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.

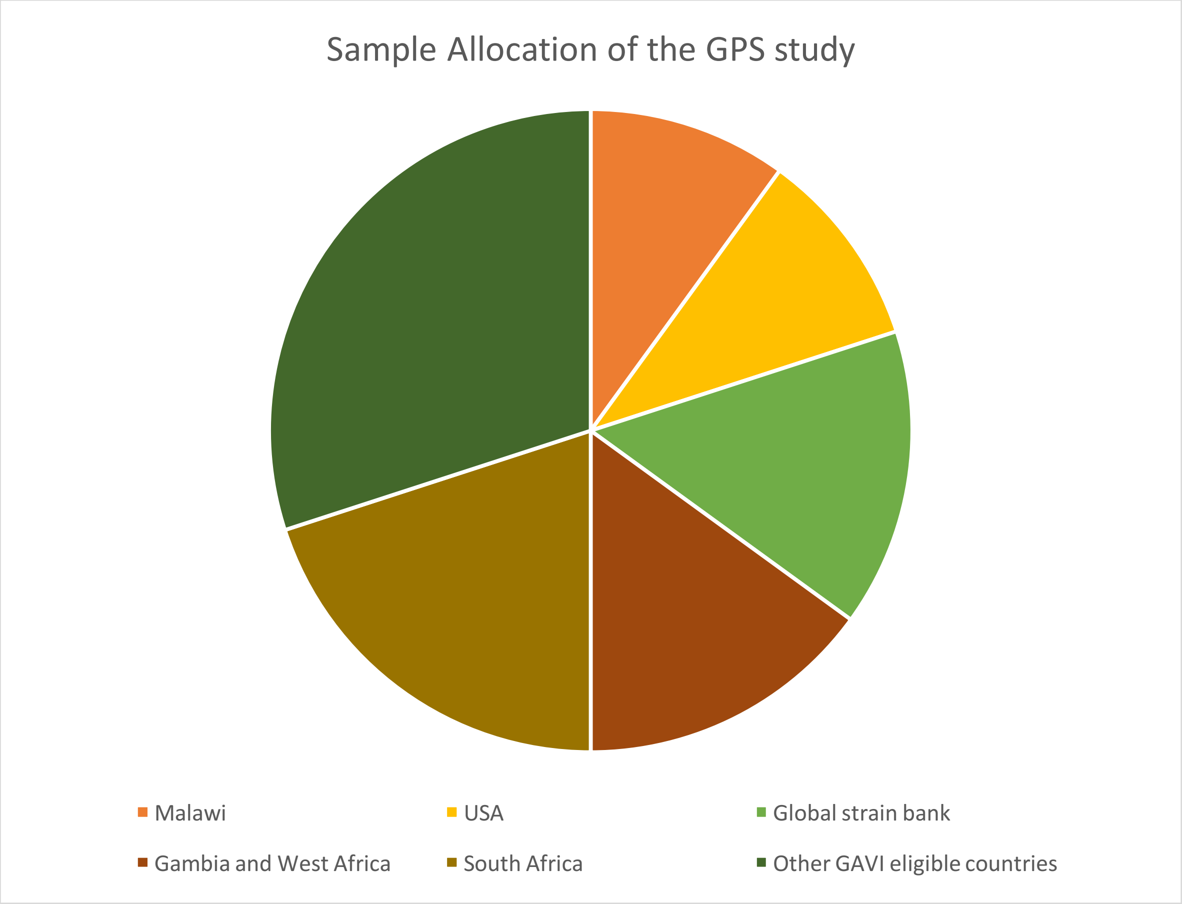

Figure 2.1 GPS sample collection sites.

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].

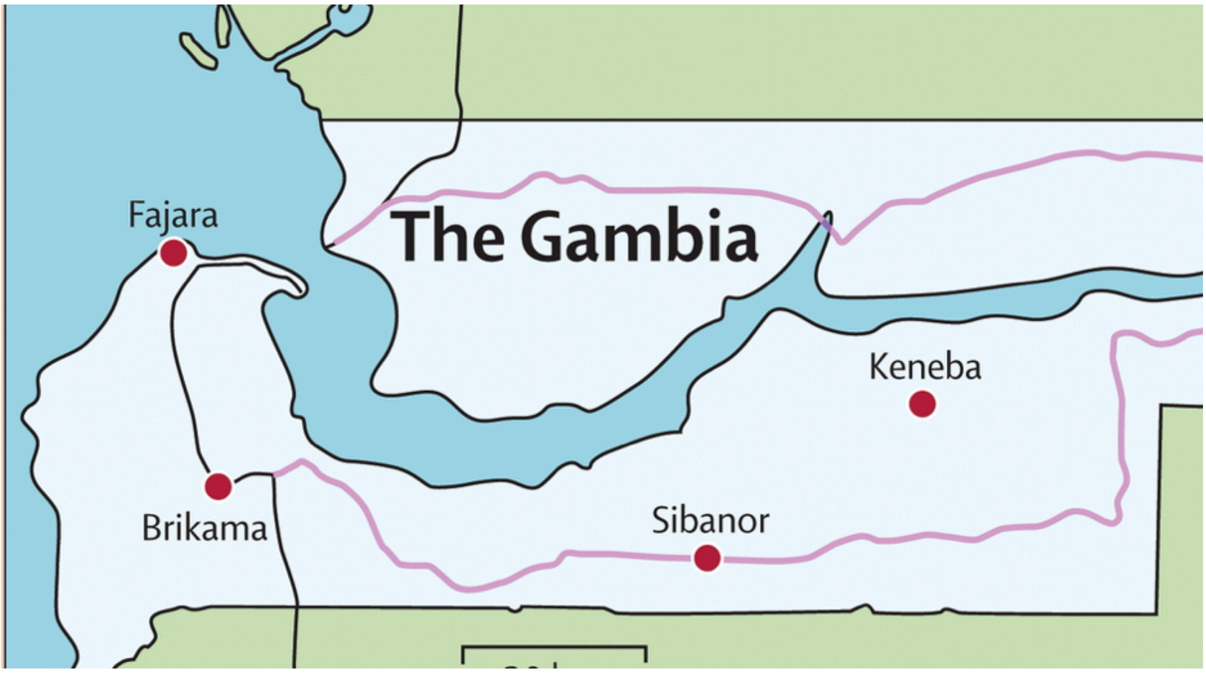

Figure 2.2 Partial map of The Gambia.

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.

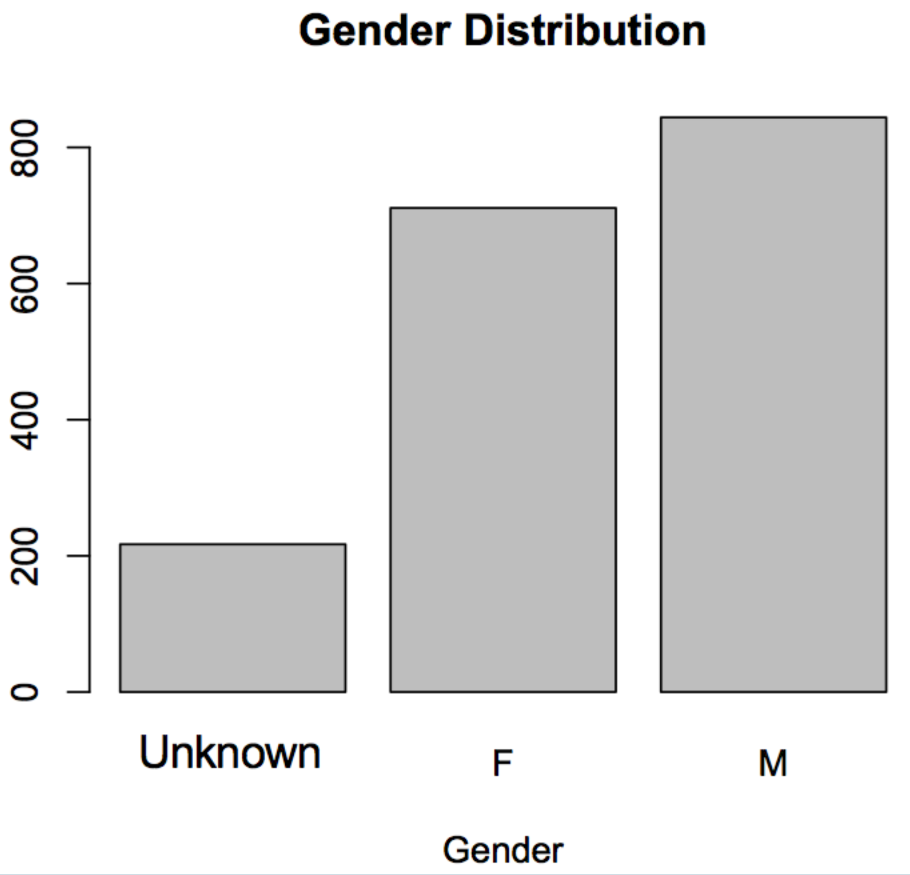

Figure 2.3 Gender distribution of the isolates in this dataset.

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].

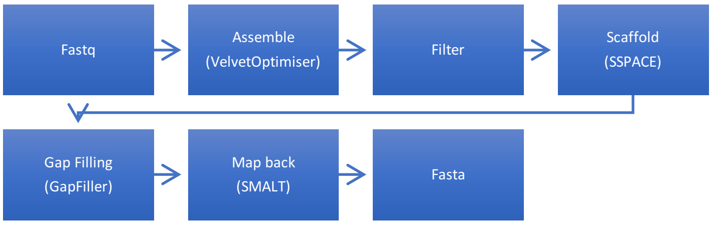

Figure 2.4 *De novo* assembly with the Sanger Institute pipeline.

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.

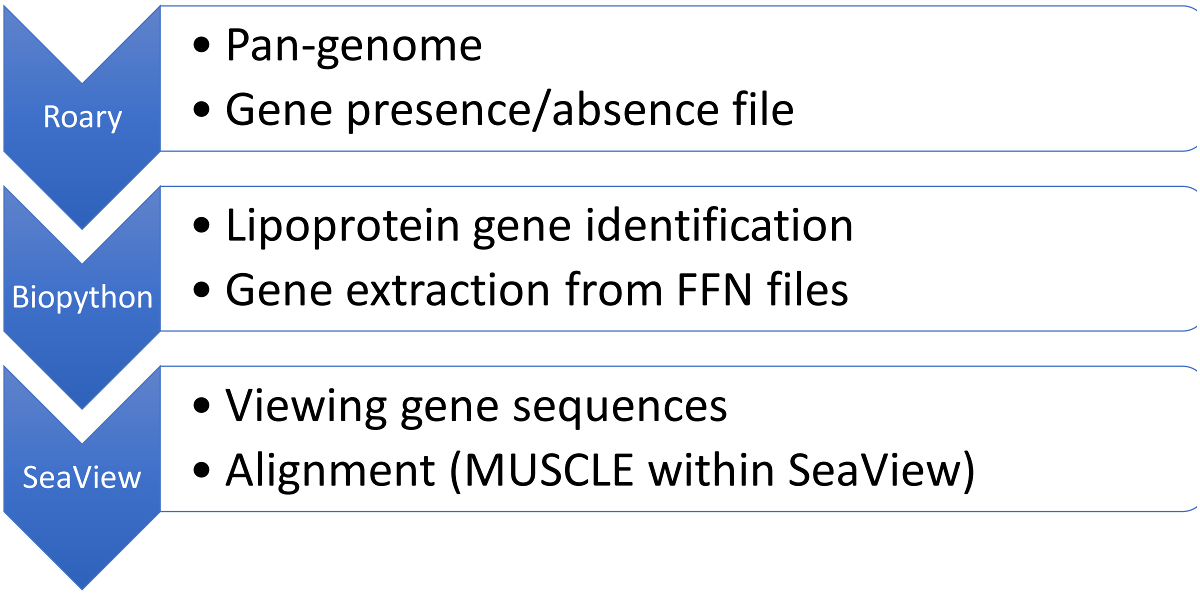
Roary
• Pan-genome
• Gene presence/absence file
Biopython
• Lipoprotein gene identification
• Gene extraction from FFN files
SeaView
• Viewing gene sequences
• Alignment (MUSCLE within SeaView)

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 α-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 $\geq$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.