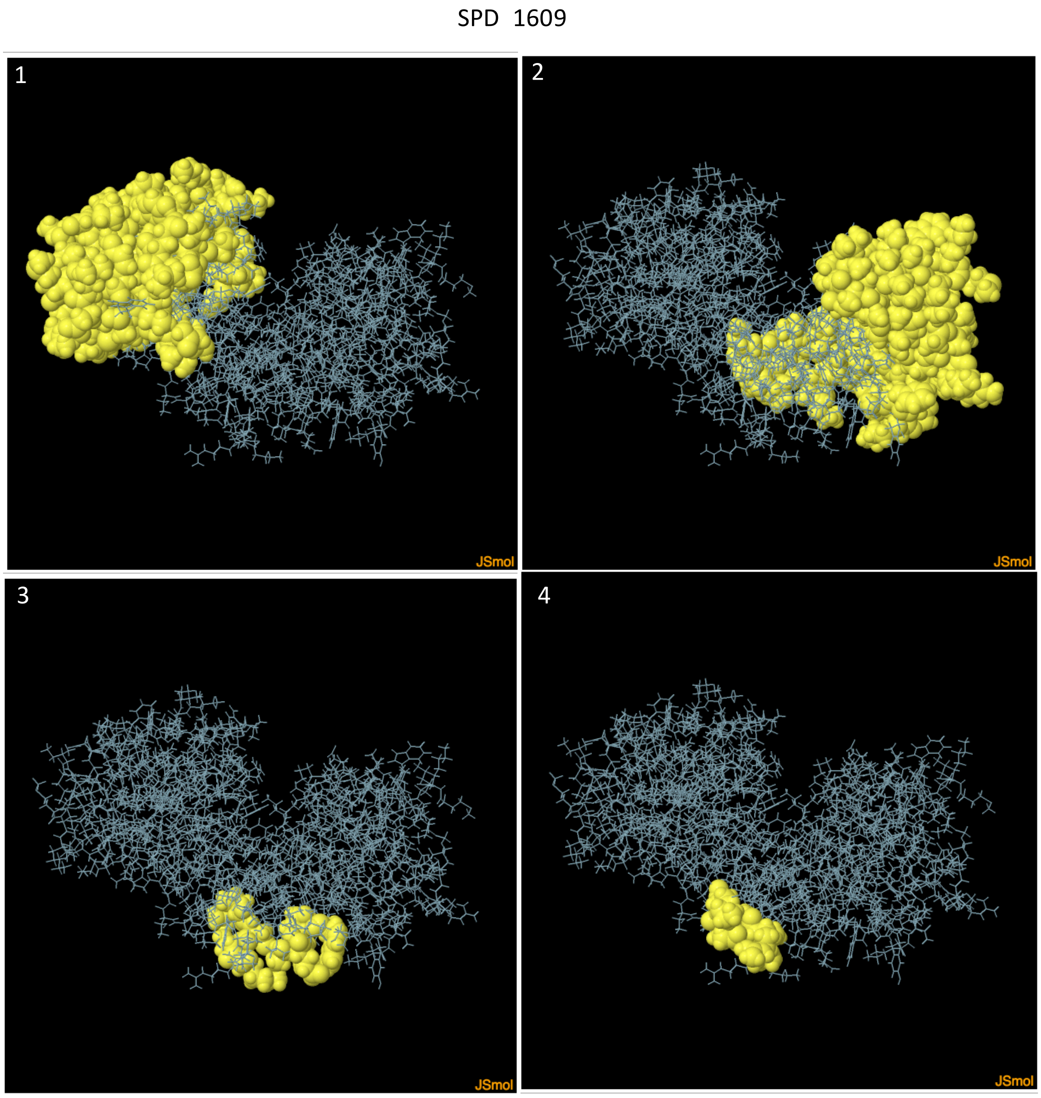


**Figure A5 ElliPro predicted discontinuous epitopes for SPD_1609.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues part of the predicted epitope.

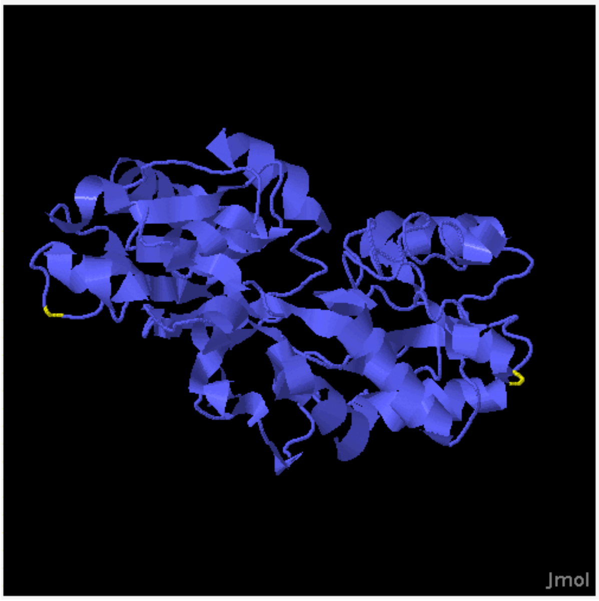


**Figure A6 DiscoTope2 predicted discontinuous epitopes for SPD_1609.**

The parts coloured yellow are the predicted epitopes.

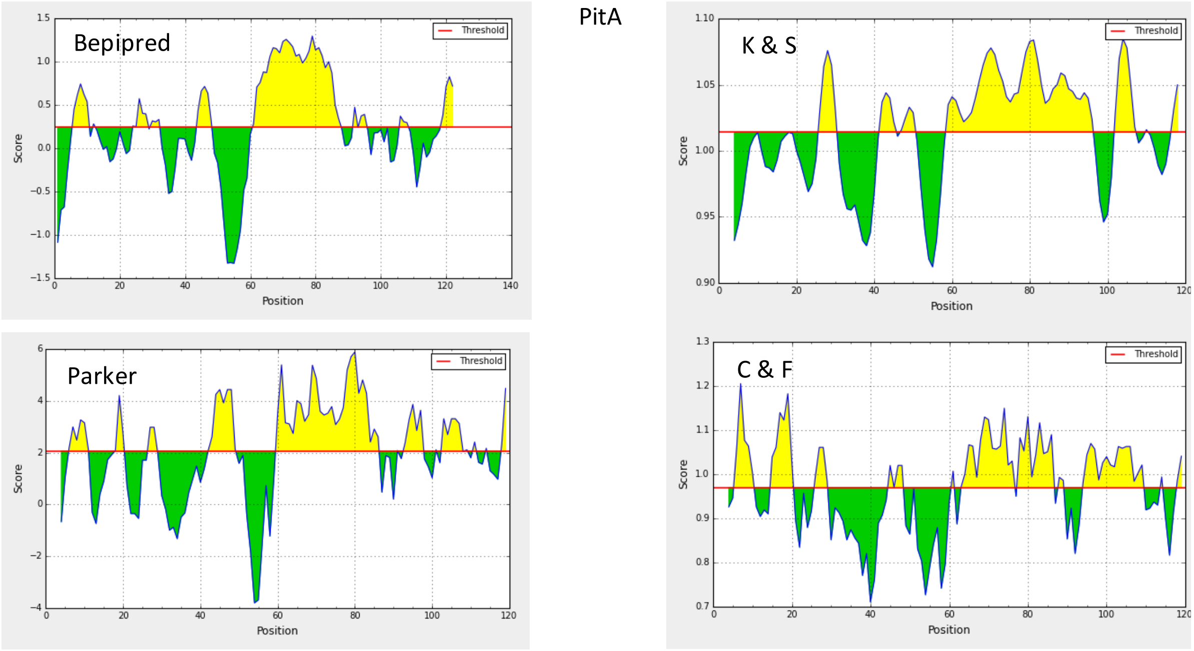


**Figure A7 Linear epitope predictions of the PitA protein.**

Pit

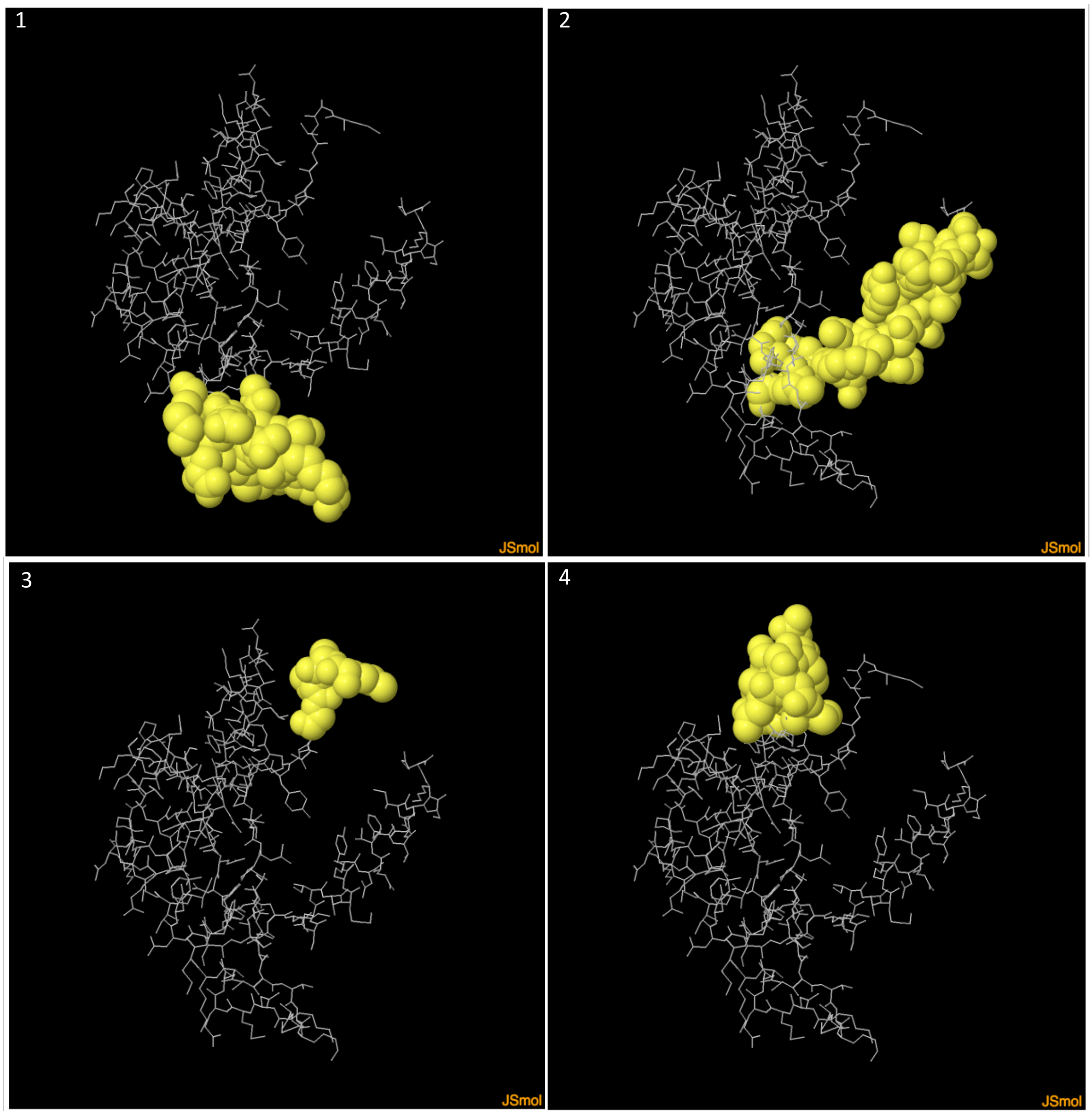


**Figure A8 ElliPro predicted discontinuous epitopes for PitA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with one having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

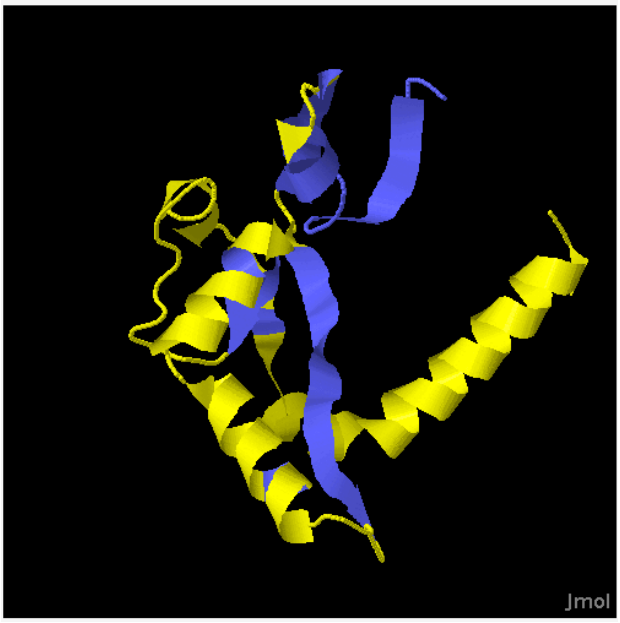


**Figure A9 DiscoTope2 predicted discontinuous epitopes for PitA.**

The parts coloured yellow are the predicted epitopes.

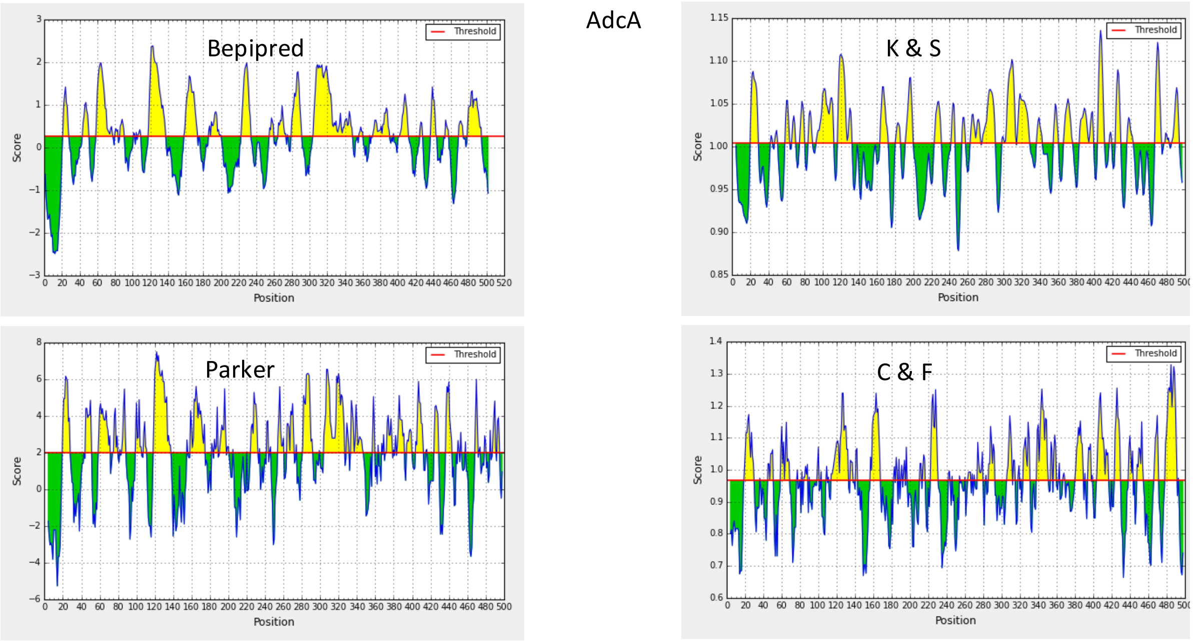


**Figure A10 Linear epitope predictions of the AdcA protein.**

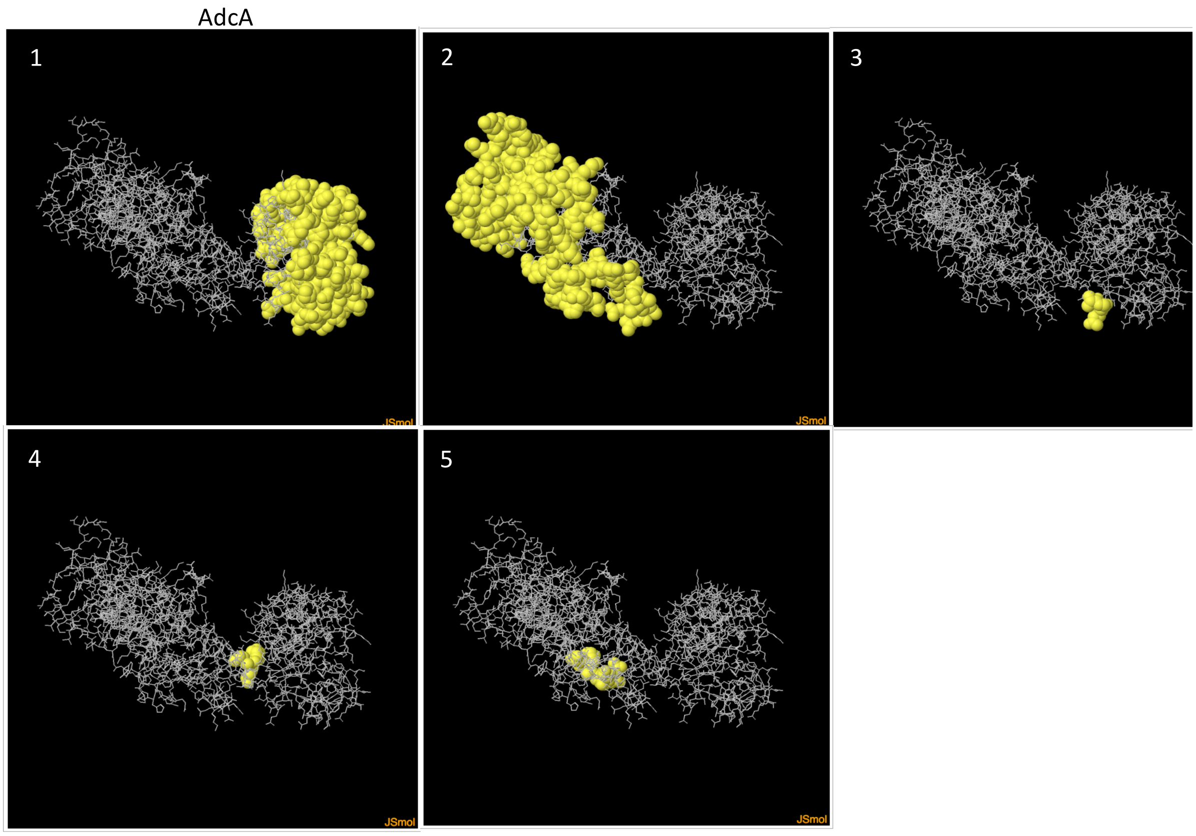


**Figure A11 ElliPro predicted discontinuous epitopes for AdcA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

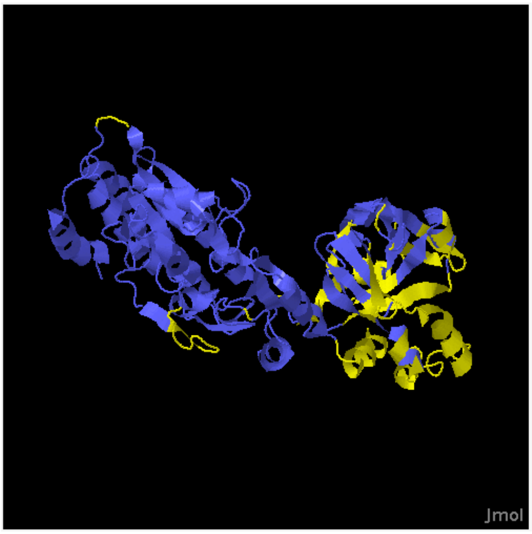


**Figure A12 DiscoTope2 predicted discontinuous epitopes for AdcA.**

The parts coloured yellow are the predicted epitopes.

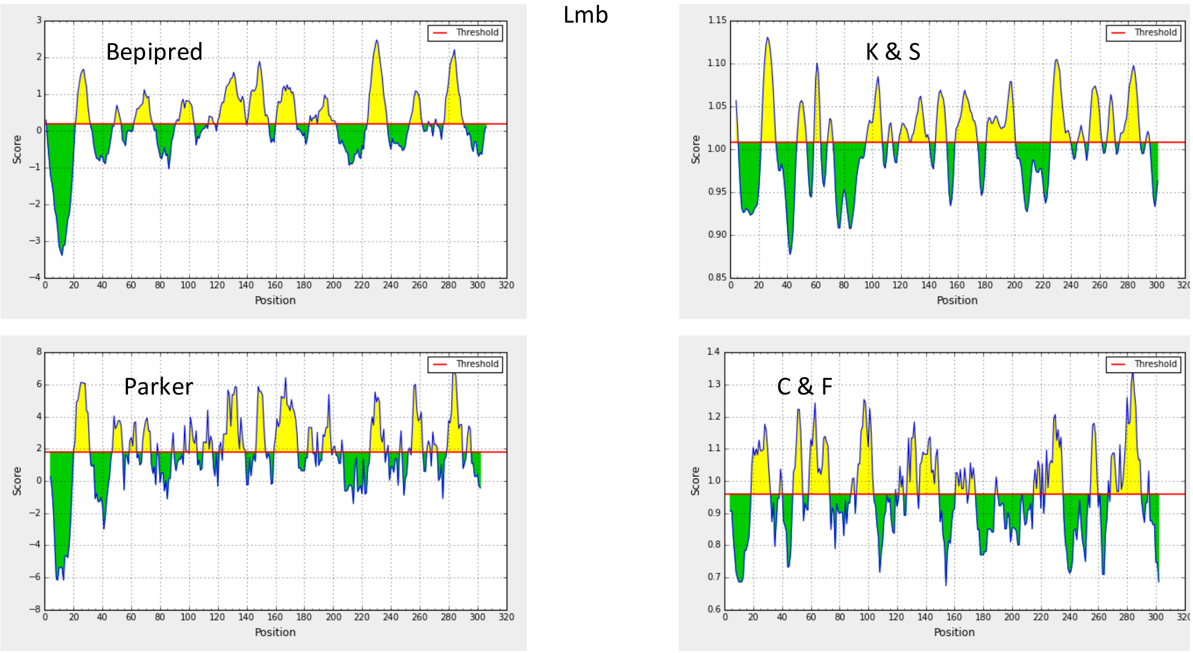


**Figure A13 Linear epitope predictions of the Lmb protein.**

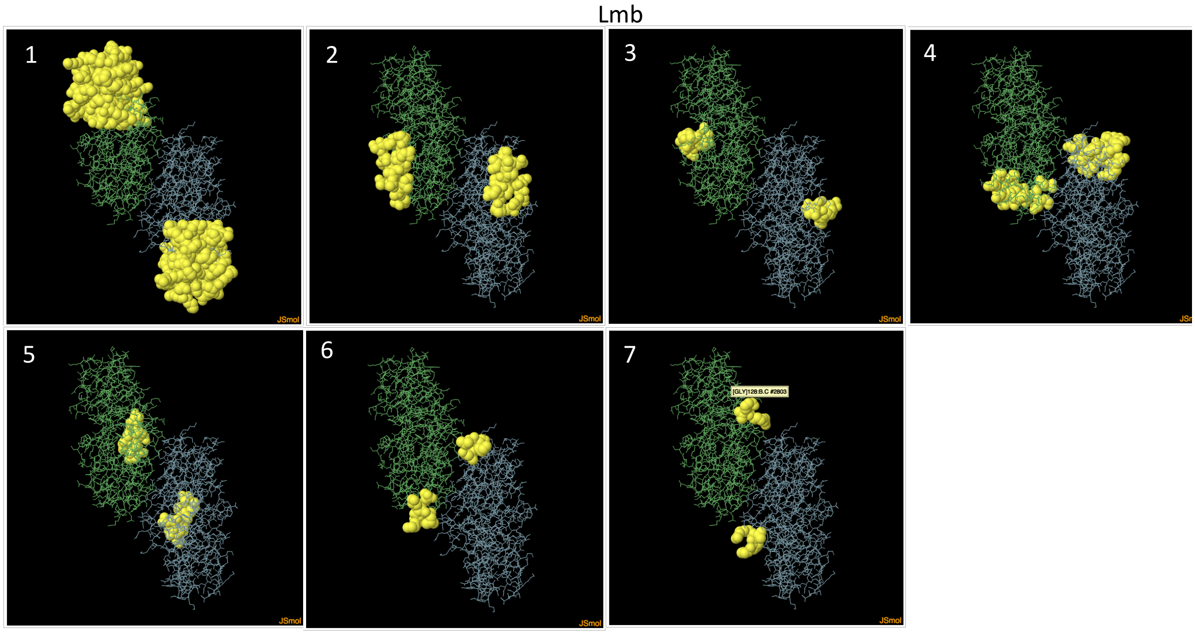


**Figure A14 ElliPro predicted discontinuous epitopes for Lmb.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

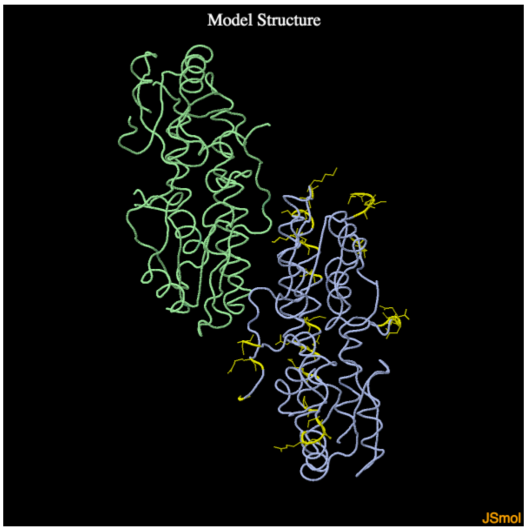


**Figure A15 DiscoTope2 predicted discontinuous epitopes for Lmb.**

The parts coloured yellow are the predicted epitopes.

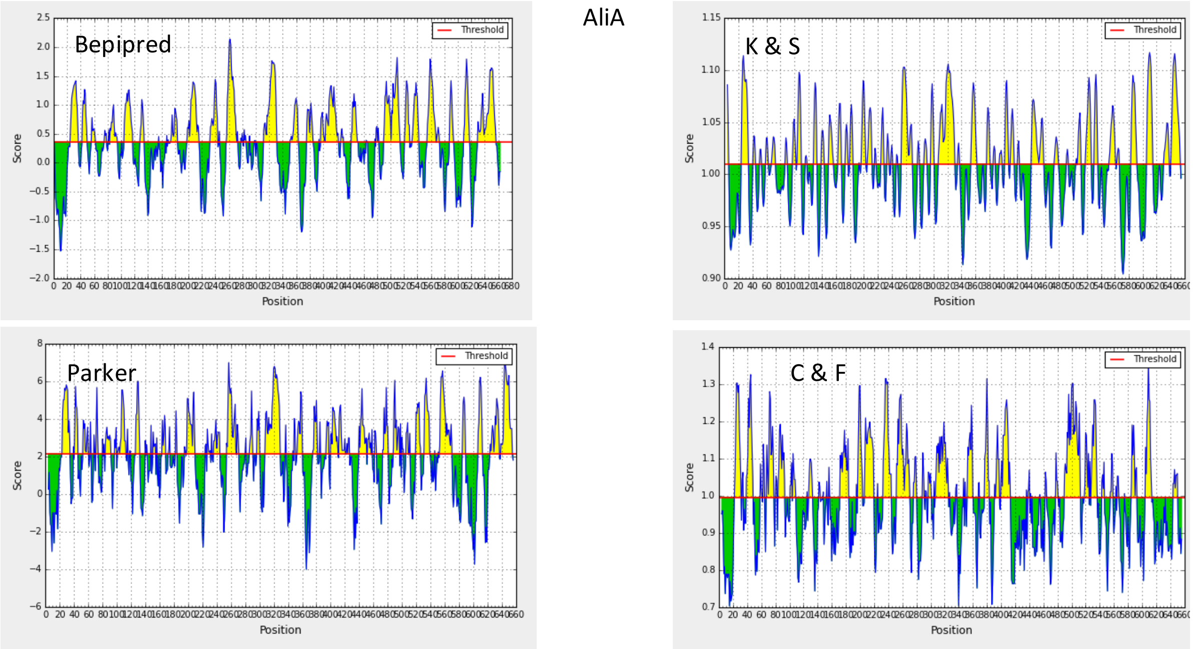


**Figure A16 Linear epitope predictions of the AliA protein.**

## AliA

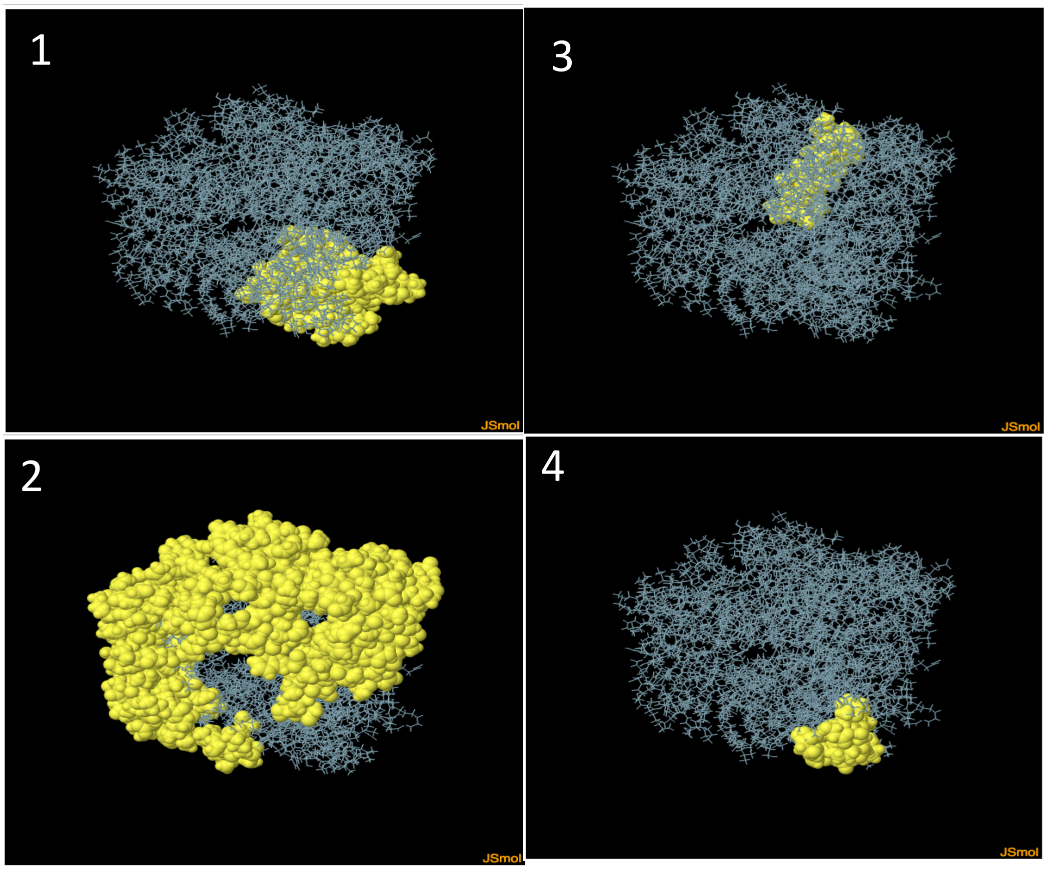


**Figure A17 ElliPro predicted discontinuous epitopes for AliA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

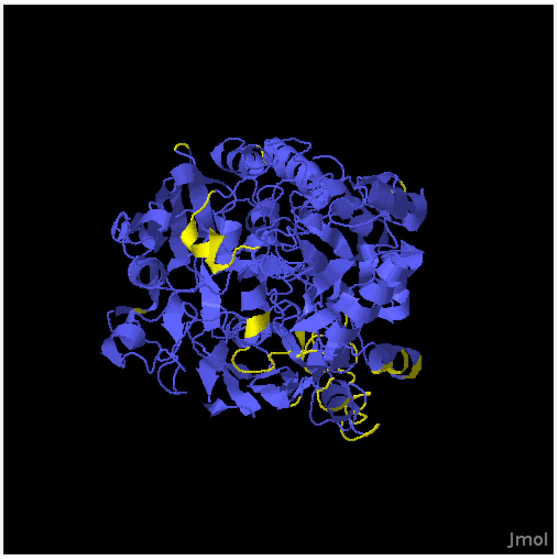


**Figure A18 DiscoTope2 predicted discontinuous epitopes for AliA.**

The parts coloured yellow are the predicted epitopes.

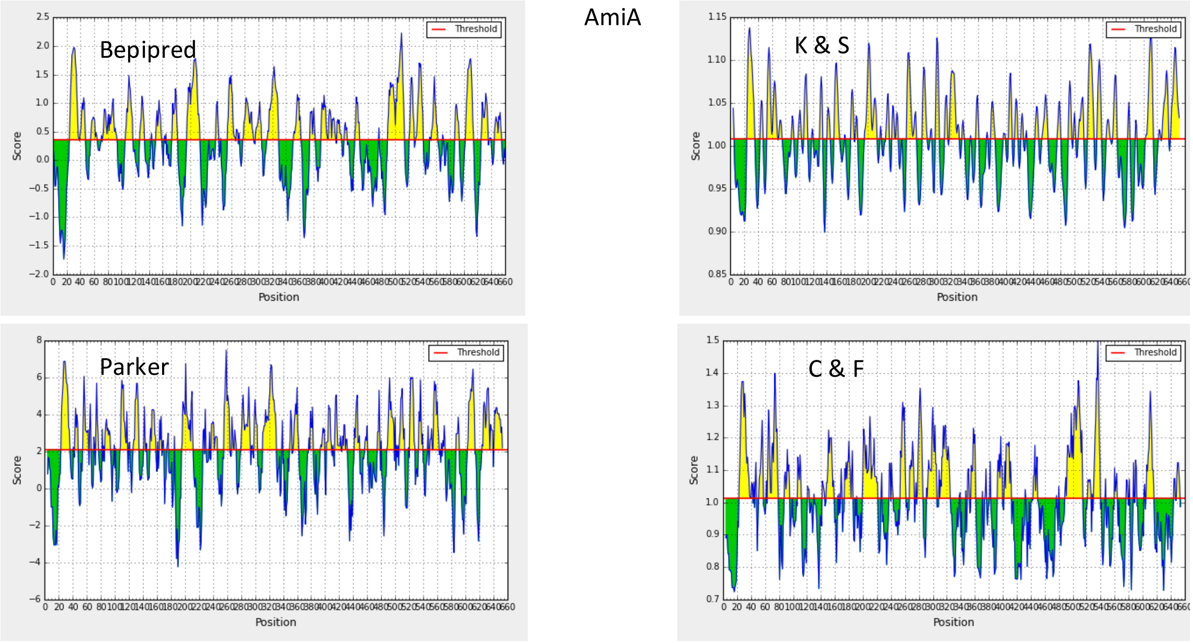


**Figure A19 Linear epitope predictions of the AmiA protein.**

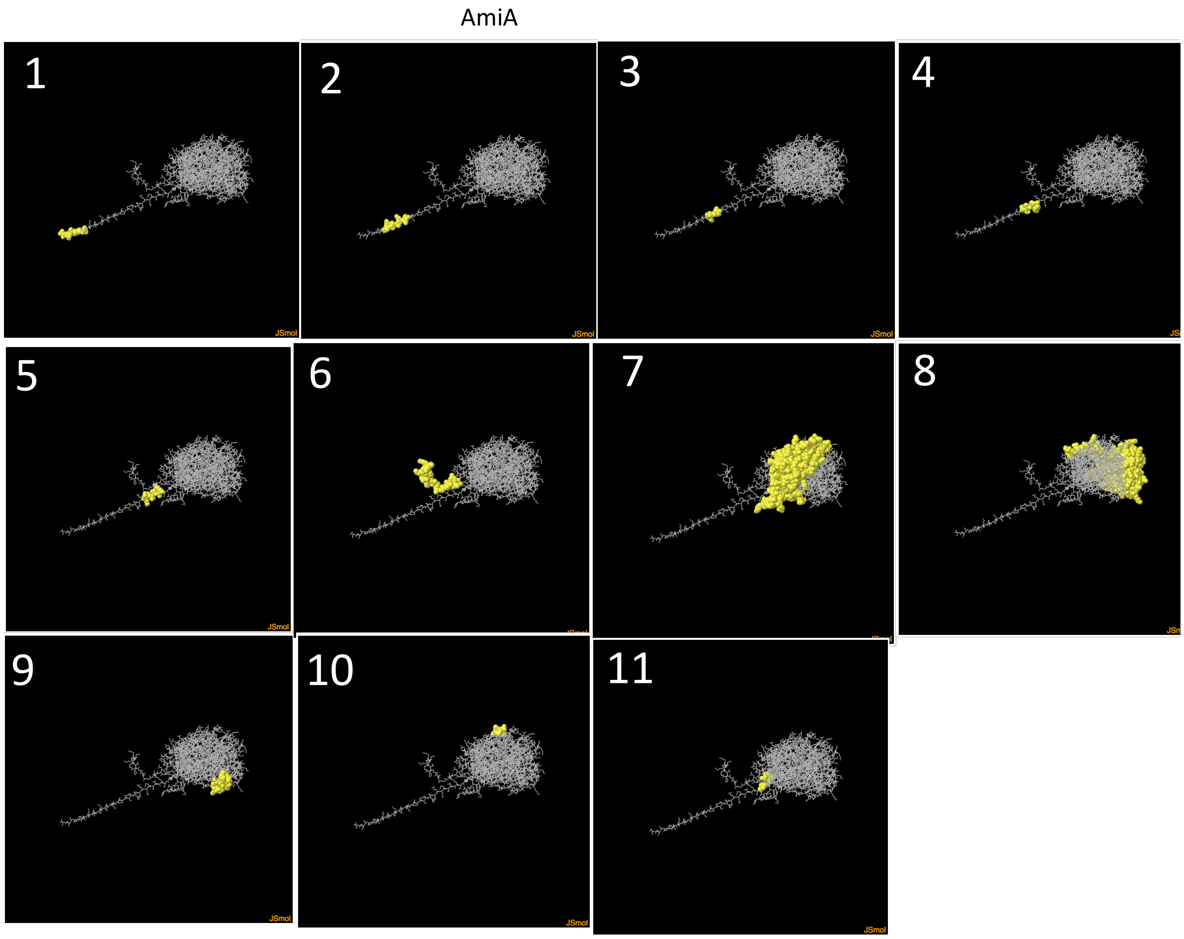


**Figure A20 ElliPro predicted discontinuous epitopes for AmiA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

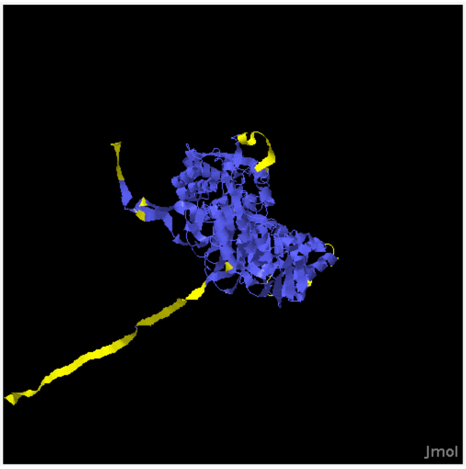


**Figure A21 DiscoTope2 predicted discontinuous epitopes for AmiA.**

The parts coloured yellow are the predicted epitopes.

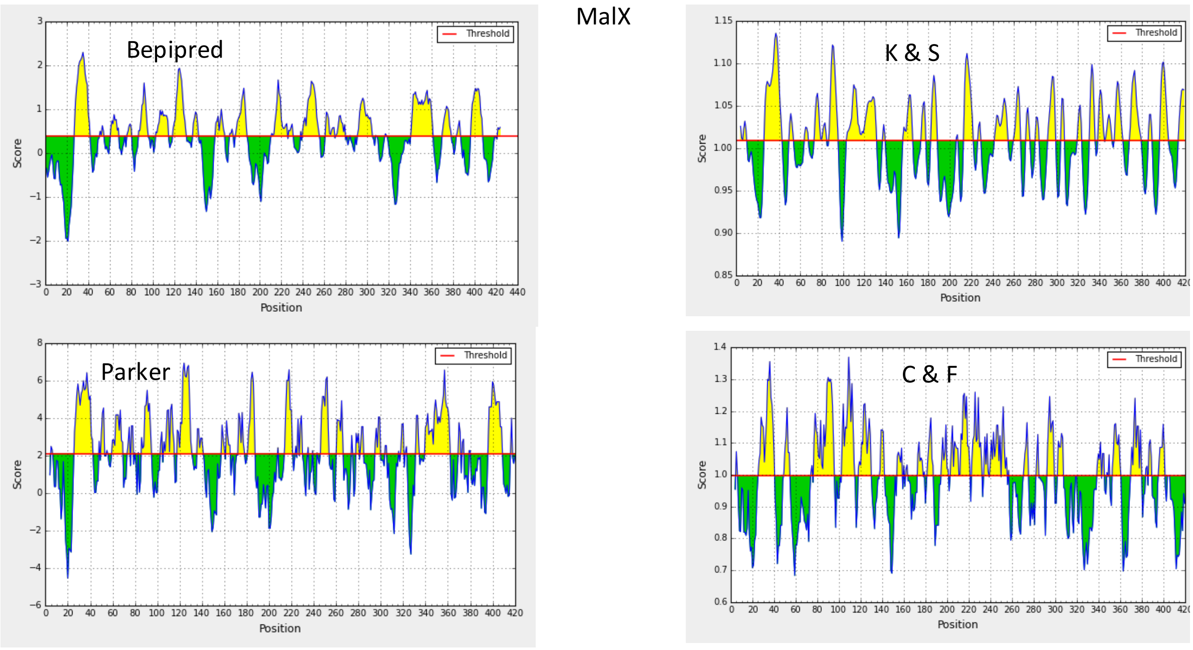


**Figure A22 Linear epitope predictions of the MalX protein.**

MalX

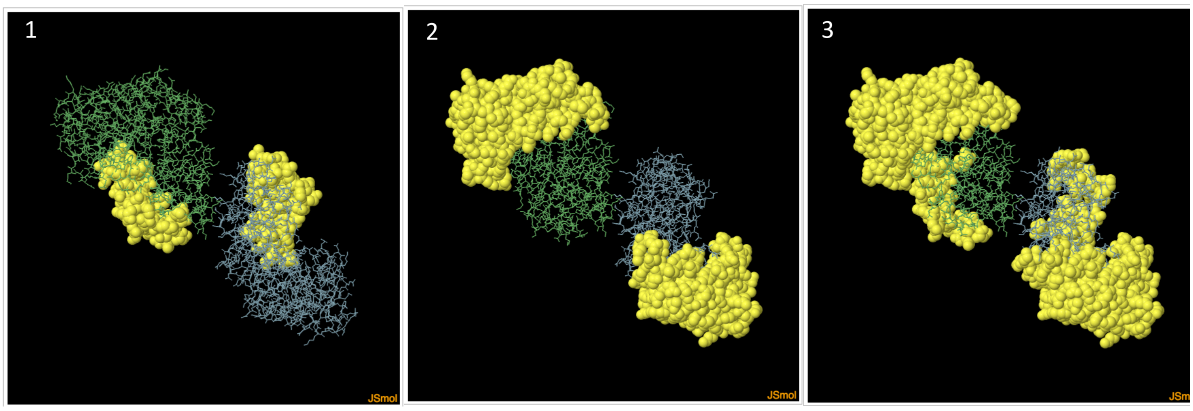


**Figure A23 ElliPro predicted discontinuous epitopes for MalX.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

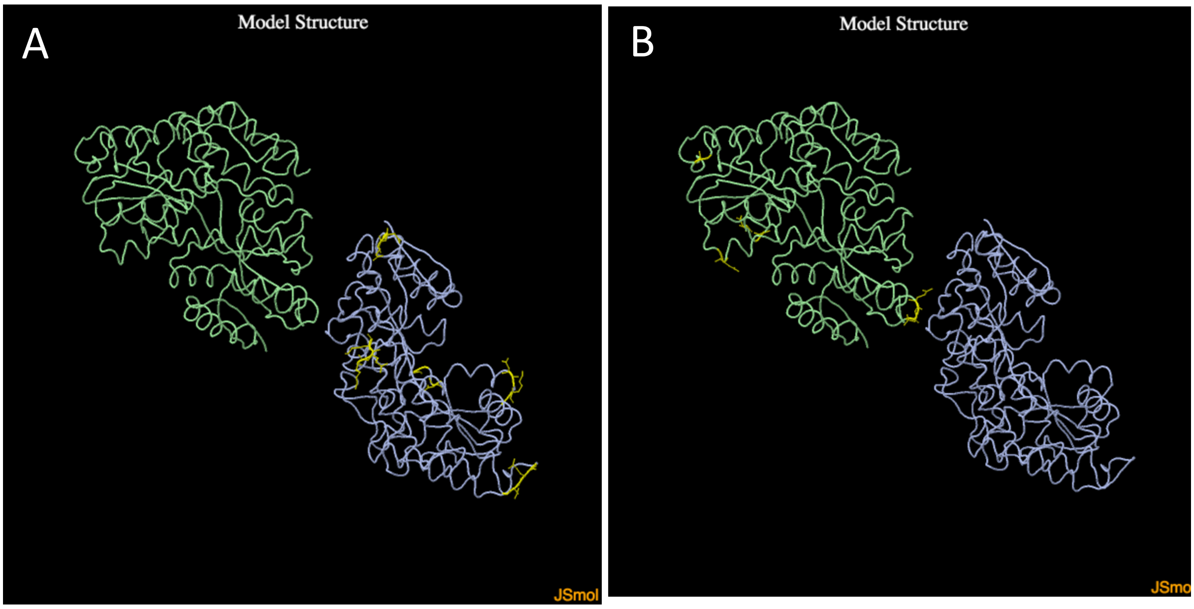


**Figure A24 DiscoTope2 predicted discontinuous epitopes for MalX.**

The parts coloured yellow are the predicted epitopes.

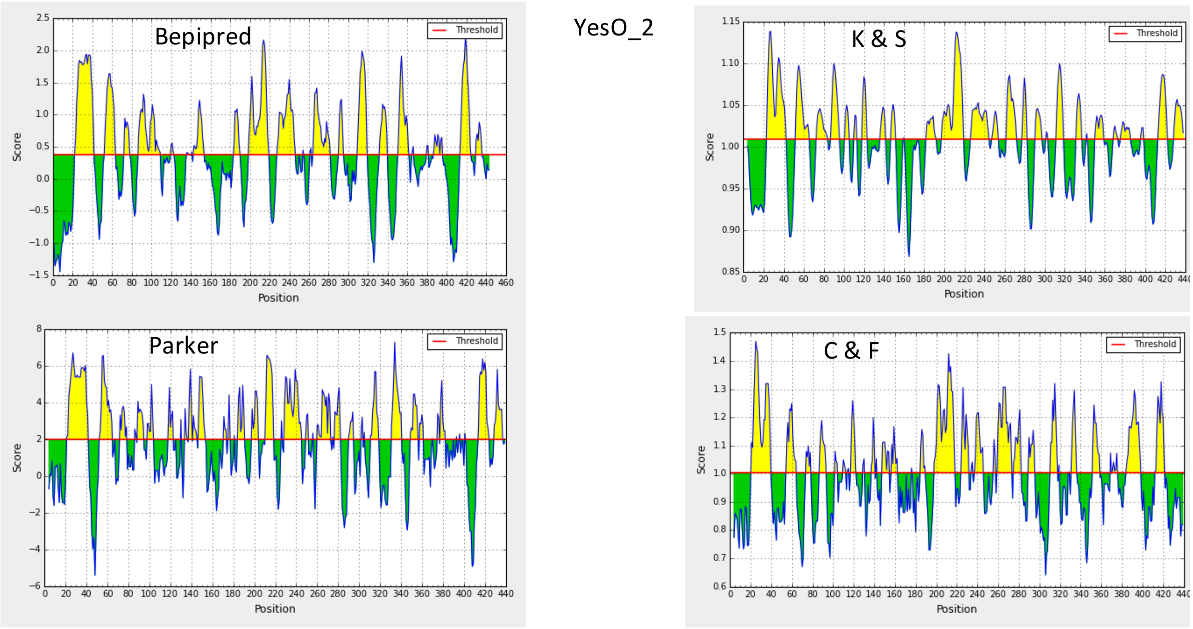


**Figure A25 Linear epitope predictions of the YesO_2 protein.**

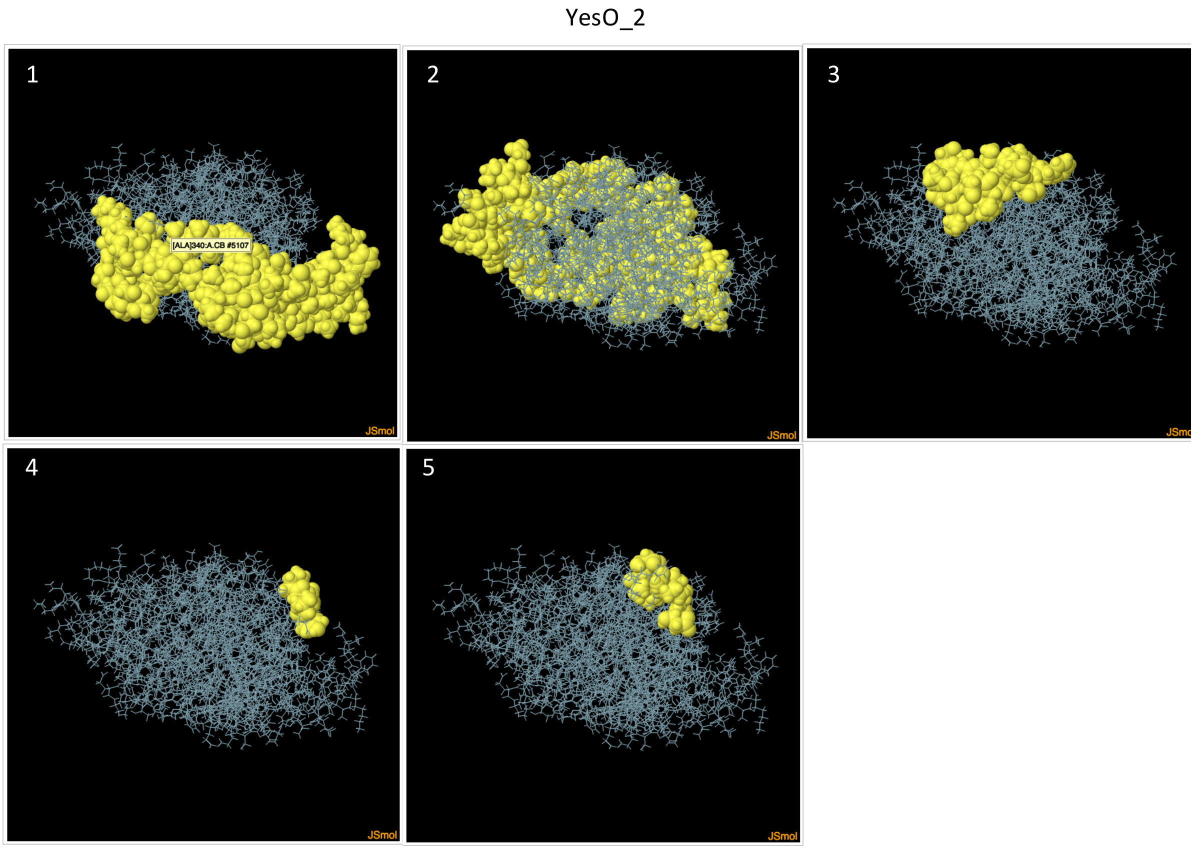


**Figure A26 ElliPro predicted discontinuous epitopes for YesO_2.**
The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

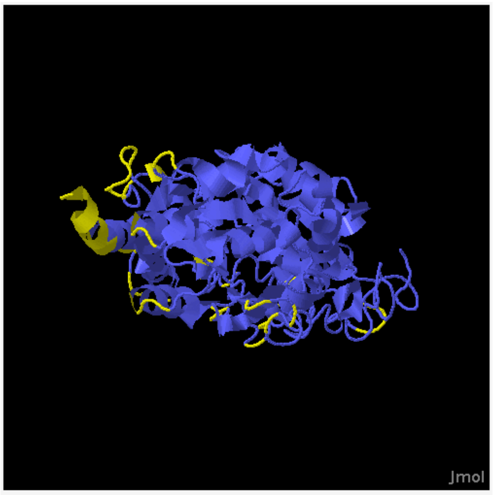


**Figure A27 DiscoTope2 predicted discontinuous epitopes for YesO_2.**
The parts coloured yellow are the predicted epitopes.

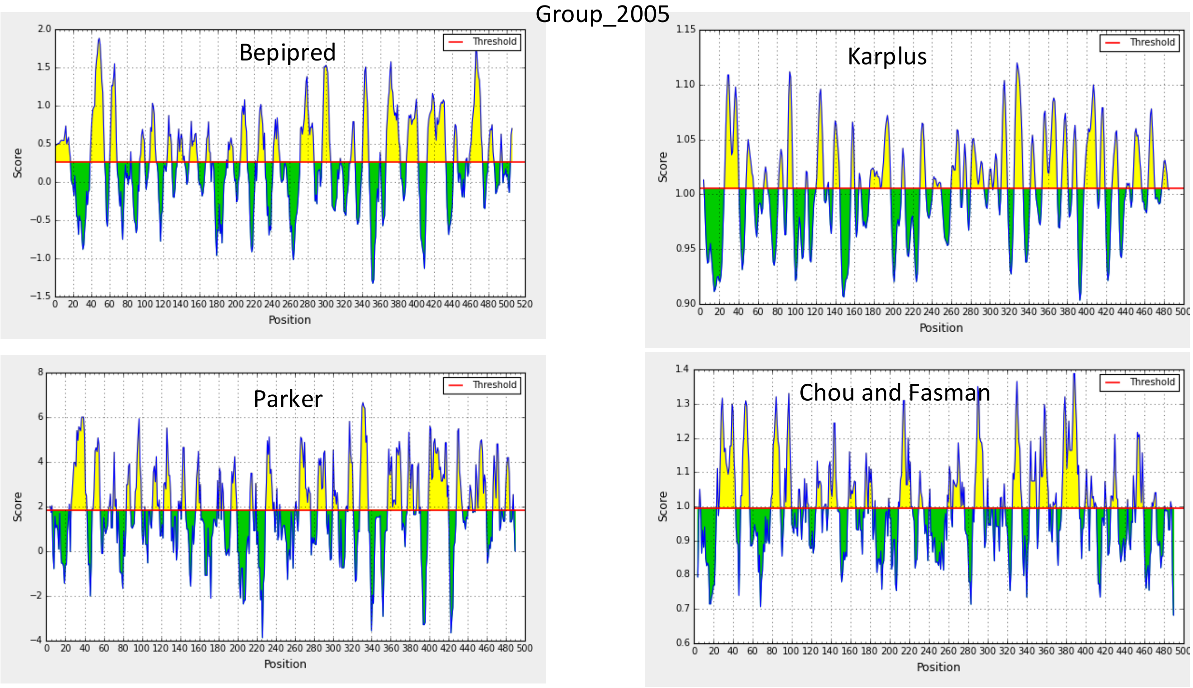


**Figure A28 Linear epitope predictions of the Group_2005.**

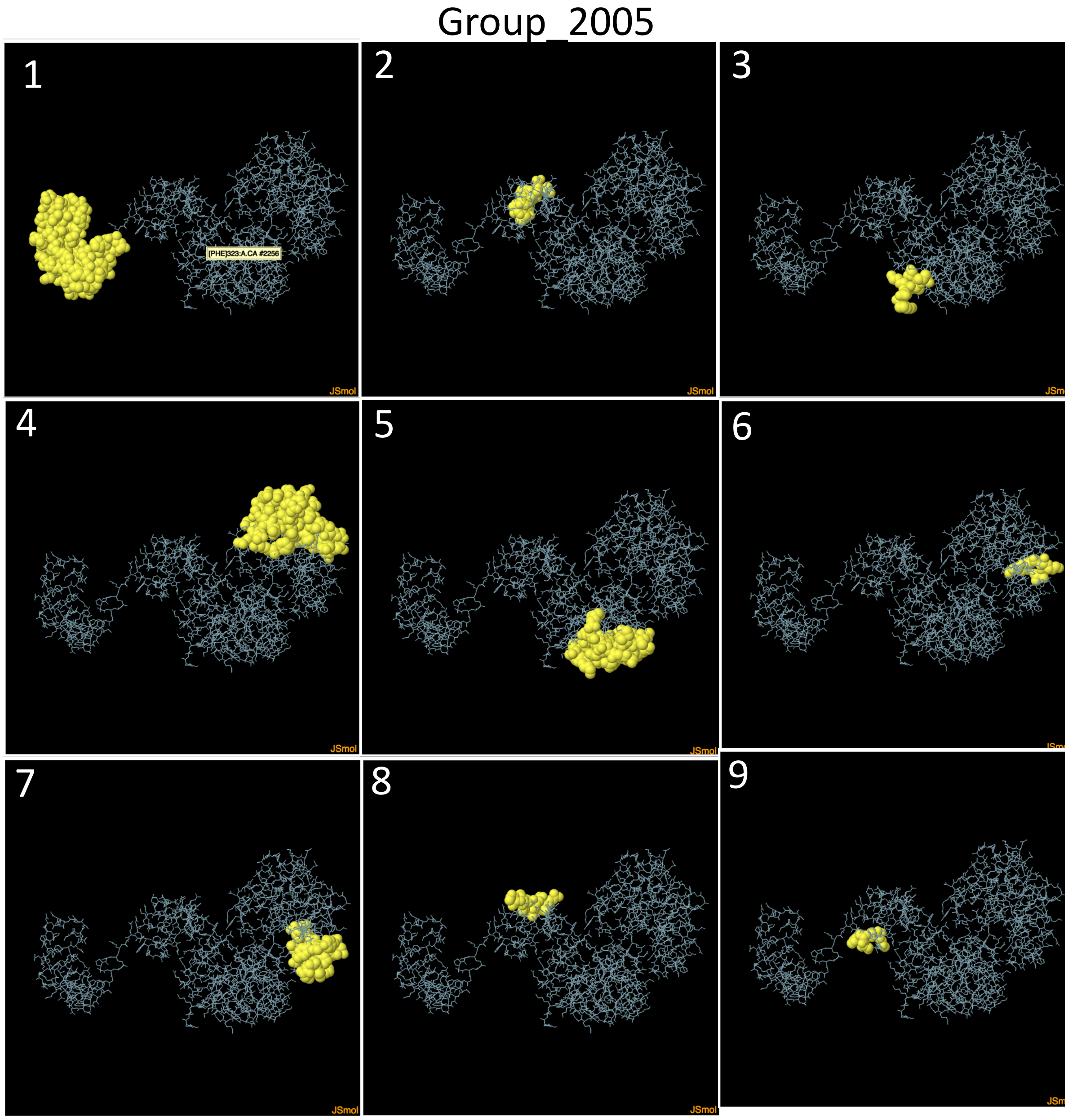


**Figure A29 ElliPro predicted discontinuous epitopes for Group_2005.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

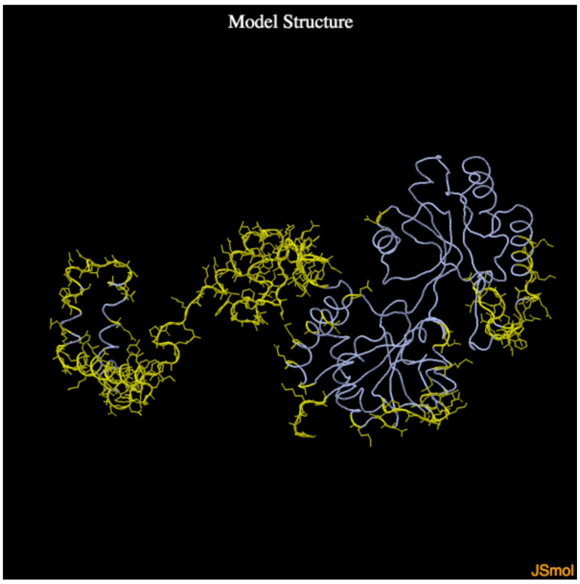


**Figure A30 DiscoTope2 predicted discontinuous epitopes for Group_2005.**

The parts coloured yellow are the predicted epitopes.

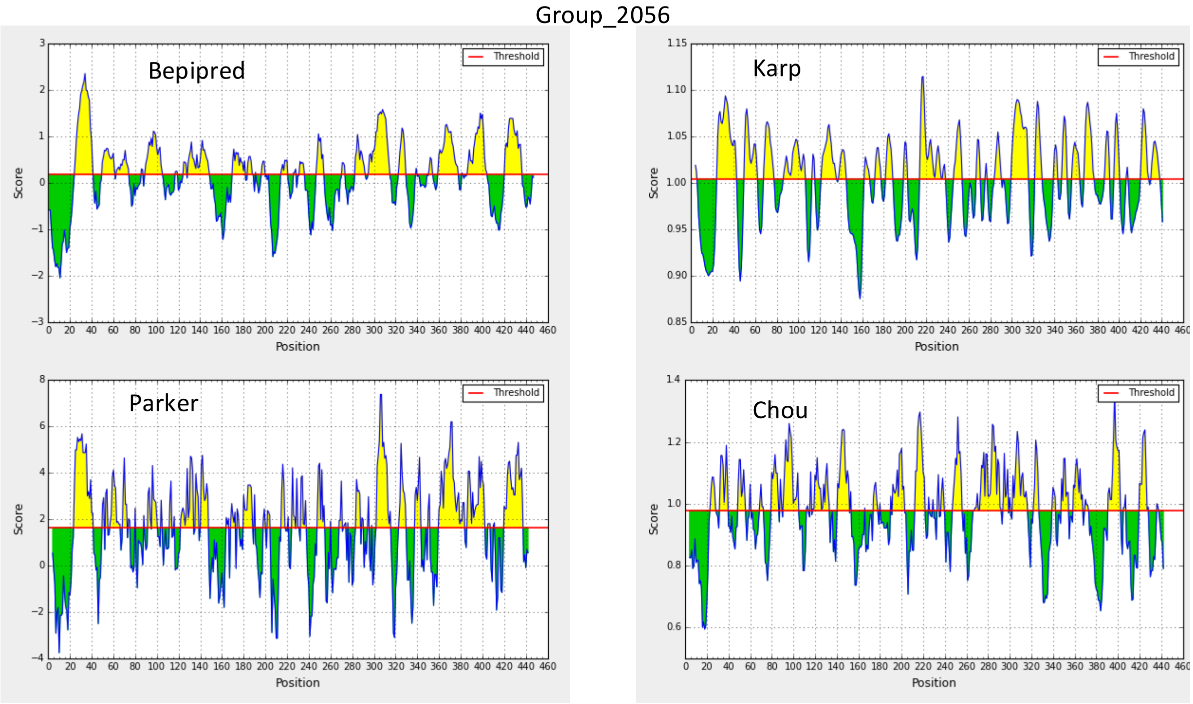


**Figure A31 Linear epitope predictions of the Group_2056.**

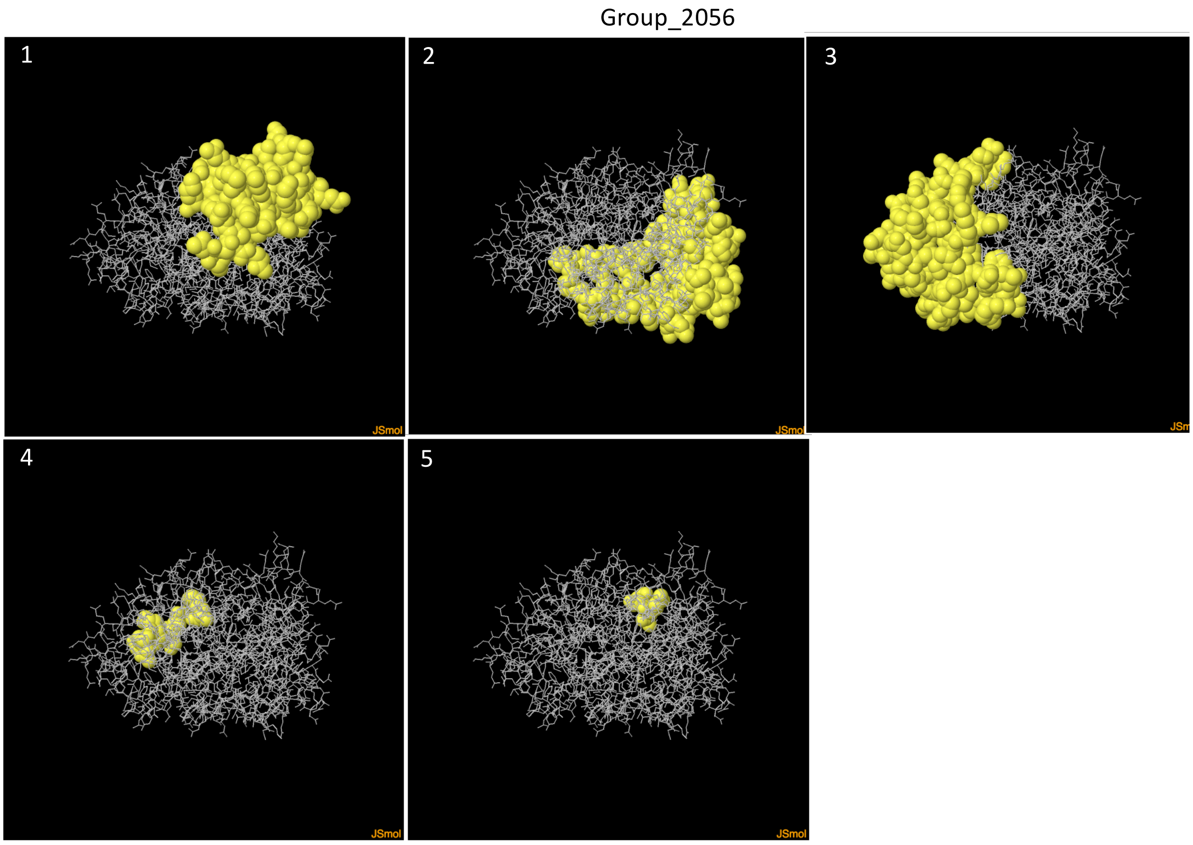


**Figure A32 ElliPro predicted discontinuous epitopes for Group_2056.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

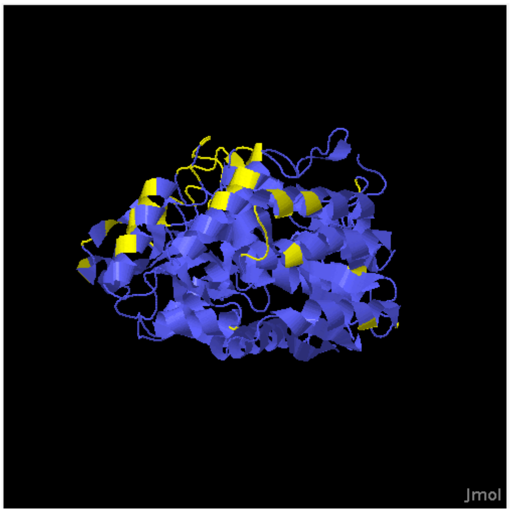


**Figure A33 DiscoTope2 predicted discontinuous epitopes for Group_2056.**

The parts coloured yellow are the predicted epitopes.

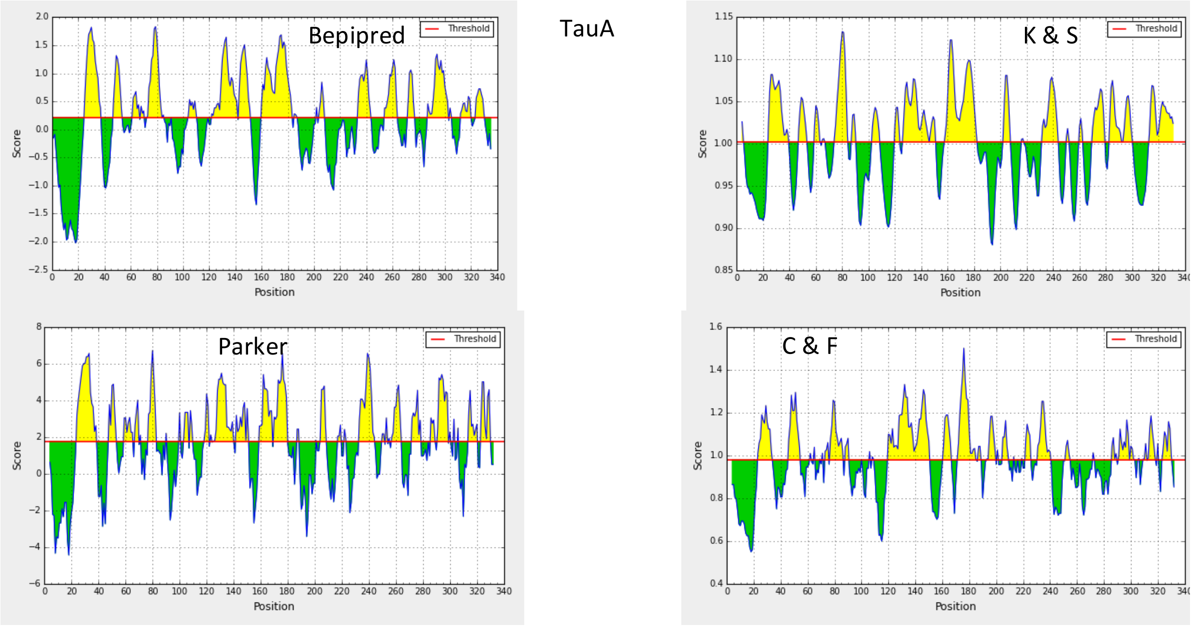


**Figure A34 Linear epitope predictions of the TauA protein.**

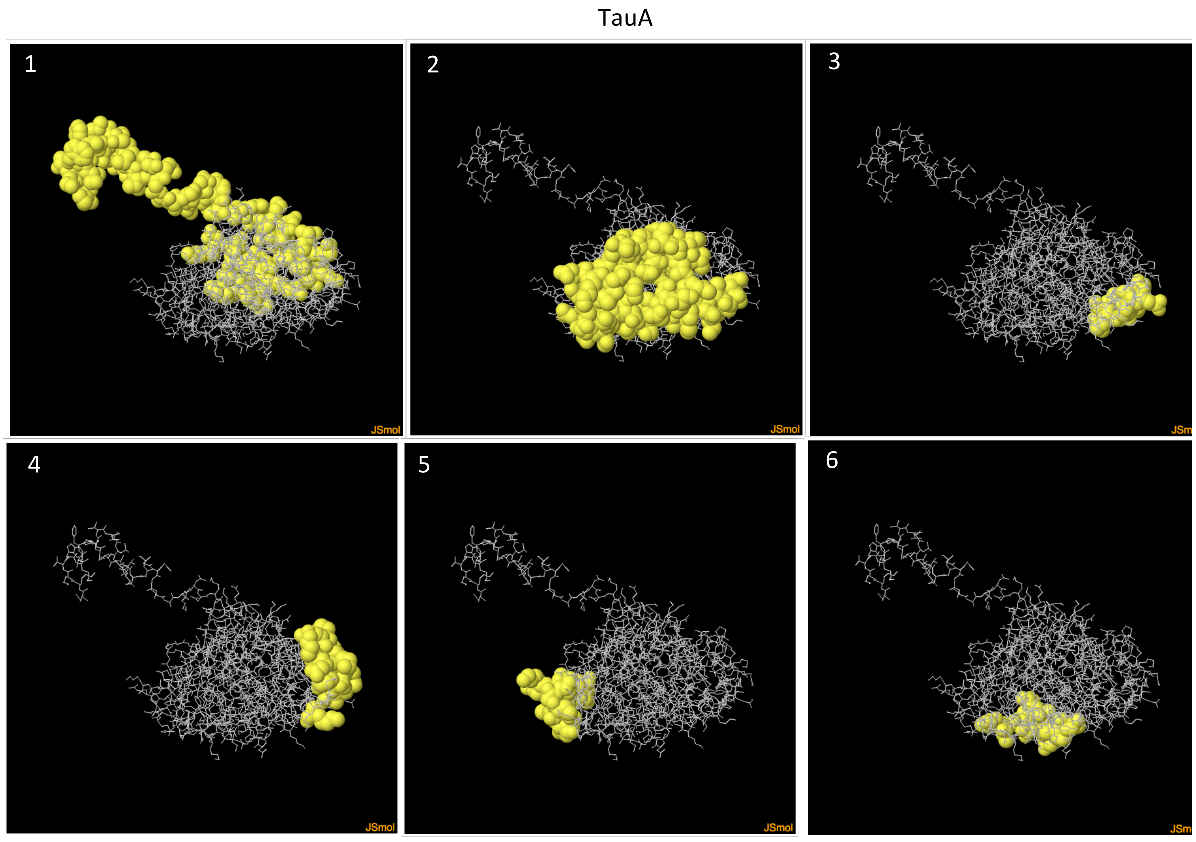


**Figure A35 ElliPro predicted discontinuous epitopes for TauA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

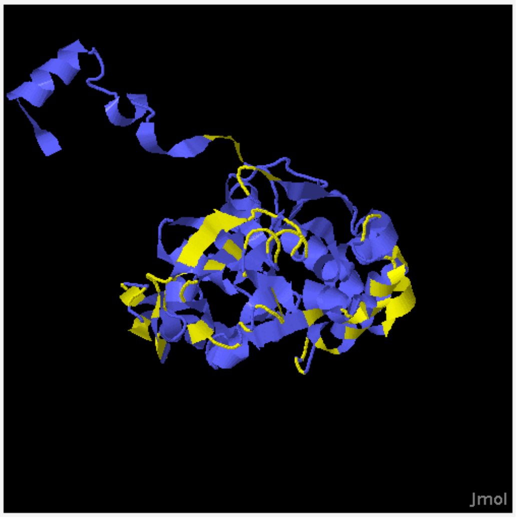


**Figure A36 DiscoTope2 predicted discontinuous epitopes for TauA.**

The parts coloured yellow are the predicted epitopes.

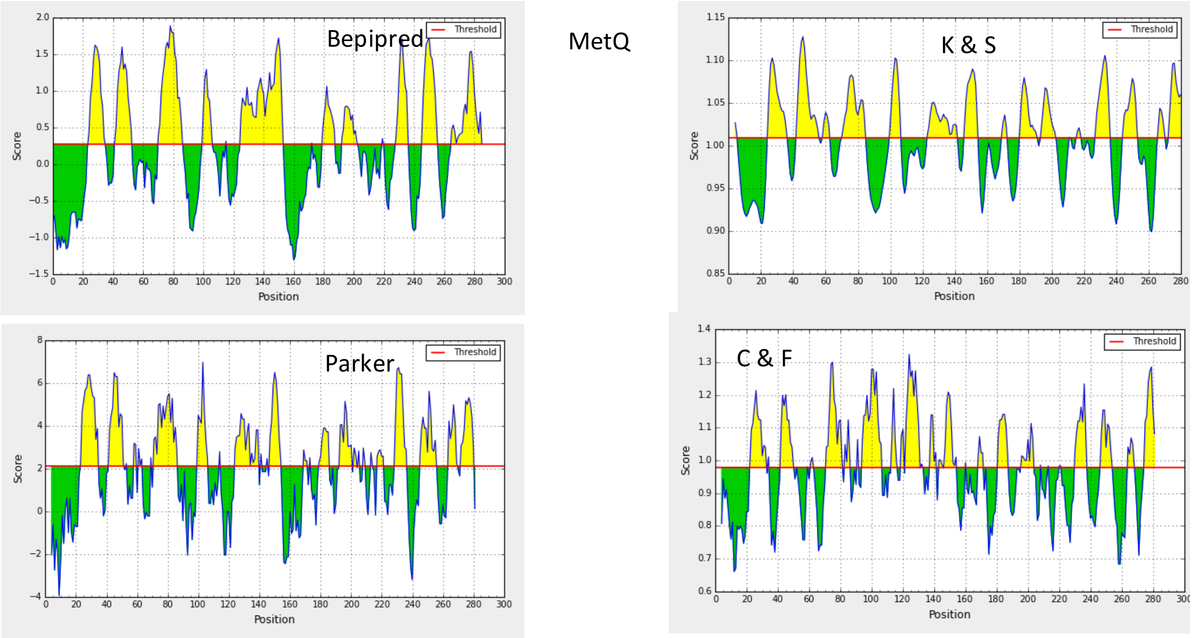


**Figure A37 Linear epitope predictions of the MetQ protein.**

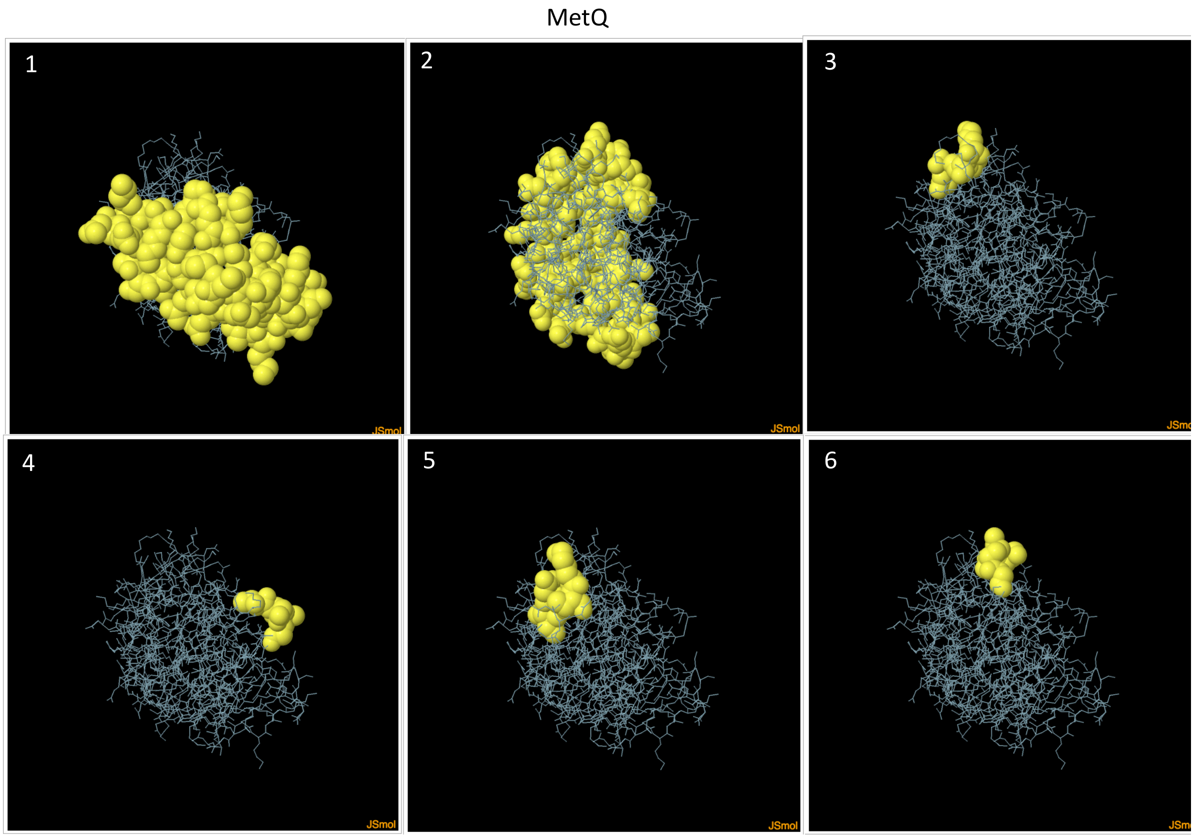


**Figure A38 ElliPro predicted discontinuous epitopes for MetQ.**

The numbers represent the different epitopes predicted in the order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

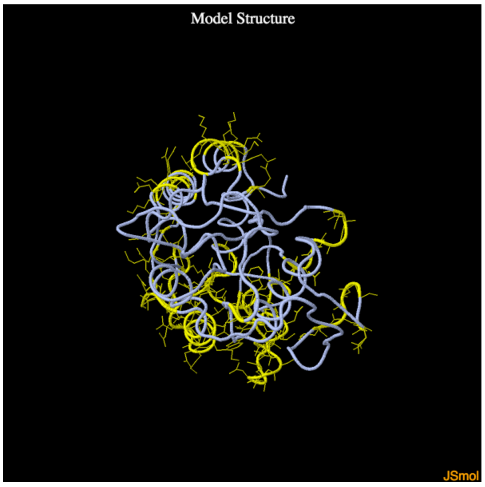


**Figure A39 DiscoTope2 predicted discontinuous epitopes for MetQ.**

The parts coloured yellow are the predicted epitopes.

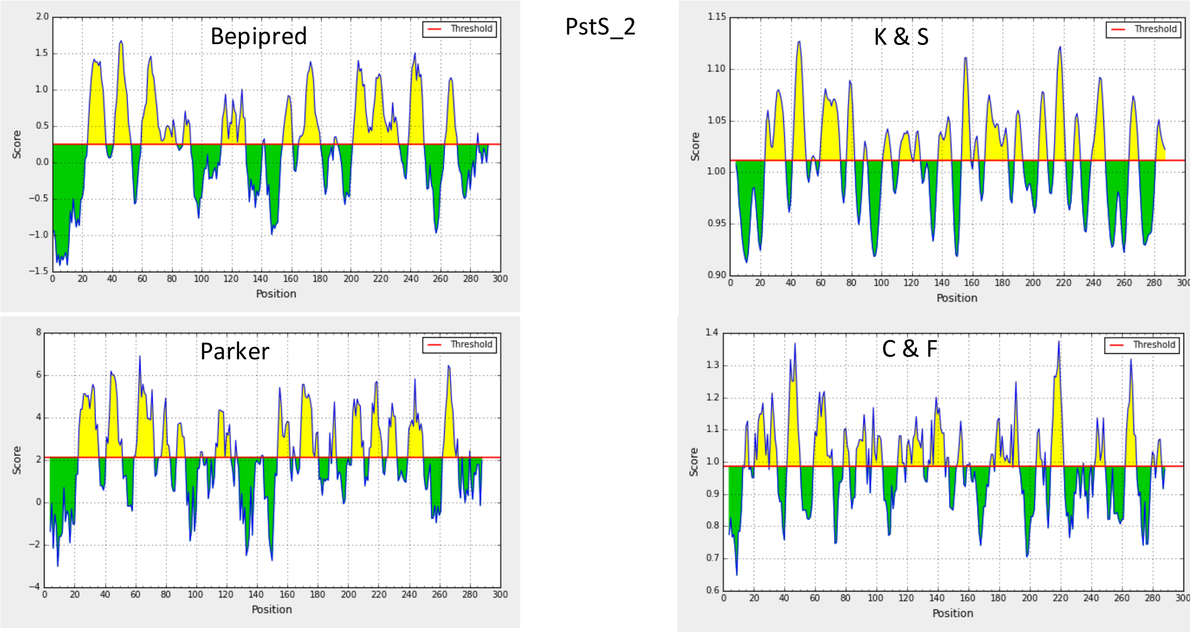


**Figure A40 Linear epitope predictions of the PstS_2 protein.**

PstS_2

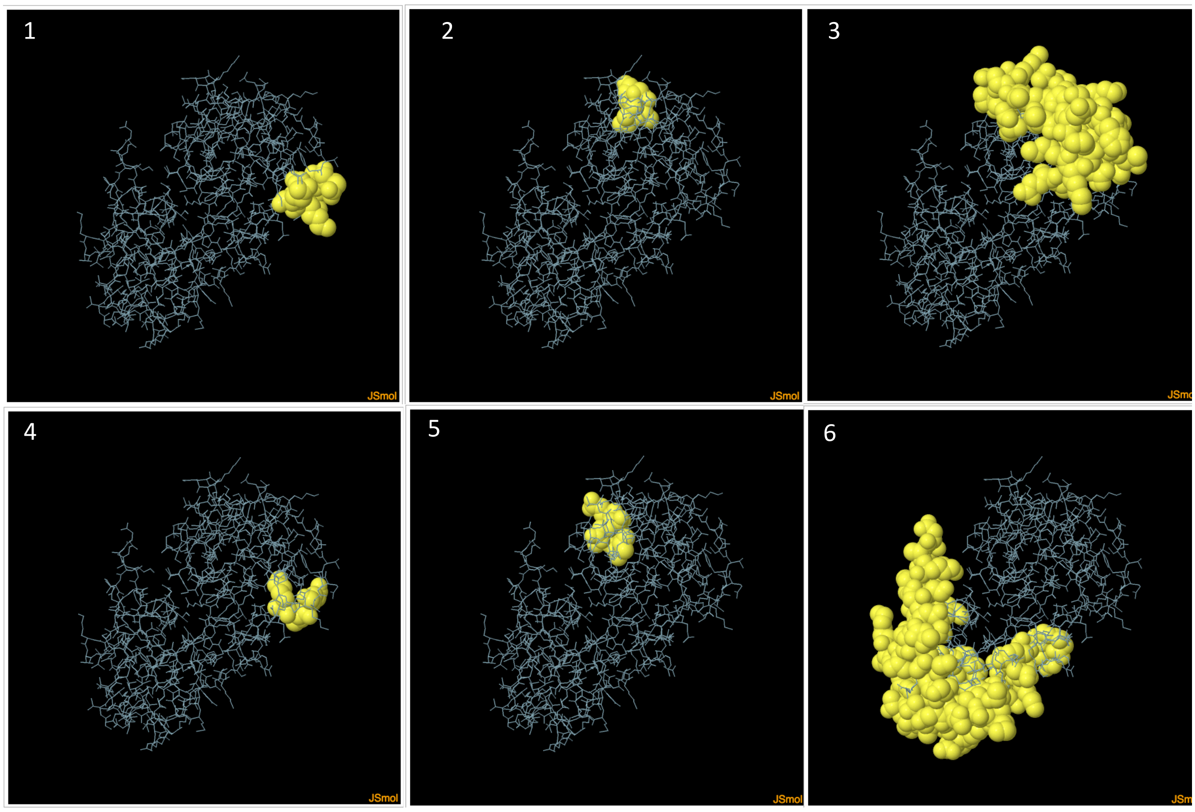


**Figure A41 ElliPro predicted discontinuous epitopes for PstS_2.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

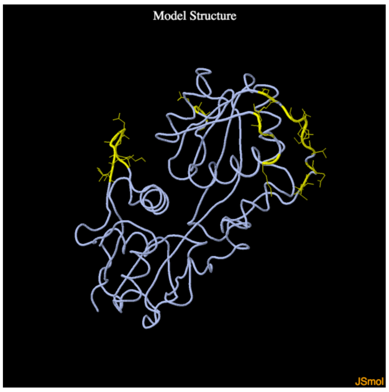


**Figure A42 DiscoTope2 predicted discontinuous epitopes for PstS_2.**

The parts coloured yellow are the predicted epitopes.

ArtP_1

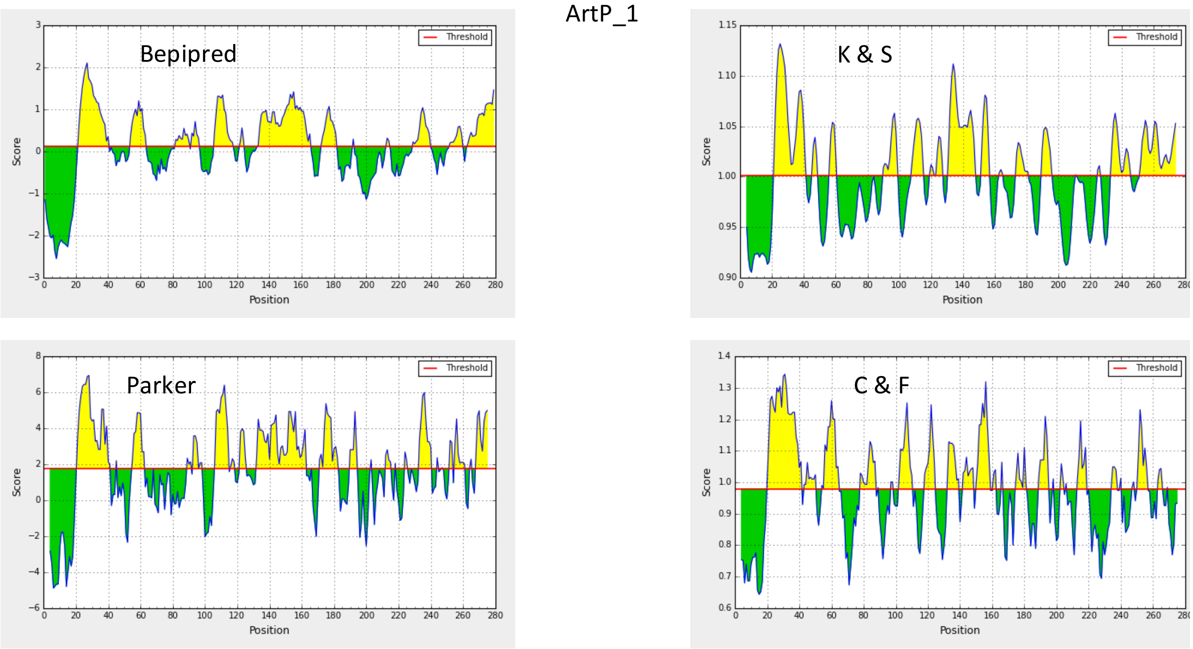


**Figure A43 Linear epitope predictions of the ArtP_1 protein.**

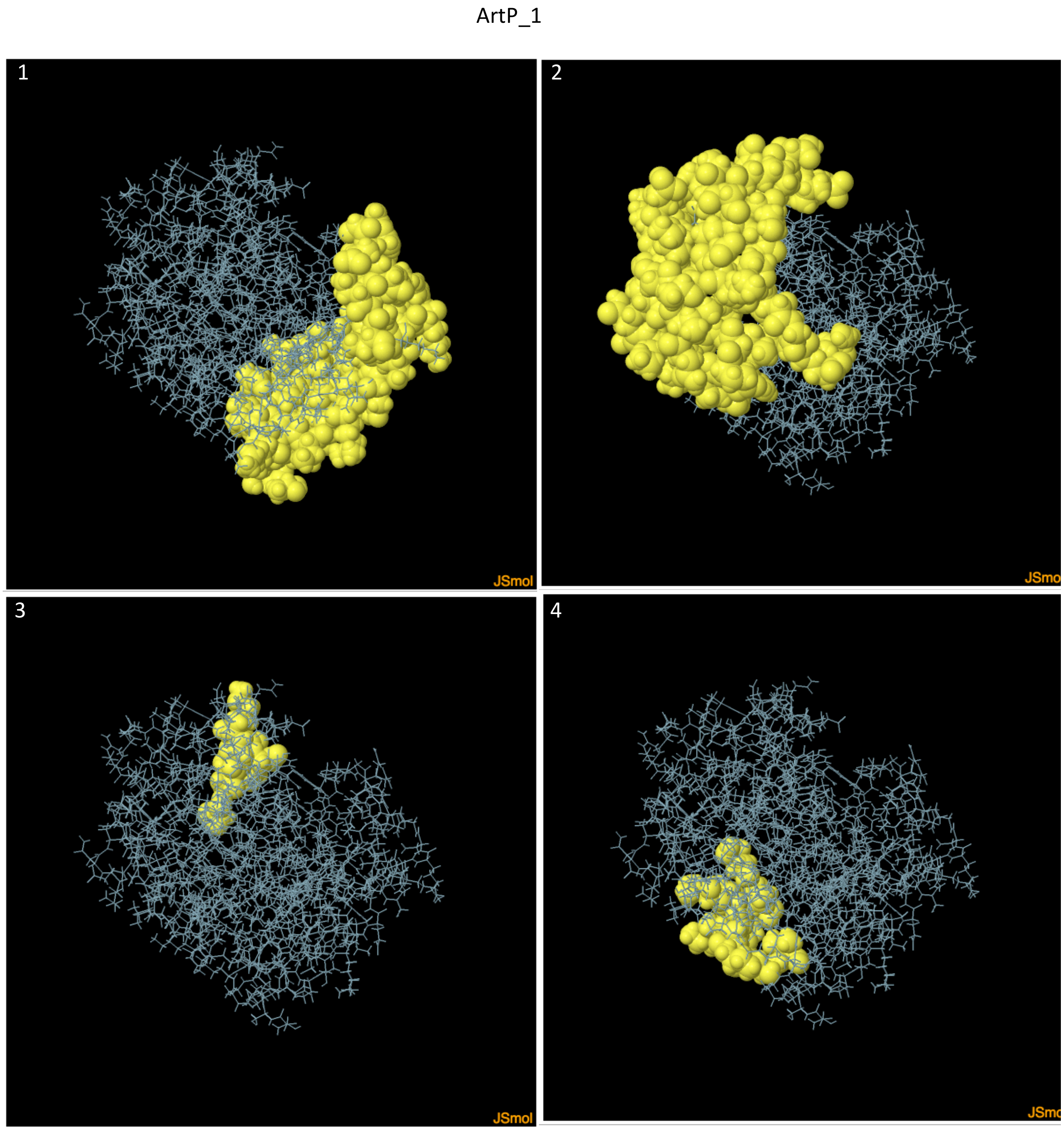


**Figure A44 ElliPro predicted discontinuous epitopes for ArtP_1.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

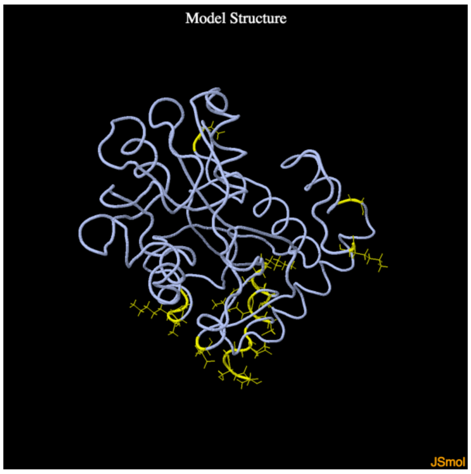


**Figure A45 DiscoTope2 predicted discontinuous epitopes for ArtP_1.**

The parts coloured yellow are the predicted epitopes.

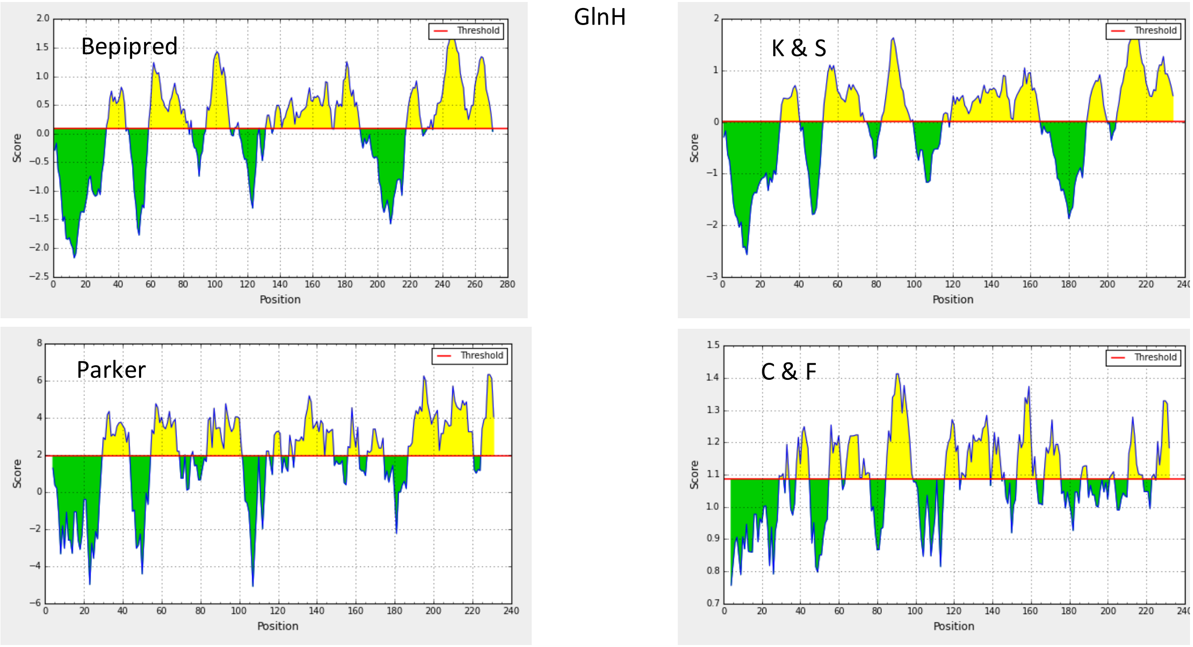


**Figure A46 Linear epitope predictions of the GlnH protein.**

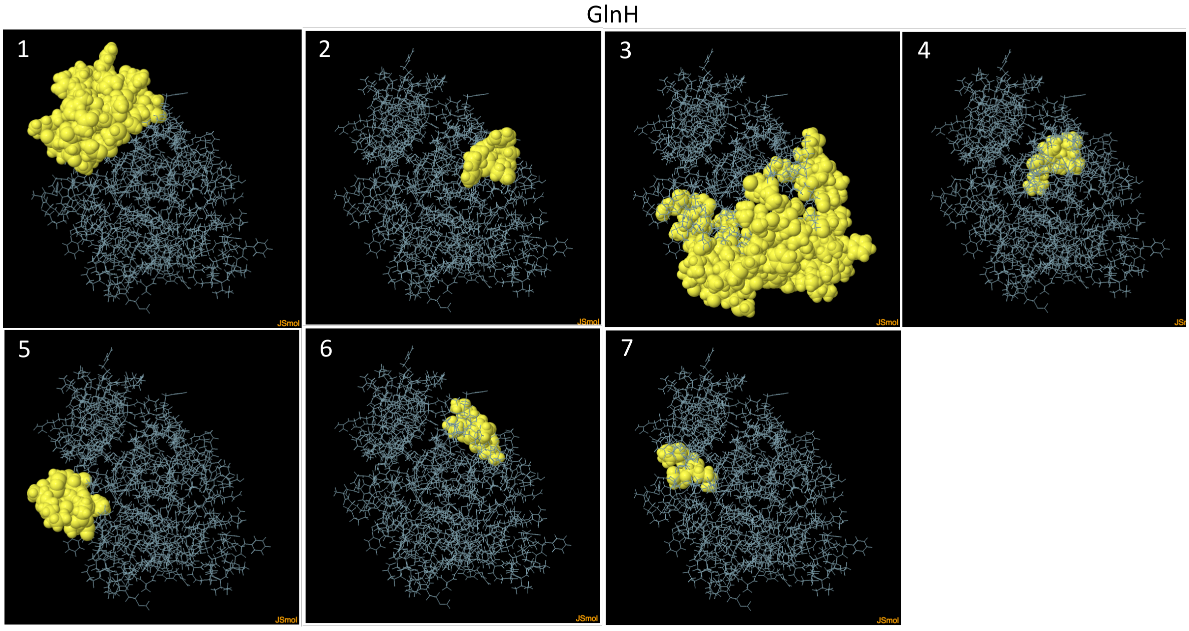


**Figure A47 ElliPro predicted discontinuous epitopes for GlnH.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

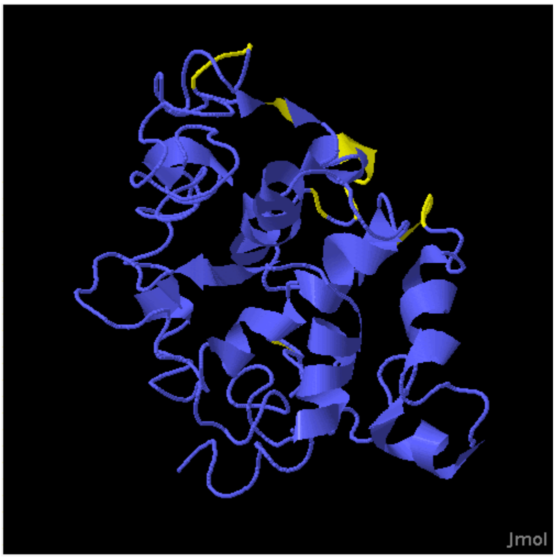


**Figure A48 DiscoTope2 predicted discontinuous epitopes for GlnH.**

The parts coloured yellow are the predicted epitopes.

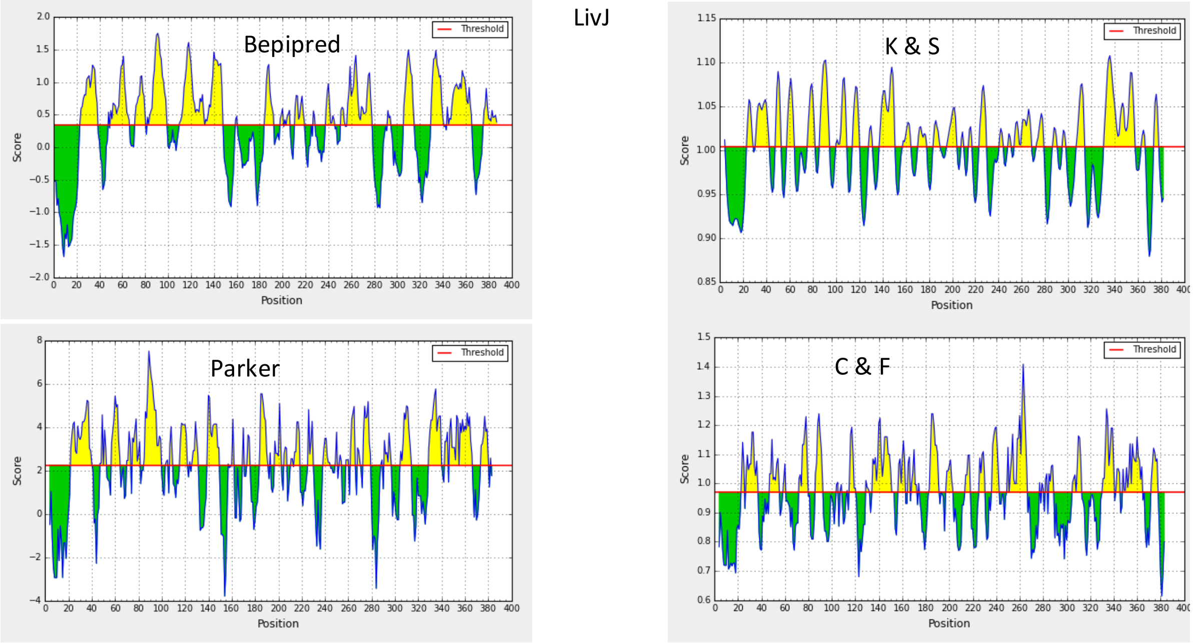


**Figure A49 Linear epitope predictions of the LivJ protein.**

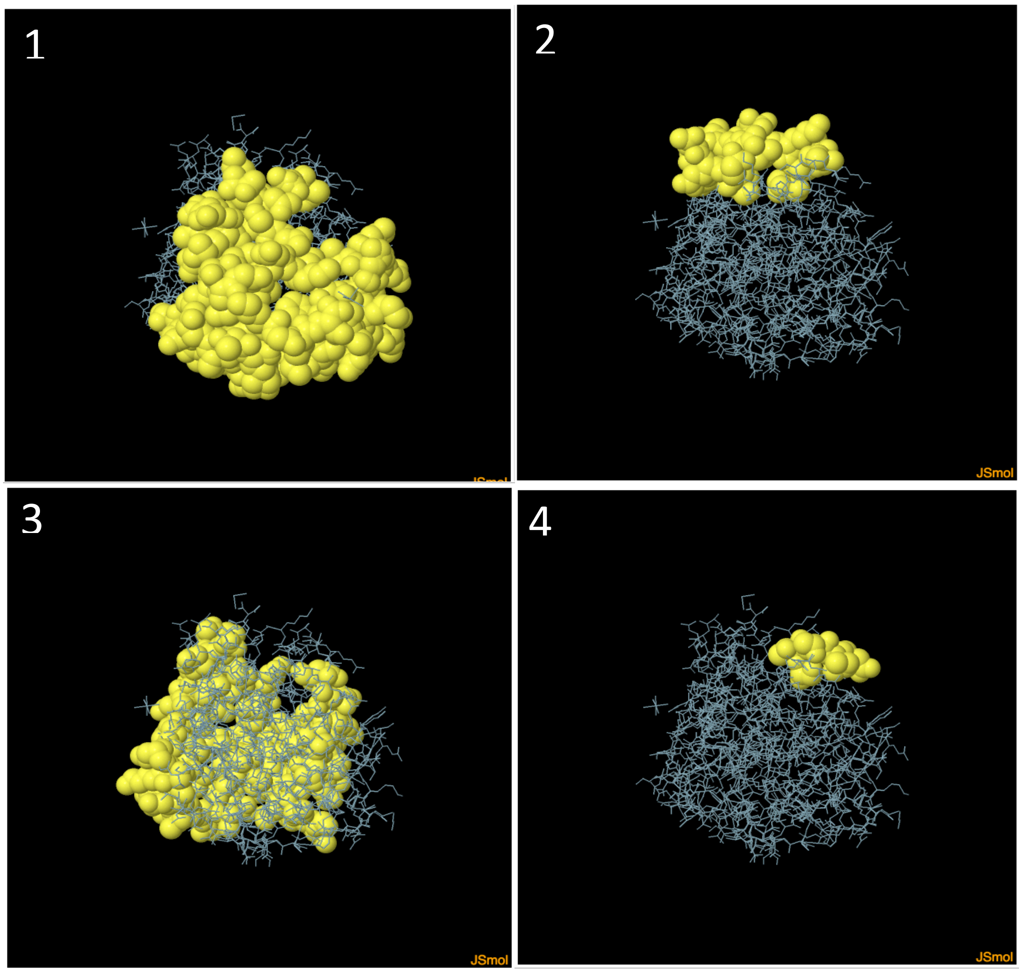


**Figure A50 ElliPro predicted discontinuous epitopes for LivJ.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

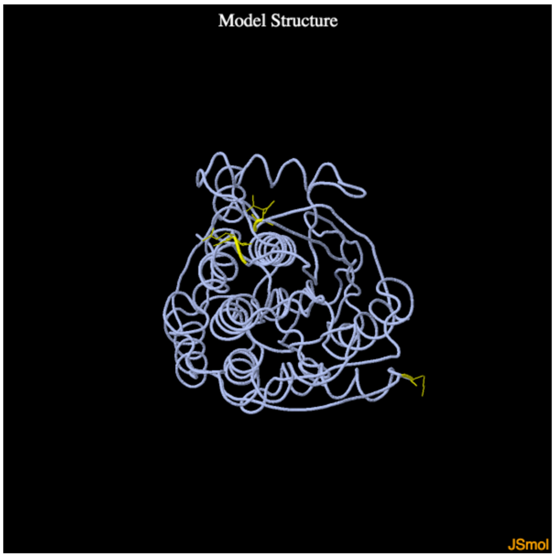


**Figure A51 DiscoTope2 predicted discontinuous epitopes for LivJ.**

The parts coloured yellow are the predicted epitopes.

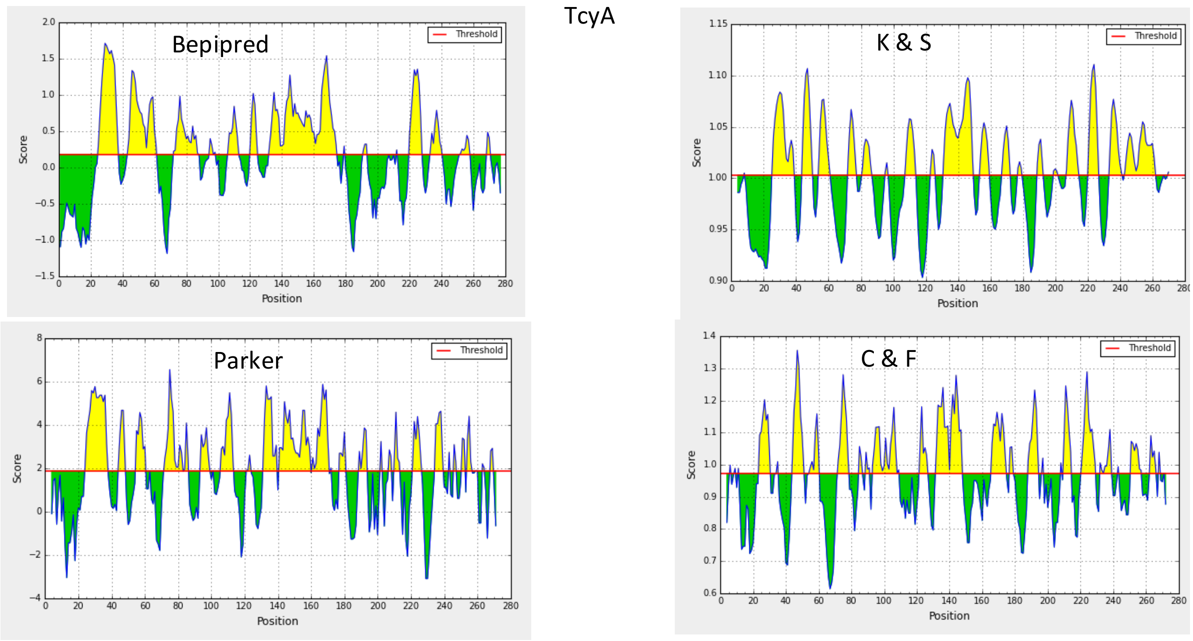


**Figure A52 Linear epitope predictions of the TcyA protein.**

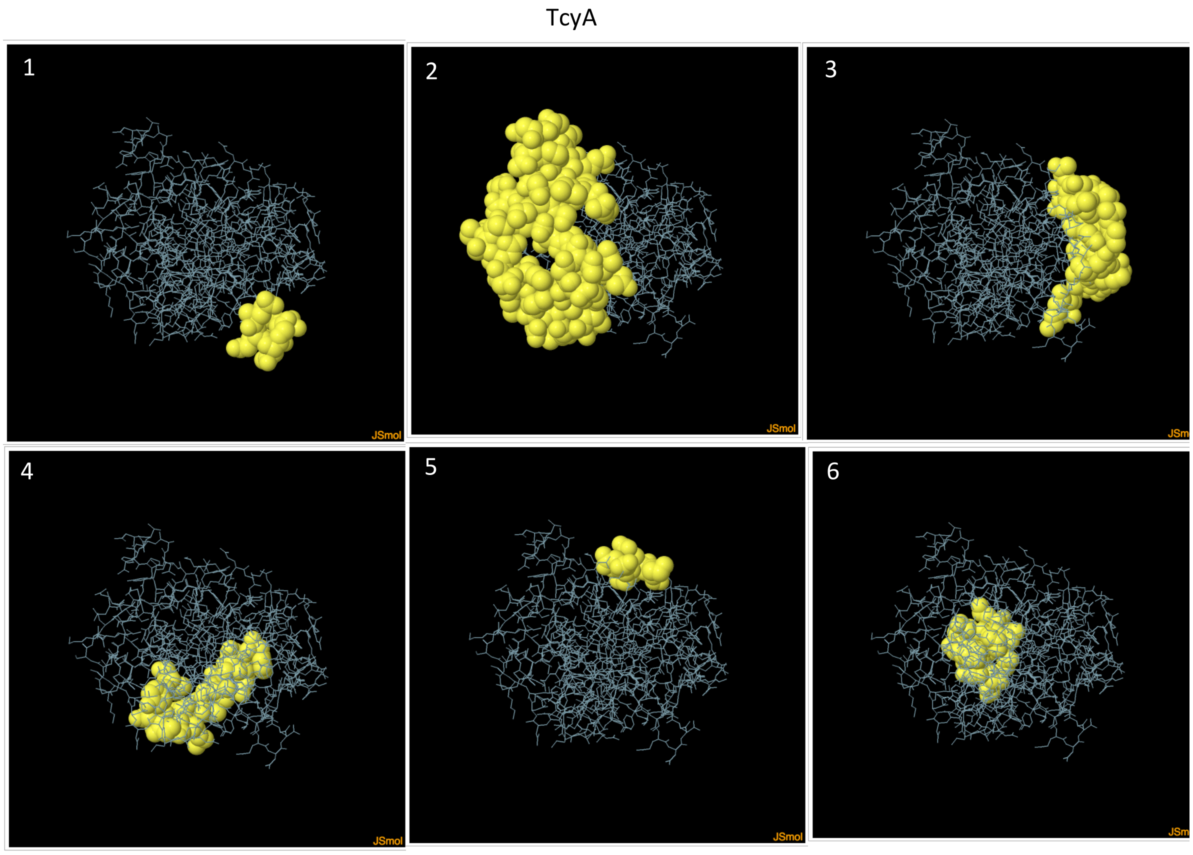


**Figure A53 ElliPro predicted discontinuous epitopes for TcyA.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

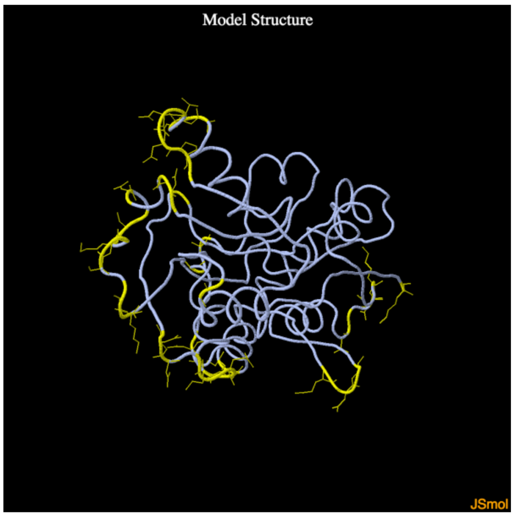


**Figure A54 DiscoTope2 predicted discontinuous epitopes for TcyA.**

The parts coloured yellow are the predicted epitopes.

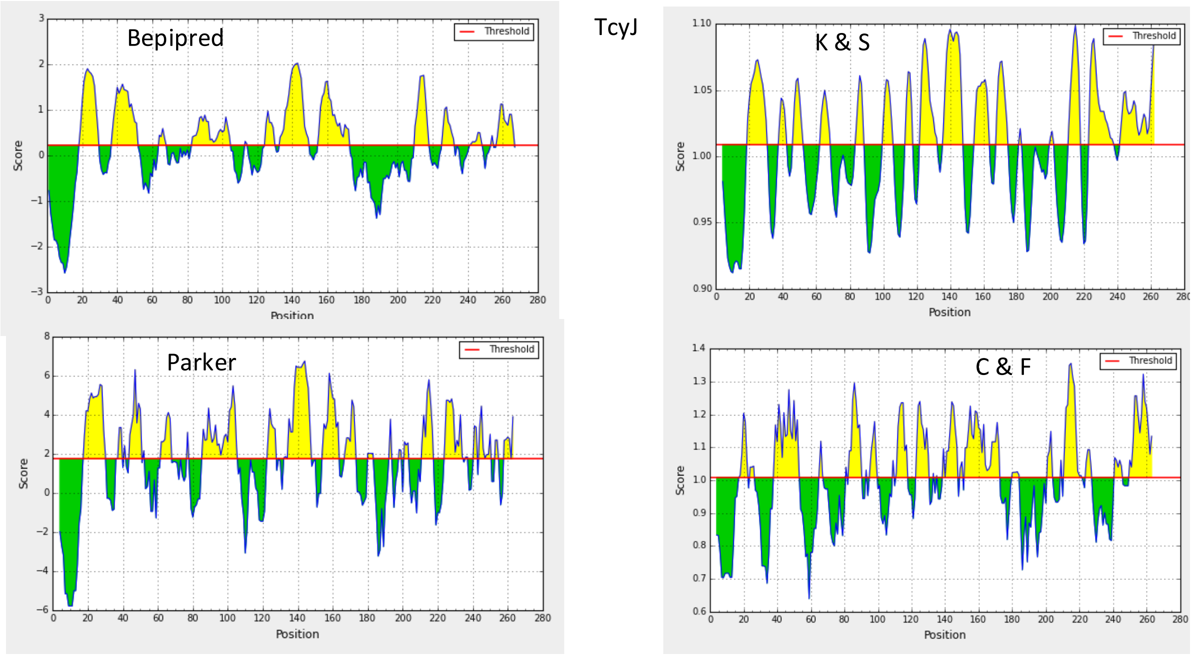


**Figure A55 Linear epitope predictions of the TcyJ protein.**

tcyJ

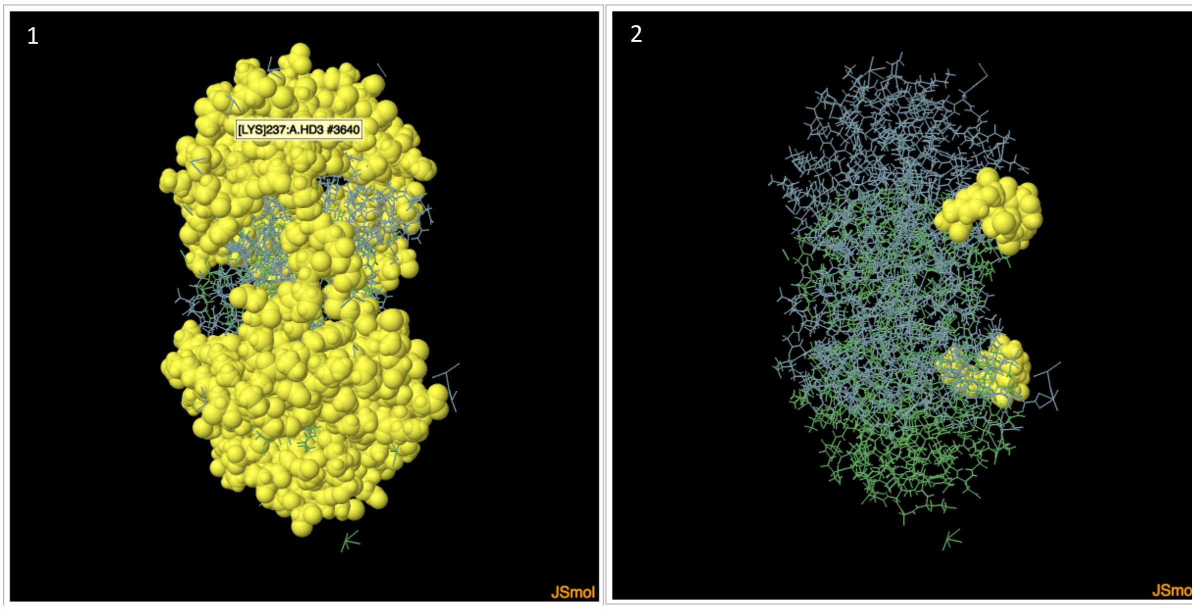


**Figure A56 ElliPro predicted discontinuous epitopes for TcyJ.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

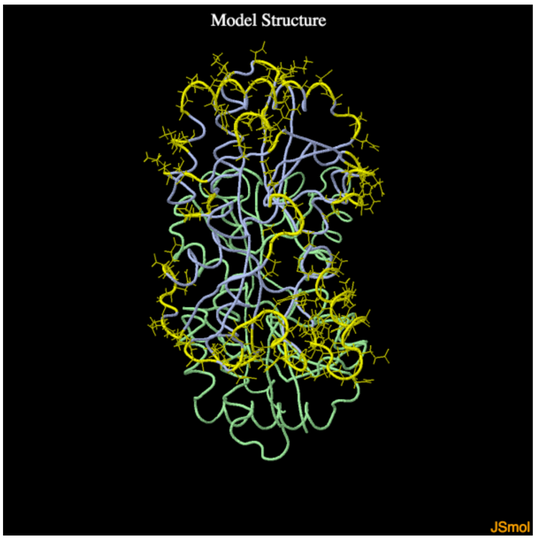


**Figure A57 DiscoTope2 predicted discontinuous epitopes for TcyJ.**

The parts coloured yellow are the predicted epitopes.

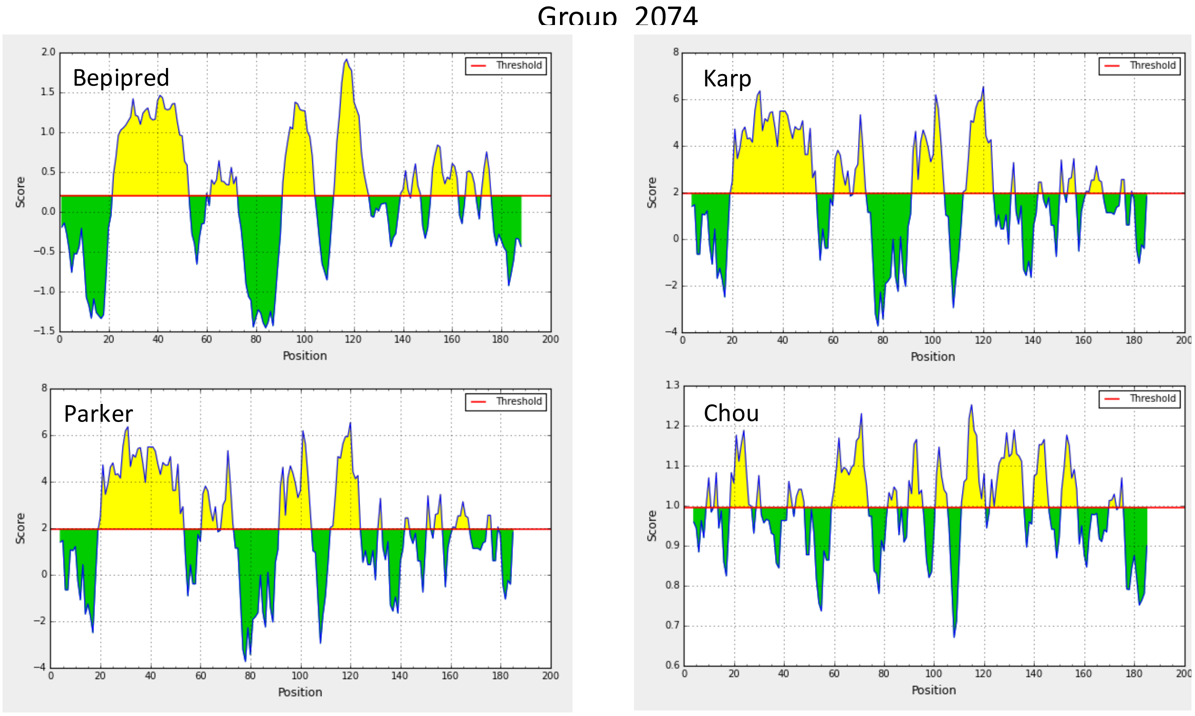


**Figure A58 Linear epitope predictions of the Group_2074 protein.**

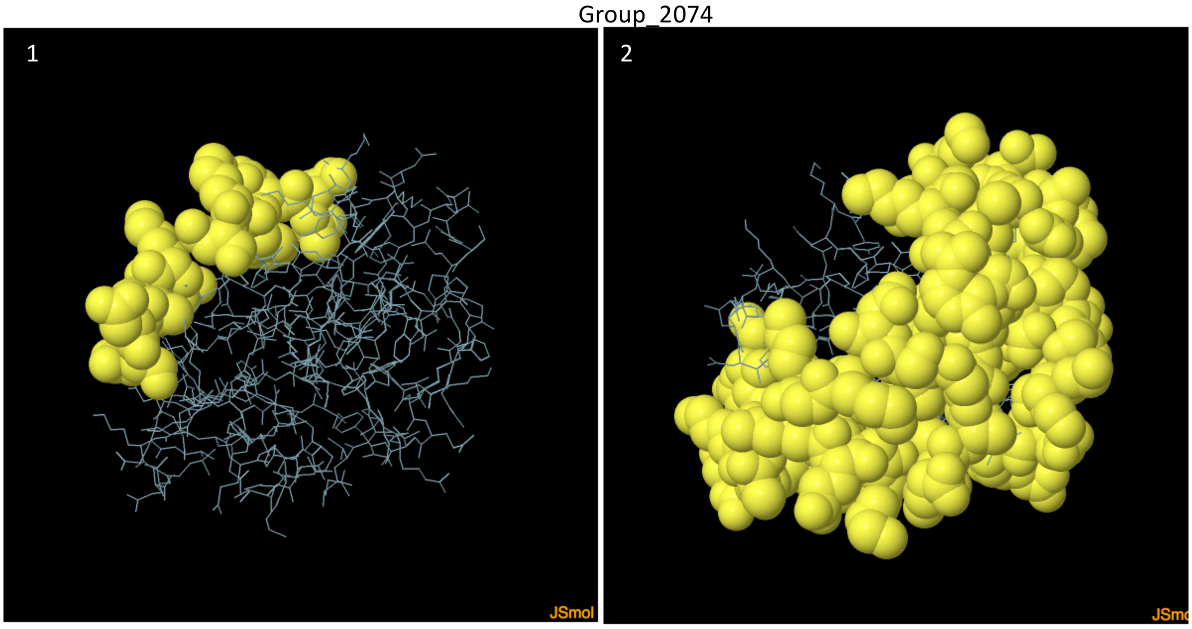


**Figure A59 ElliPro predicted discontinuous epitopes for Group_2074.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

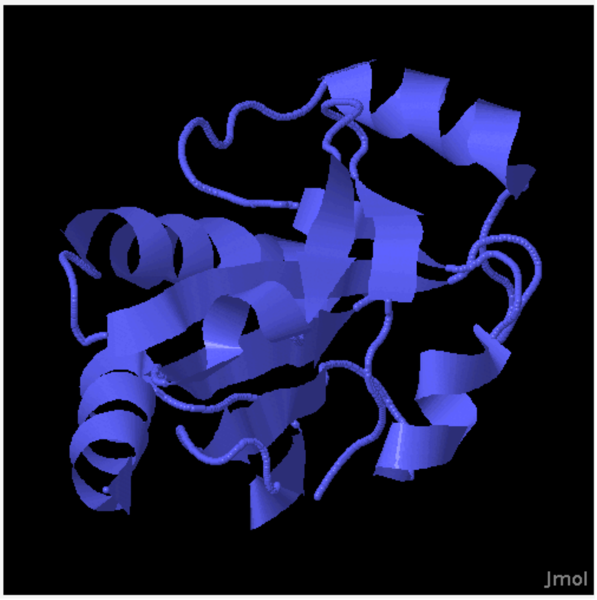


**Figure A60 DiscoTope2 predicted discontinuous epitopes for Group_2074.** The parts coloured yellow are the predicted epitopes.

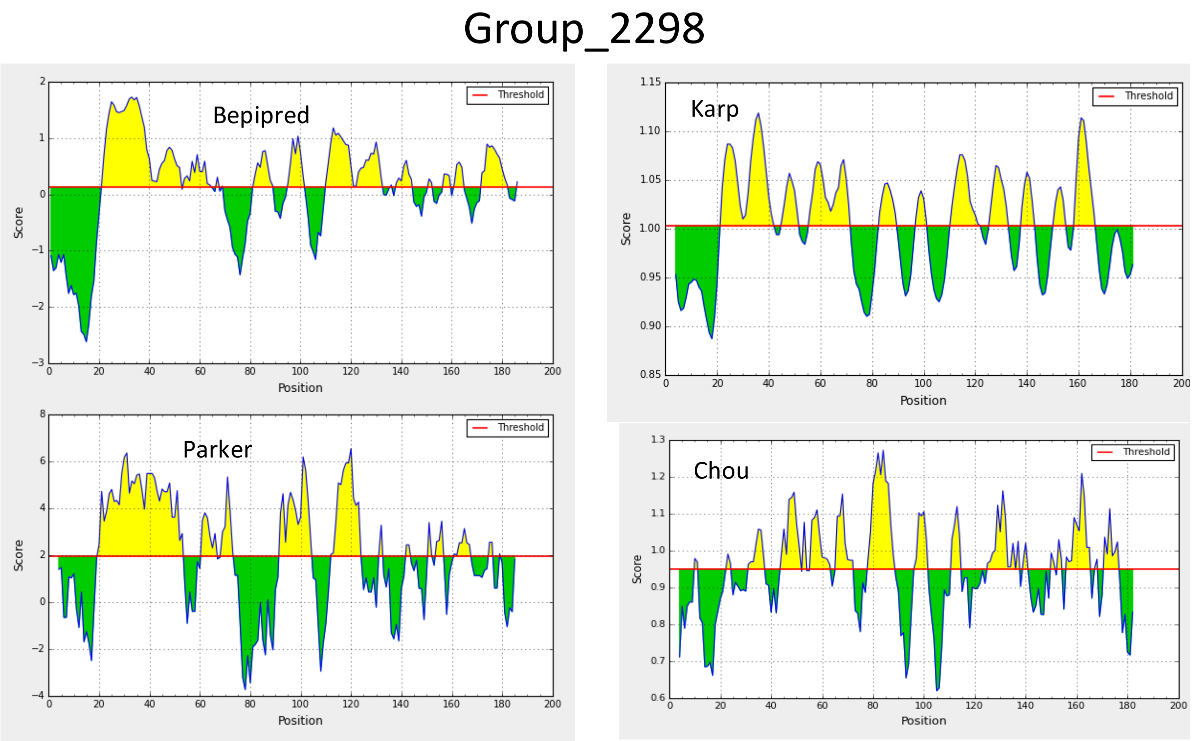


**Figure A61 Linear epitope predictions of the Group_2298 protein.**

## Group_2298

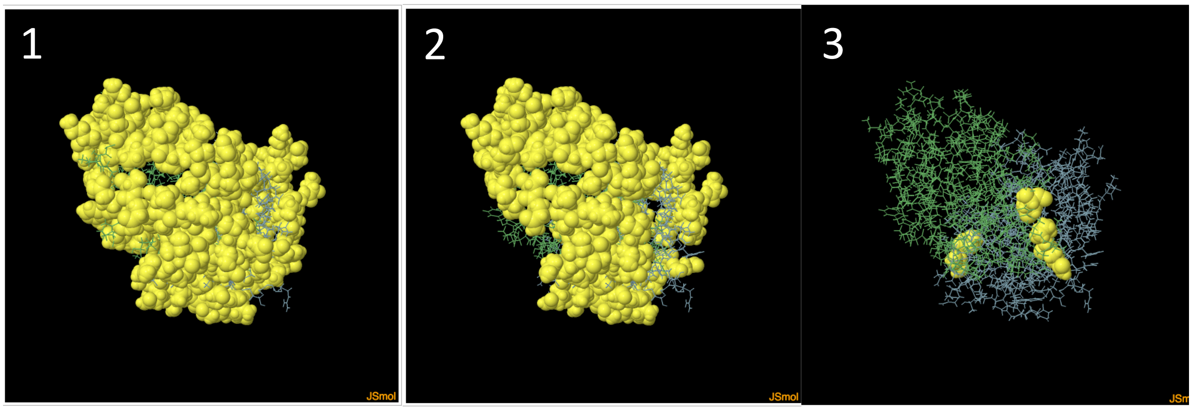


**Figure A62 ElliPro predicted discontinuous epitopes for Group_2298.**
The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

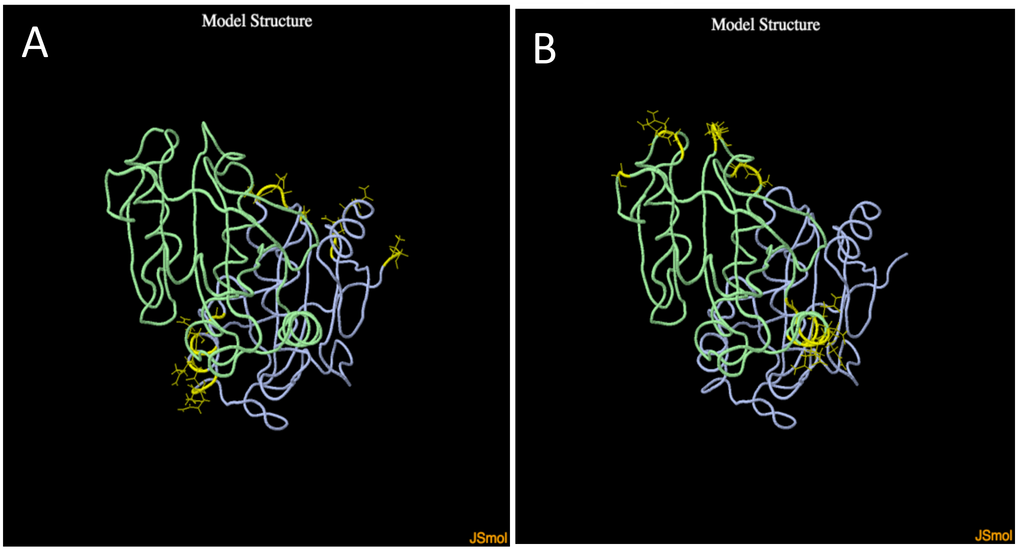


**Figure A63 DiscoTope2 predicted discontinuous epitopes for Group_2298.**
The parts coloured yellow are the predicted epitopes.

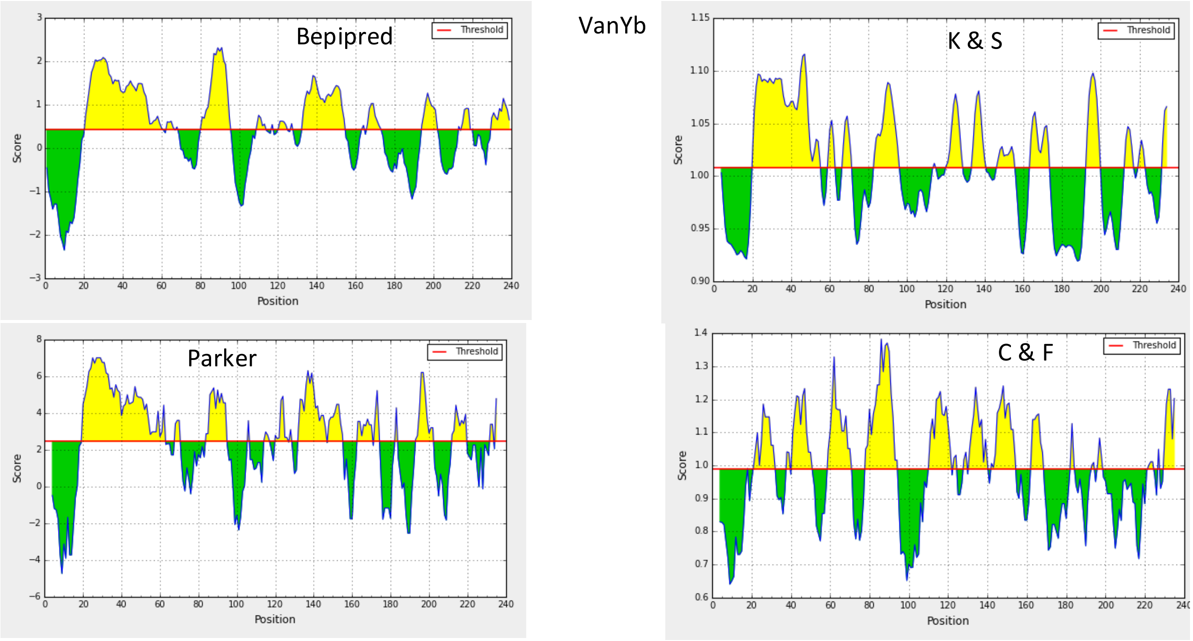


**Figure A64 Linear epitope predictions of the VanYb protein.**

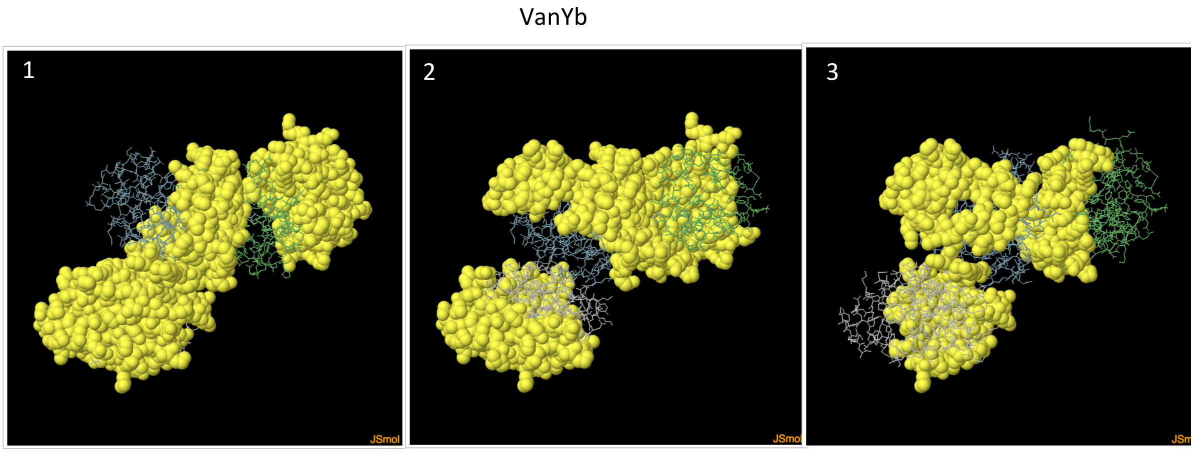


**Figure A65 ElliPro predicted discontinuous epitopes for VanYb.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

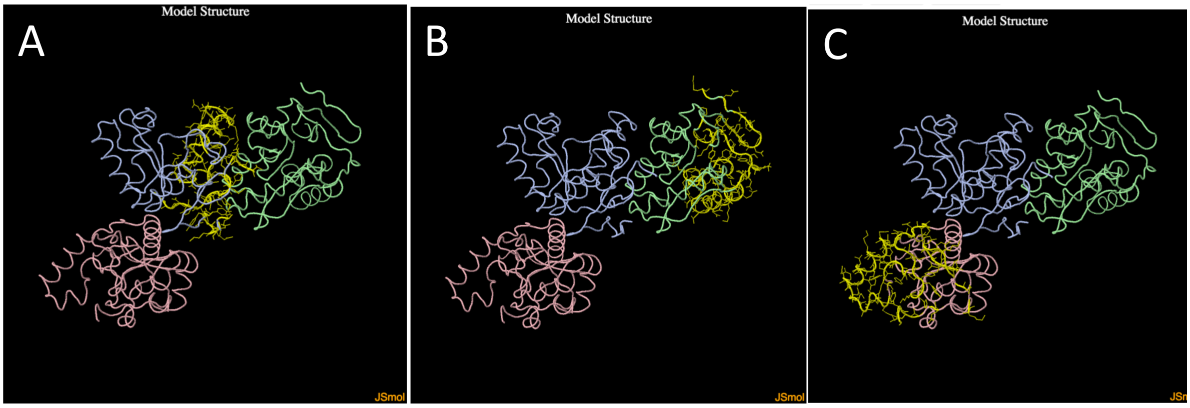


**Figure A66 DiscoTope2 predicted discontinuous epitopes for VanYb.**

The parts coloured yellow are the predicted epitopes.

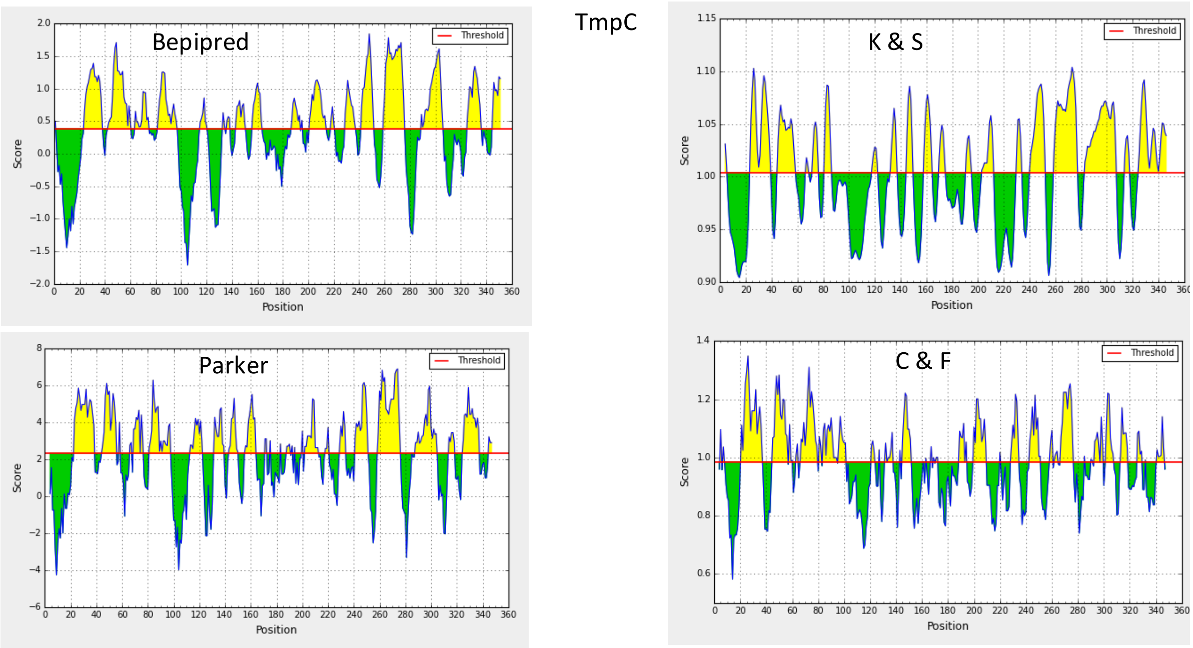


**Figure A67 Linear epitope predictions of the TmpC protein.**

TmpC

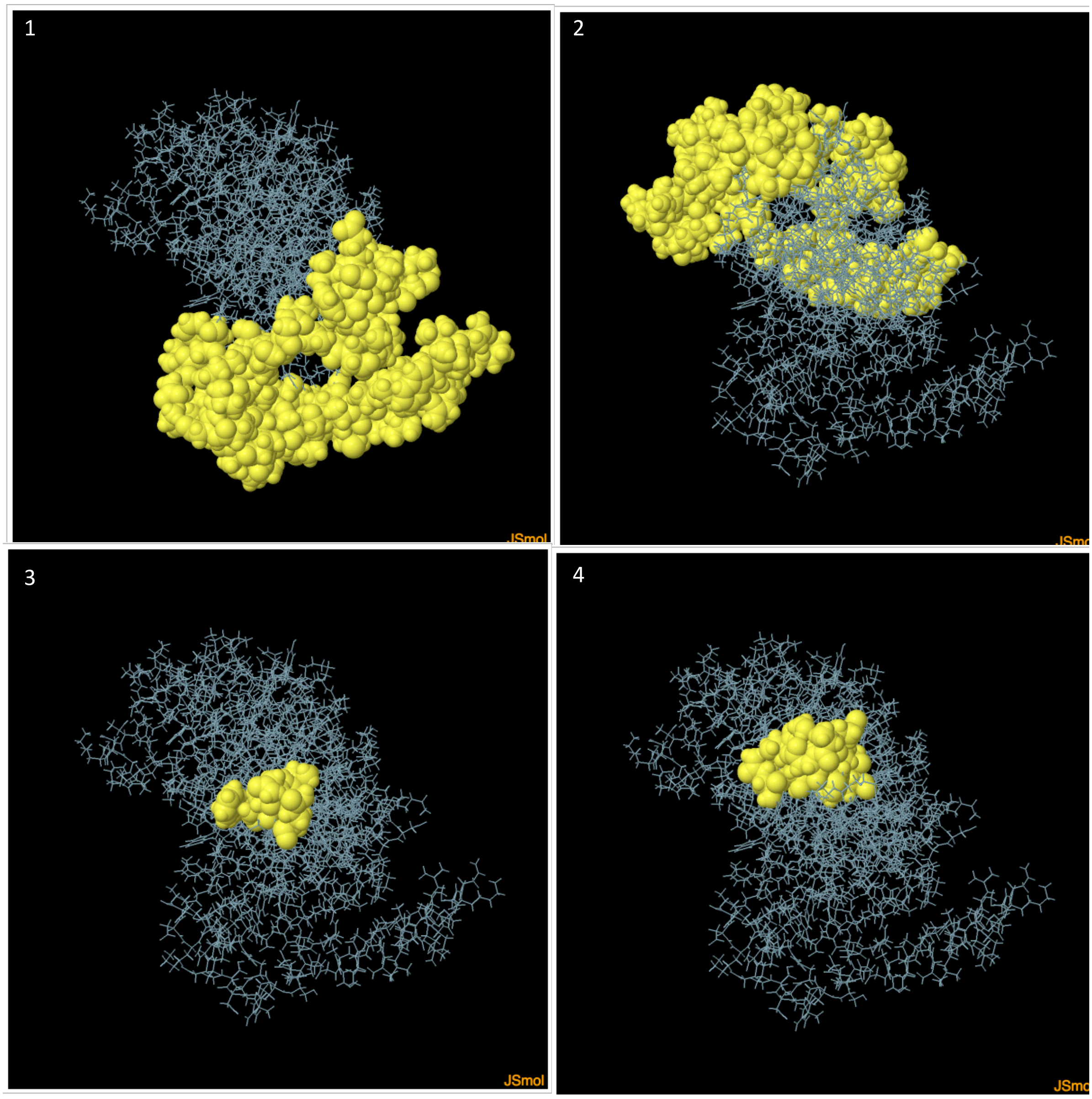


**Figure A68 ElliPro predicted discontinuous epitopes for TmpC.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

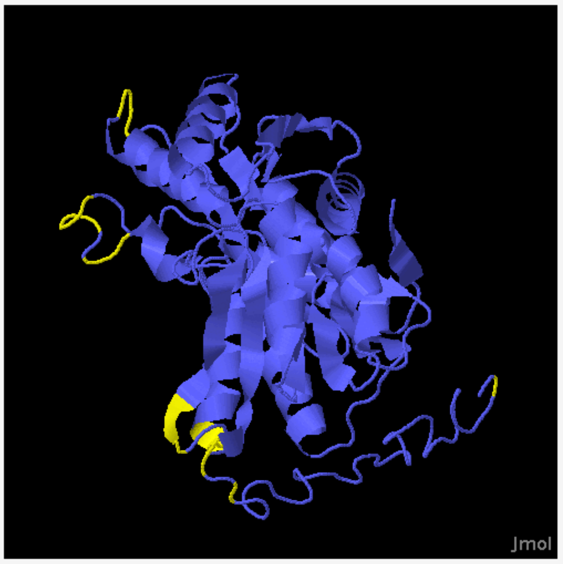


**Figure A69 DiscoTope2 predicted discontinuous epitopes for TmpC.**

The parts coloured yellow are the predicted epitopes.

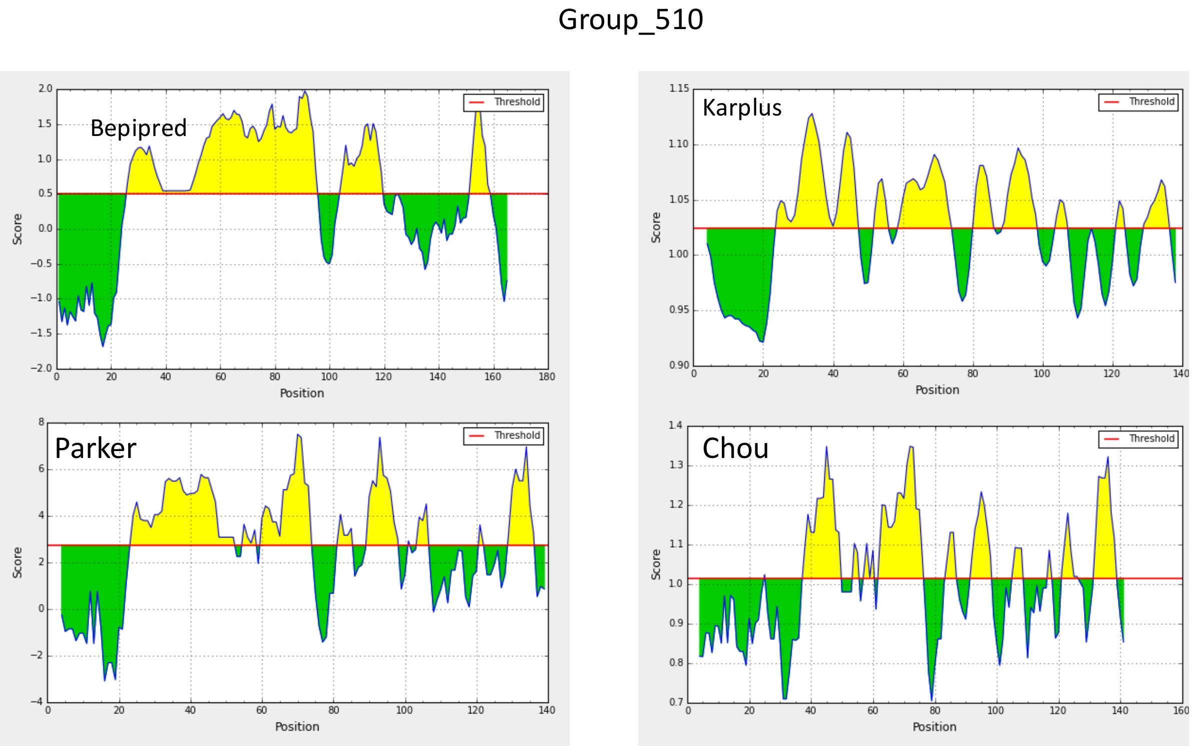


**Figure A70 Linear epitope predictions of the Group_510 protein.**

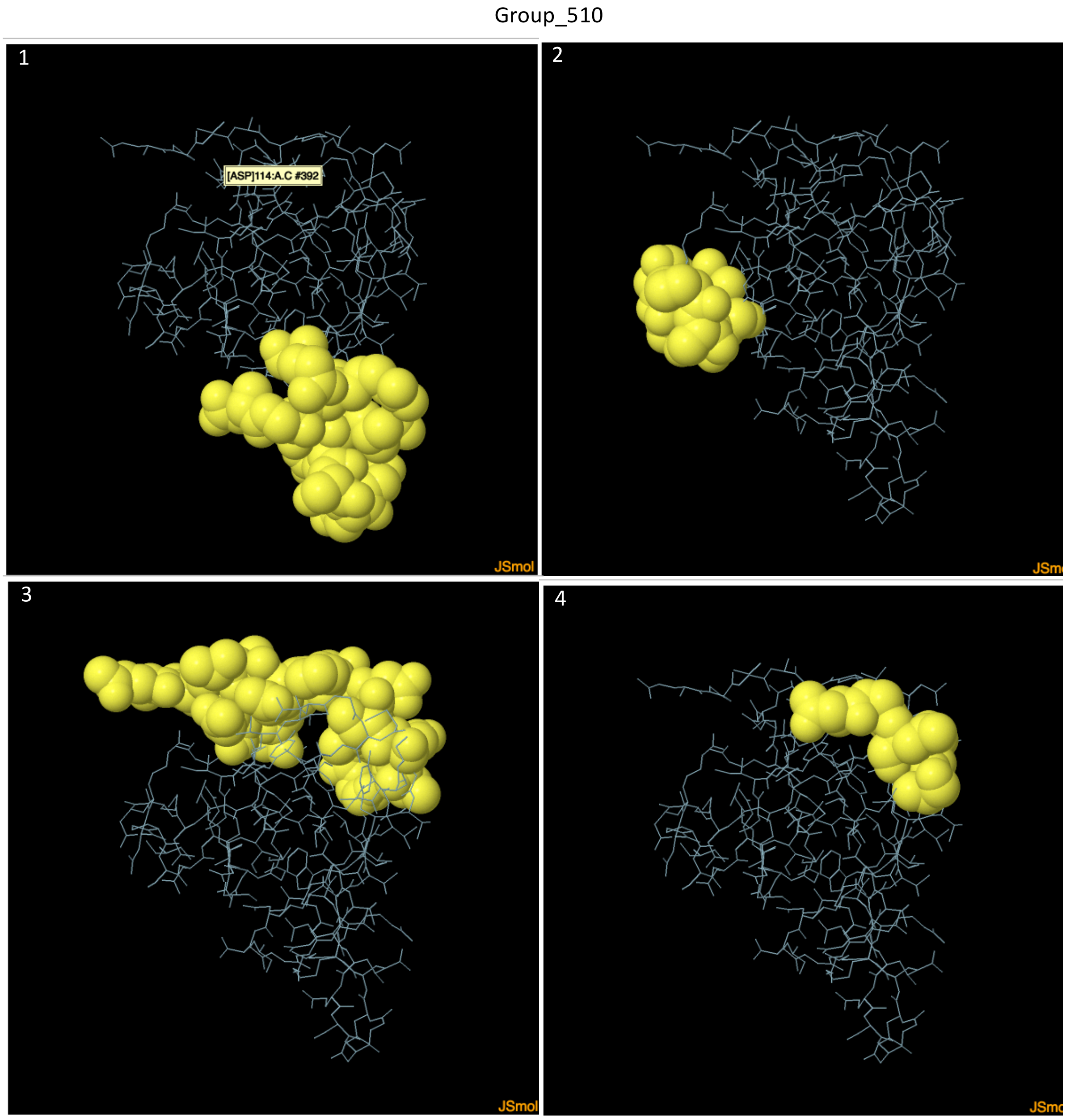


**Figure A71 ElliPro predicted discontinuous epitopes for Group_510.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

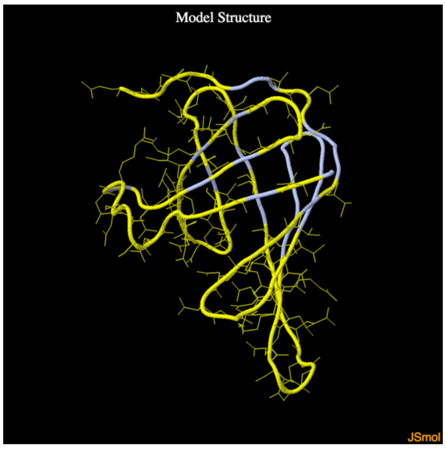


**Figure A72 DiscoTope2 predicted discontinuous epitopes for Group_510.**

The parts coloured yellow are the predicted epitopes.

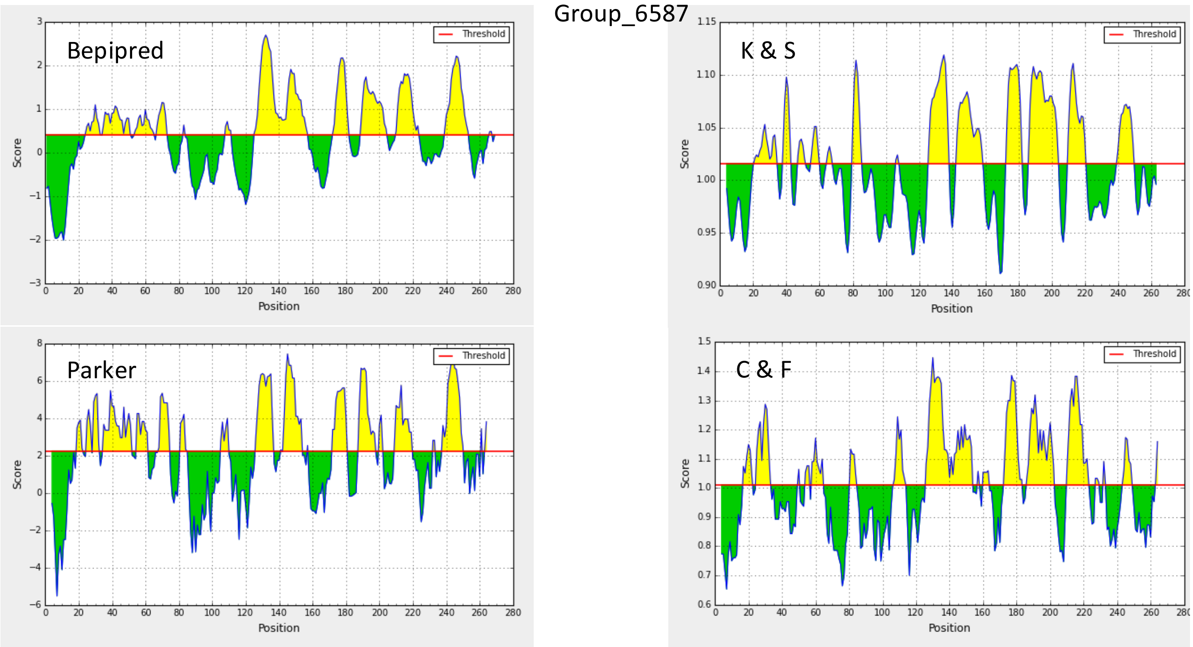


**Figure A73 Linear epitope predictions of the Group_6587 protein.**

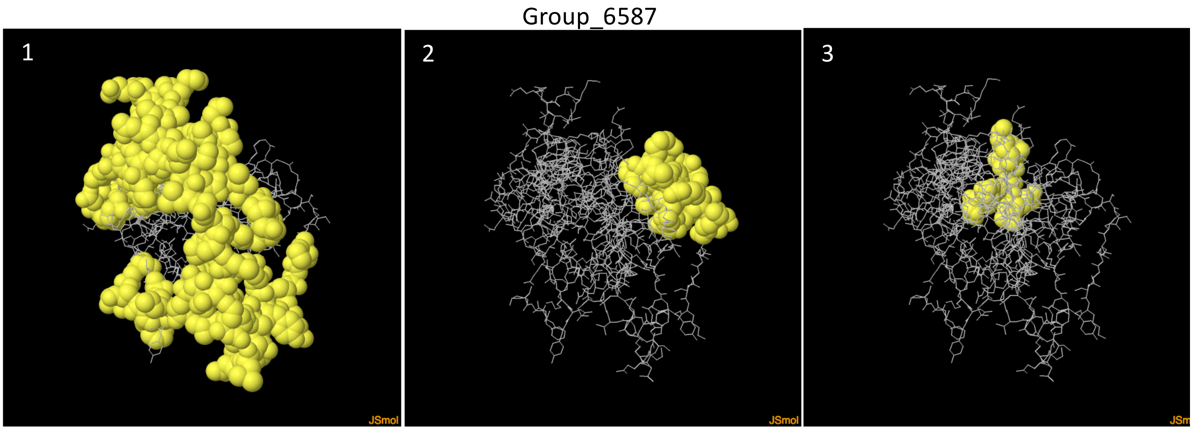


**Figure A74 ElliPro predicted discontinuous epitopes for Group_6587.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

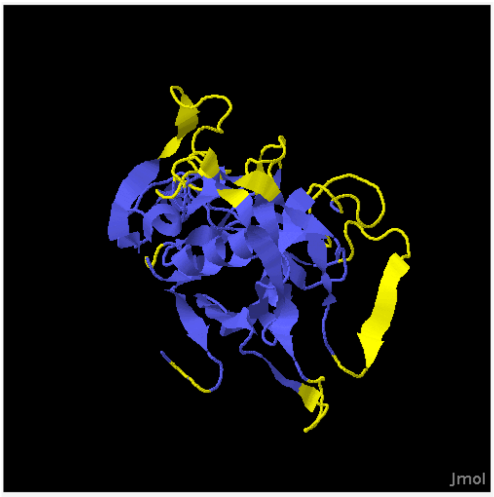


**Figure A75 DiscoTope2 predicted discontinuous epitopes for Group_6587.**

The parts coloured yellow are the predicted epitopes.

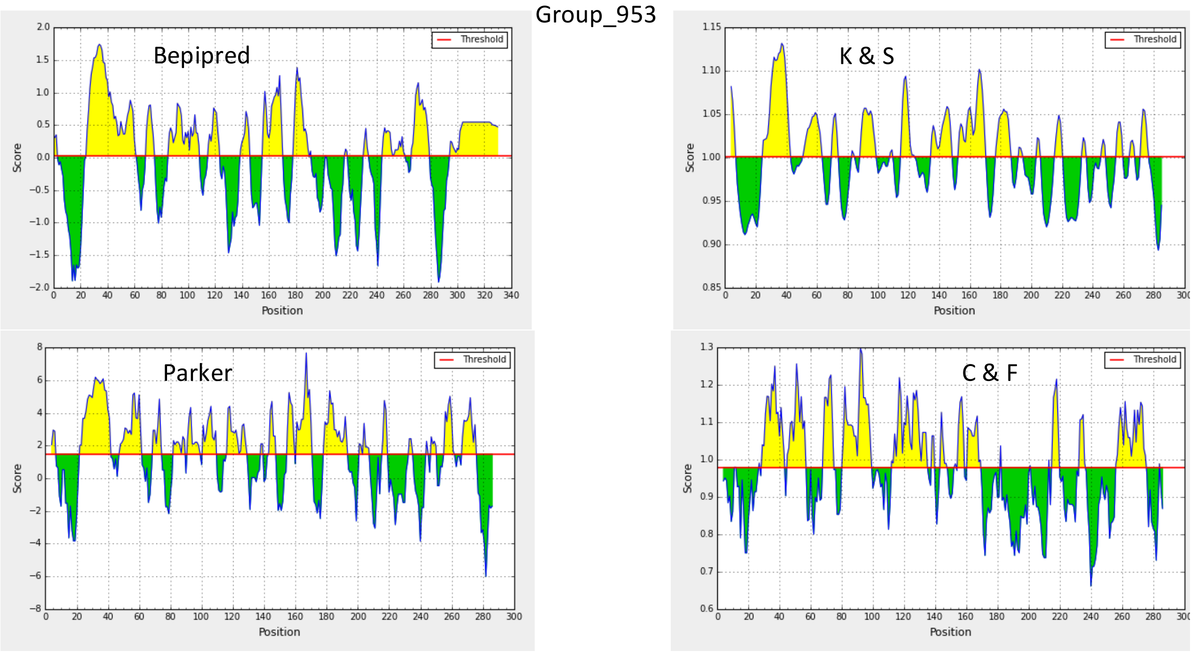


**Figure A76 Linear epitope predictions of the Group_953 protein.**

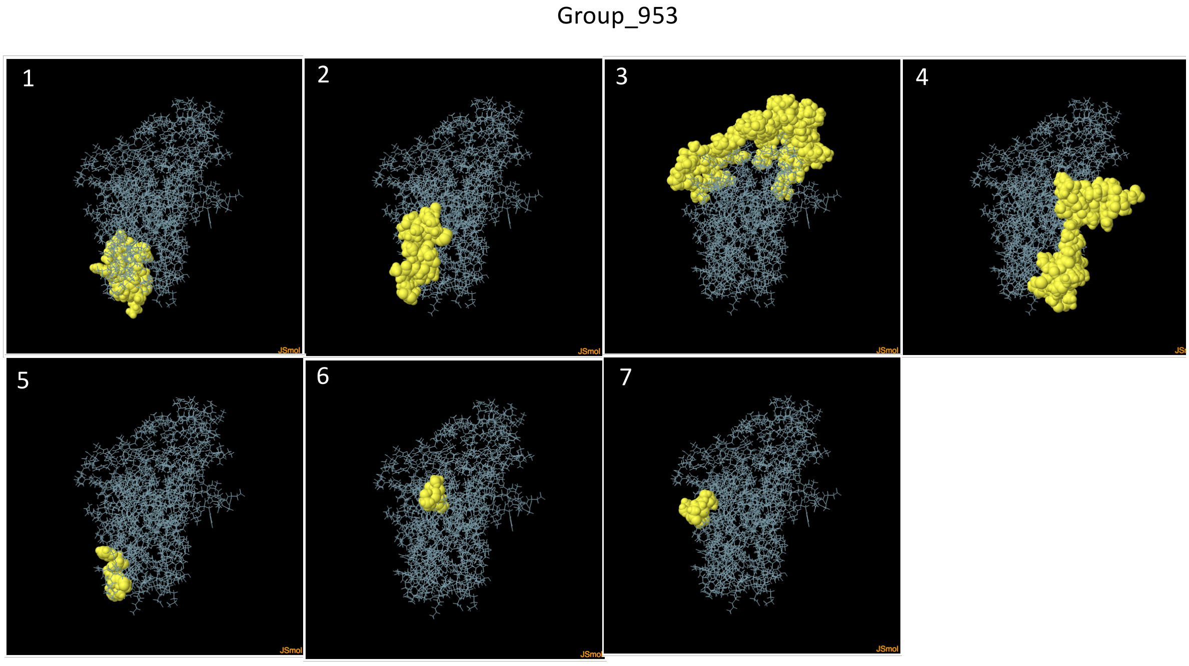


**Figure A77 ElliPro predicted discontinuous epitopes for Group_953.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

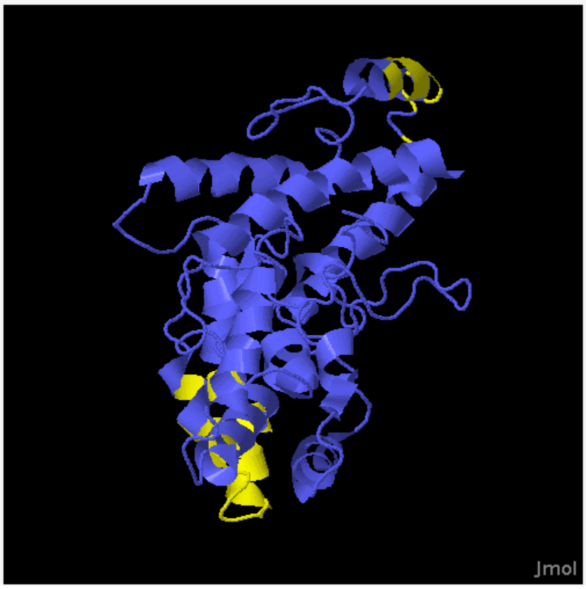


**Figure A78 DiscoTope2 predicted discontinuous epitopes for Group_953.**

The parts coloured yellow are the predicted epitopes.

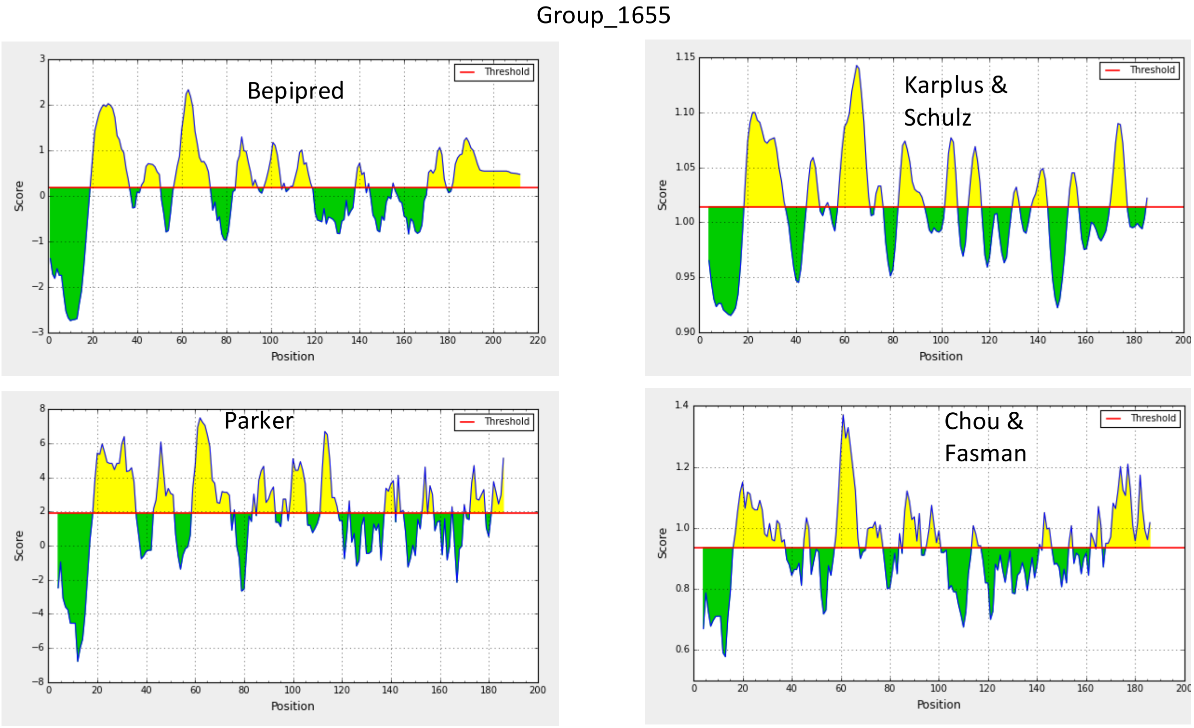


**Figure A79 Linear epitope predictions of the Group_1655 protein.**

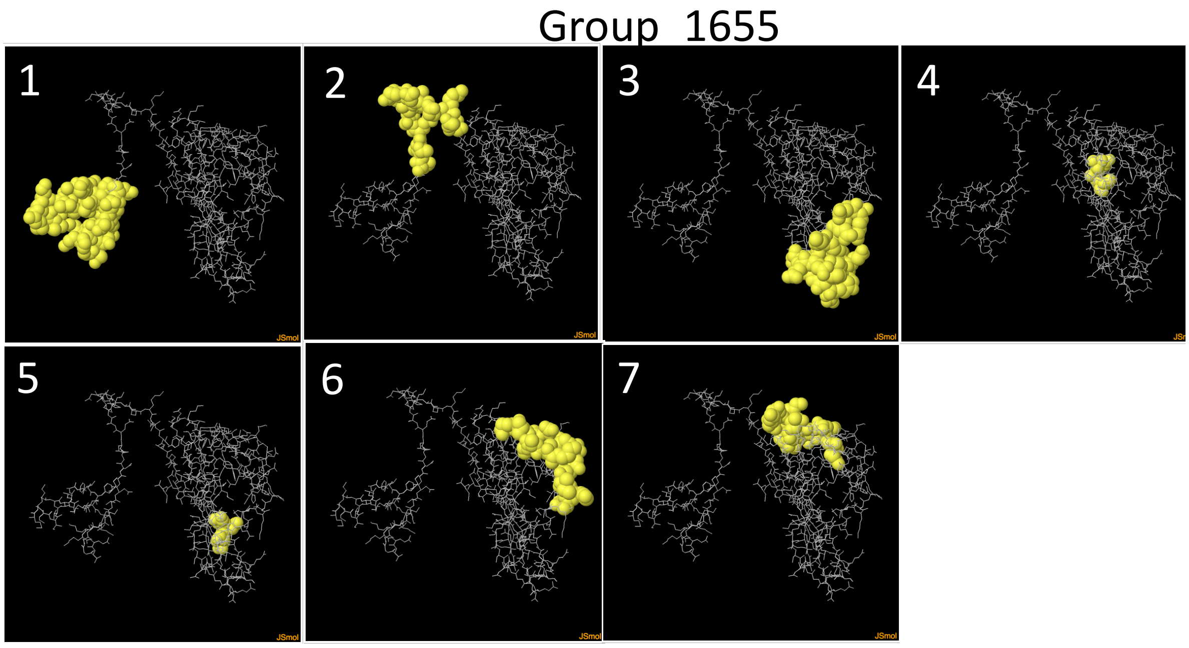


**Figure A80 ElliPro predicted discontinuous epitopes for Group_1655.**

The numbers represent the different epitopes predicted in order of decreasing overall score with 1 having the highest score. The yellow spheres represent residues that are part of the predicted epitope.

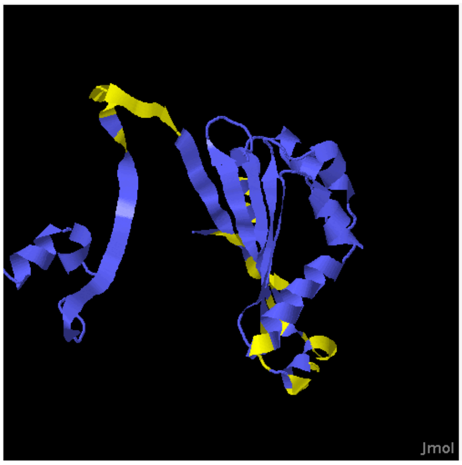


**Figure A81 DiscoTope2 predicted discontinuous epitopes for Group_1655.**

The parts coloured yellow are the predicted epitopes.