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.

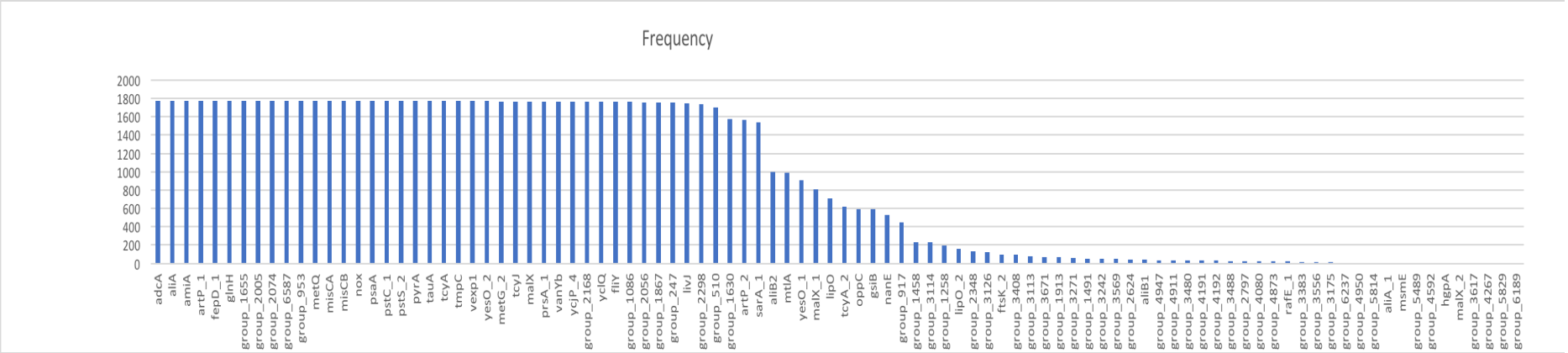


Figure 3.1 This figure shows a subset of the lipoproteins from the pattern searches and their prevalence.

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.

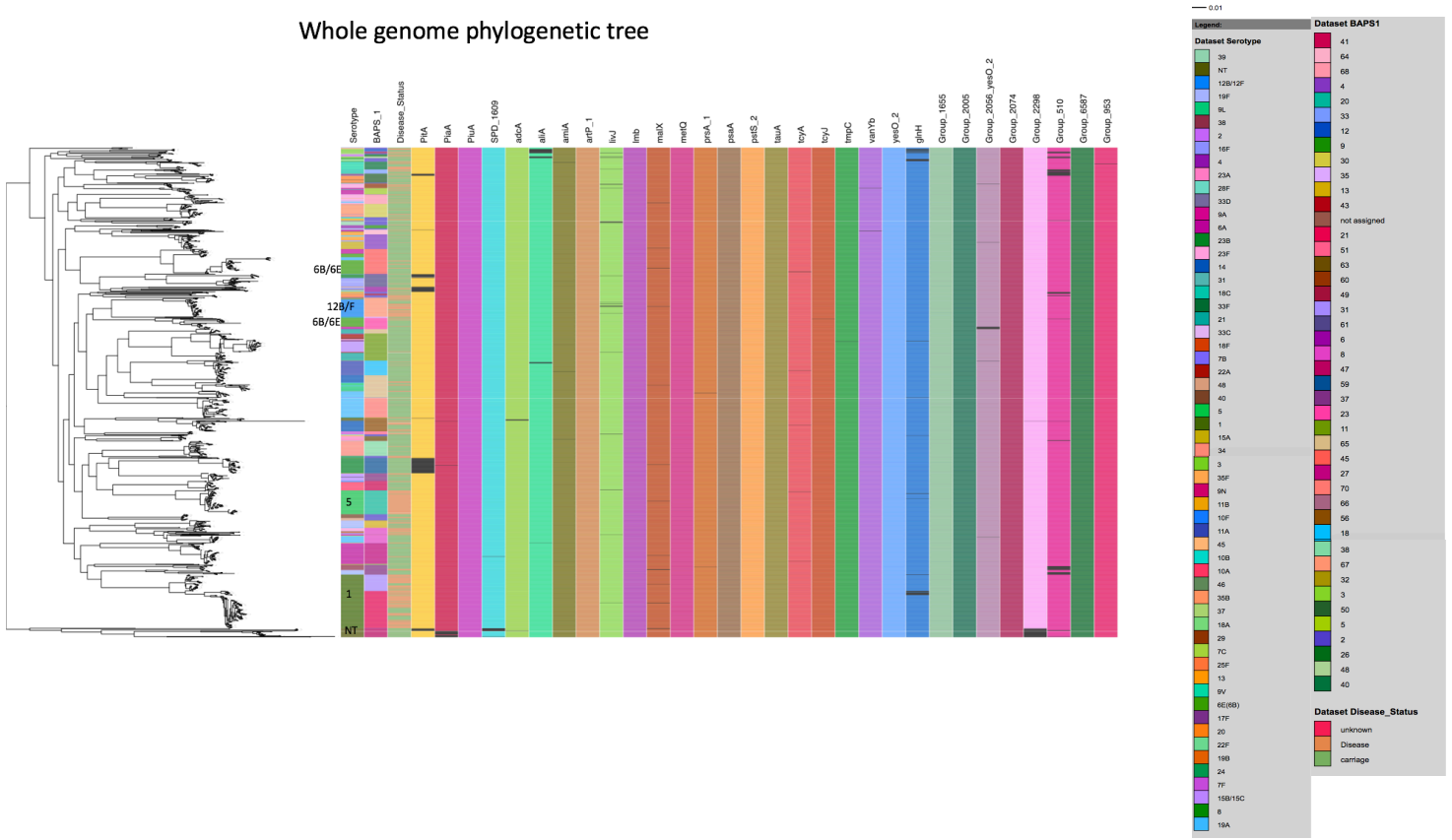


Figure 3.2 Whole genome phylogenetic tree with candidate gene presence/absence profiles.

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.

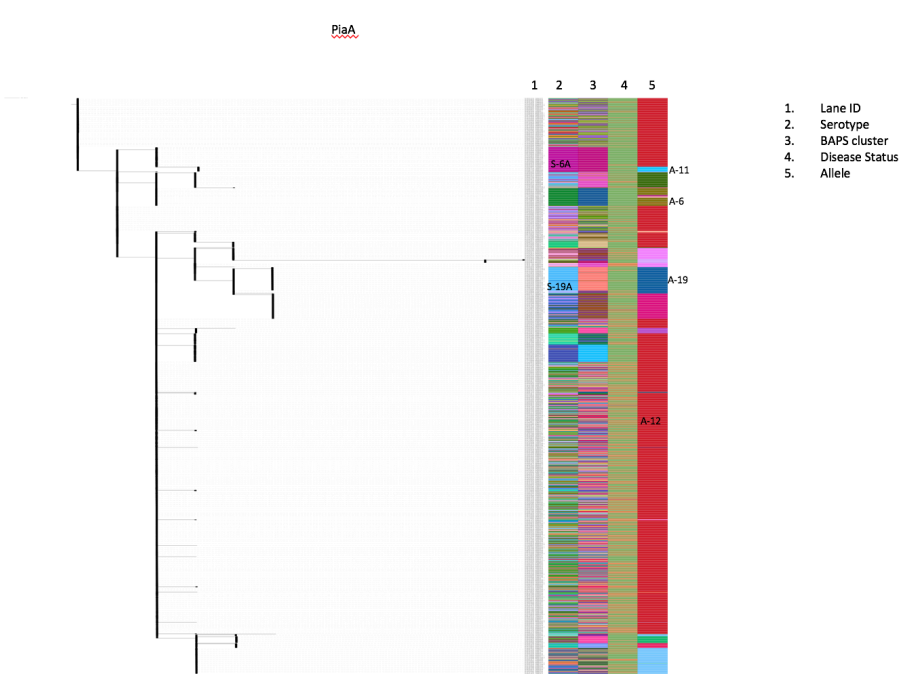
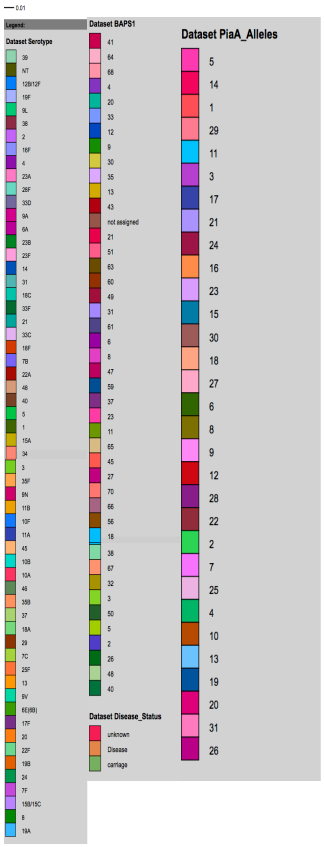


Figure 3.3 Phylogenetic gene tree of *piaA*.

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.

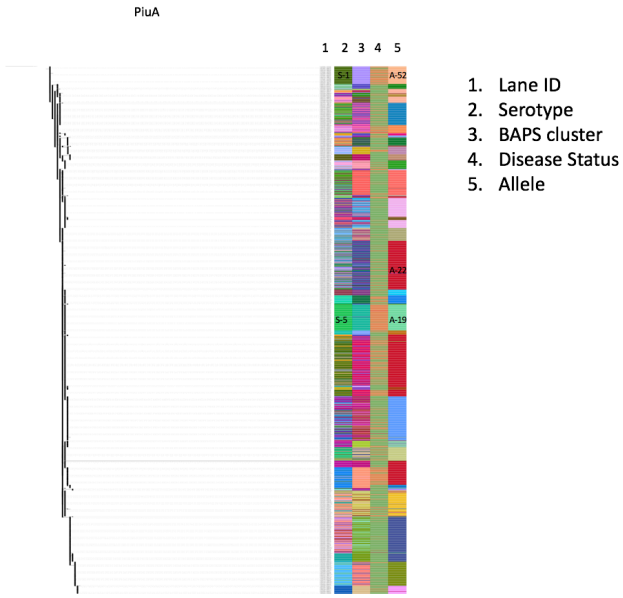
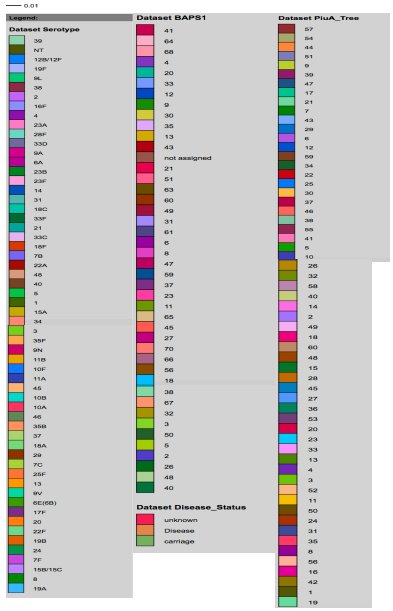


Figure 3.4 Phylogenetic gene tree of *piuA*.

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.

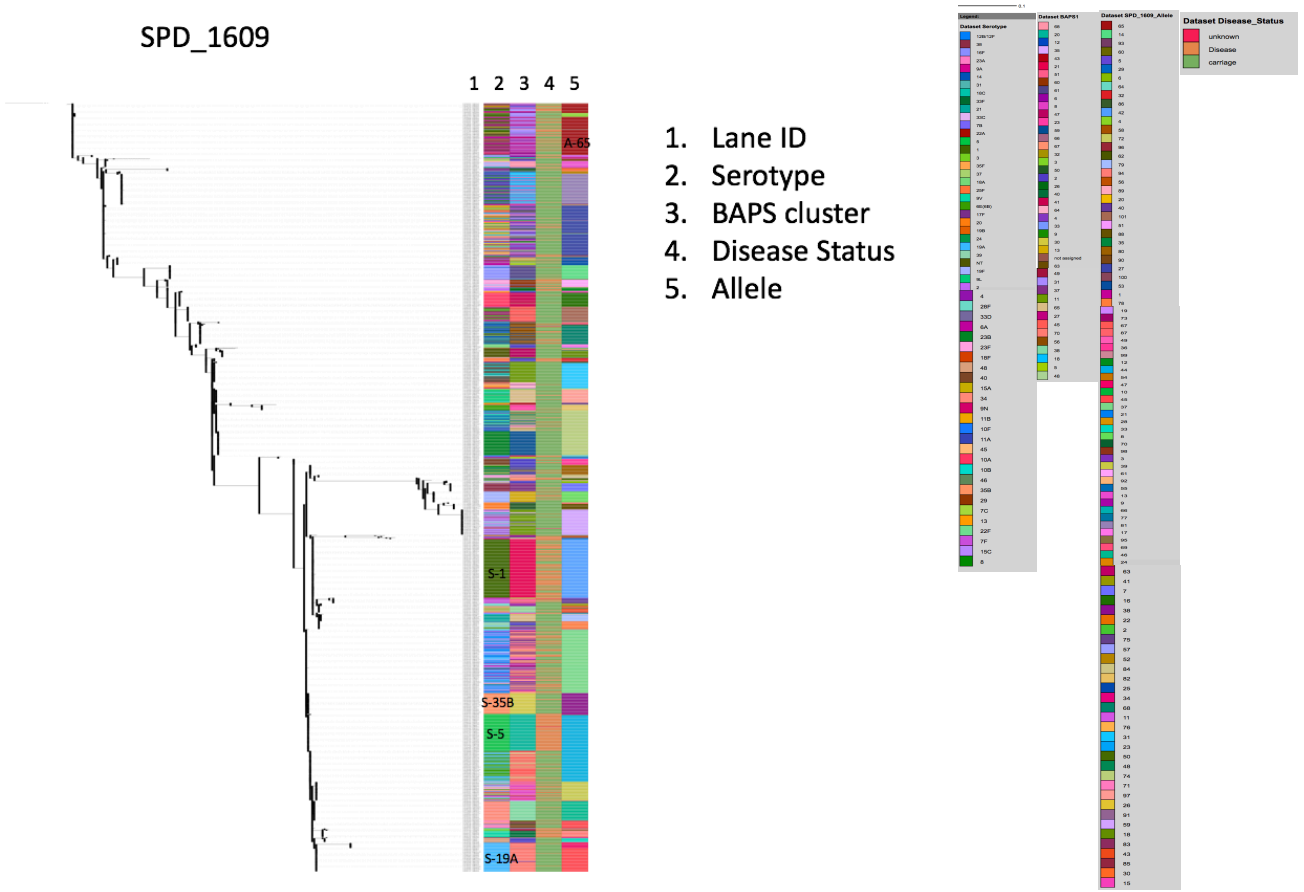


Figure 3.5 Phylogenetic gene tree of *SPD_1609*.

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.

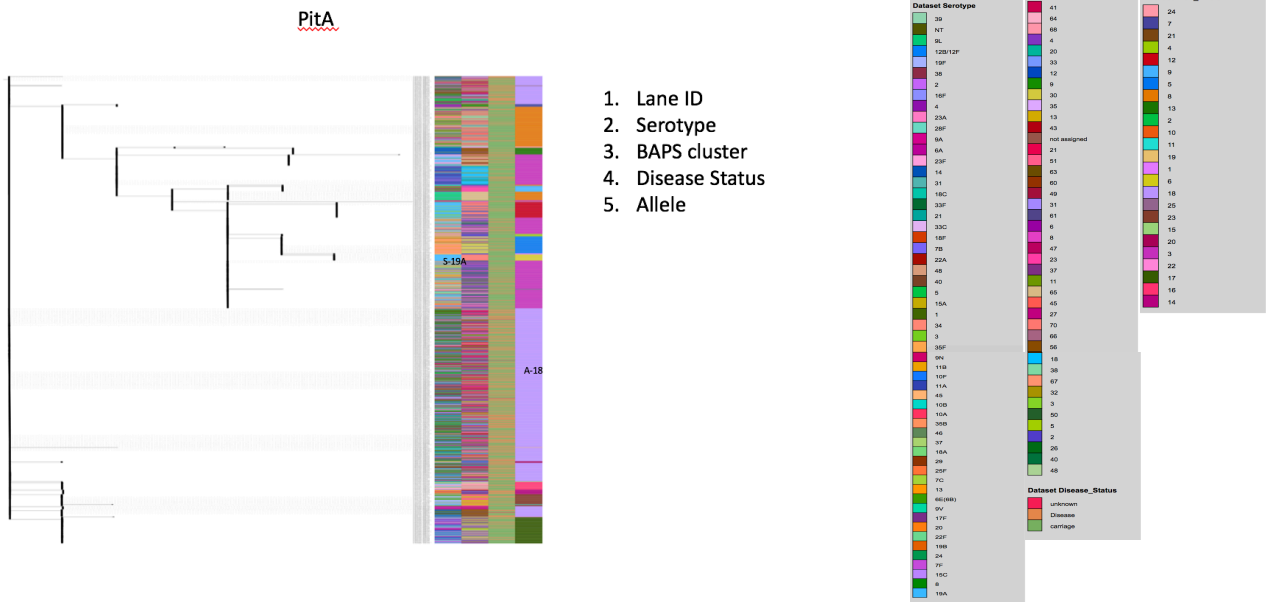


Figure 3.6 Phylogenetic gene tree of *pitA*.

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.

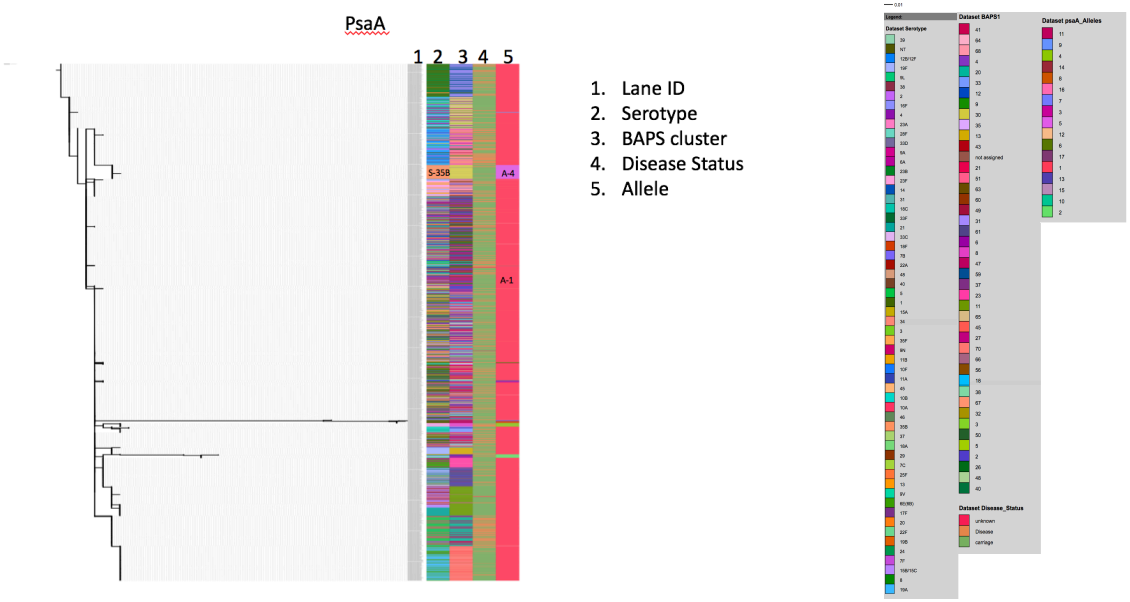


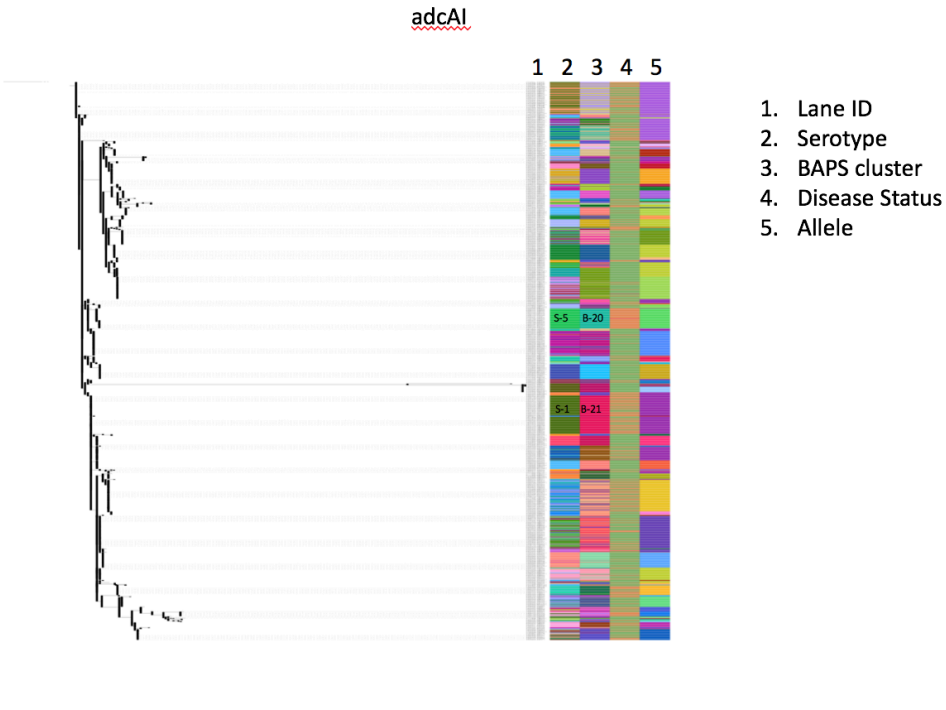
Figure 3.7 Phylogenetic gene tree of psaA.

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).



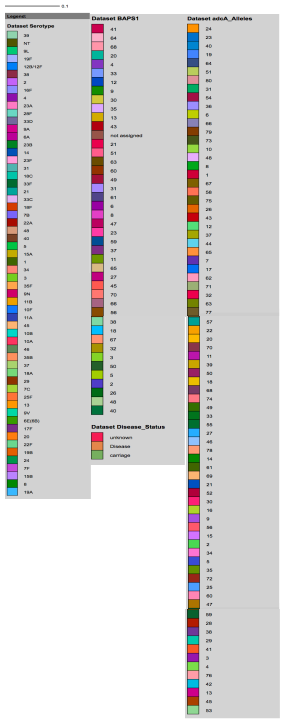
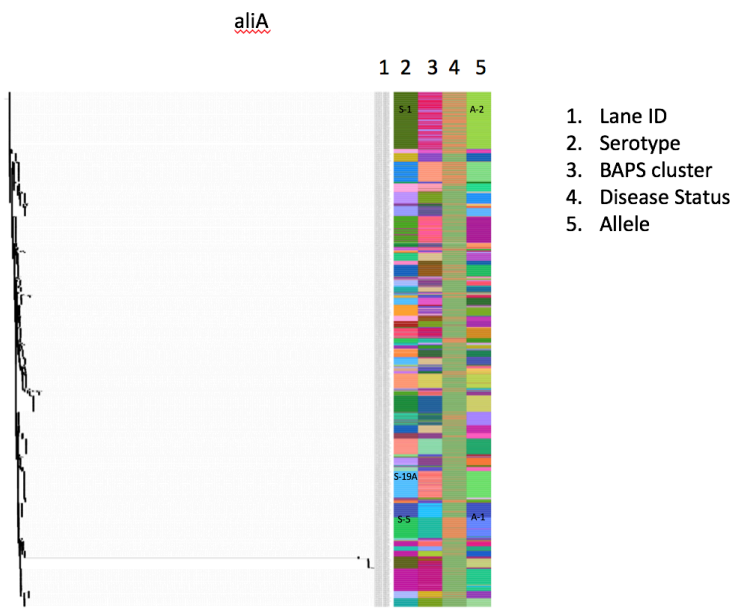


Figure 3.8 Phylogenetic gene tree of *adcA*.

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

190 SNP sites used to reconstruct this gene tree.

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.



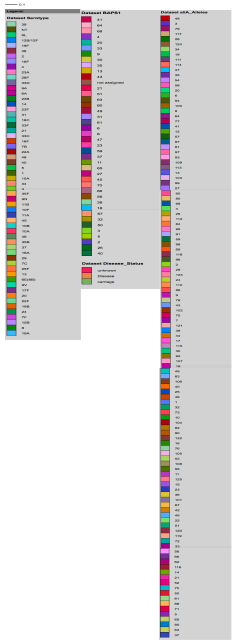


Figure 3.9 Phylogenetic gene tree of *aliA*.

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).

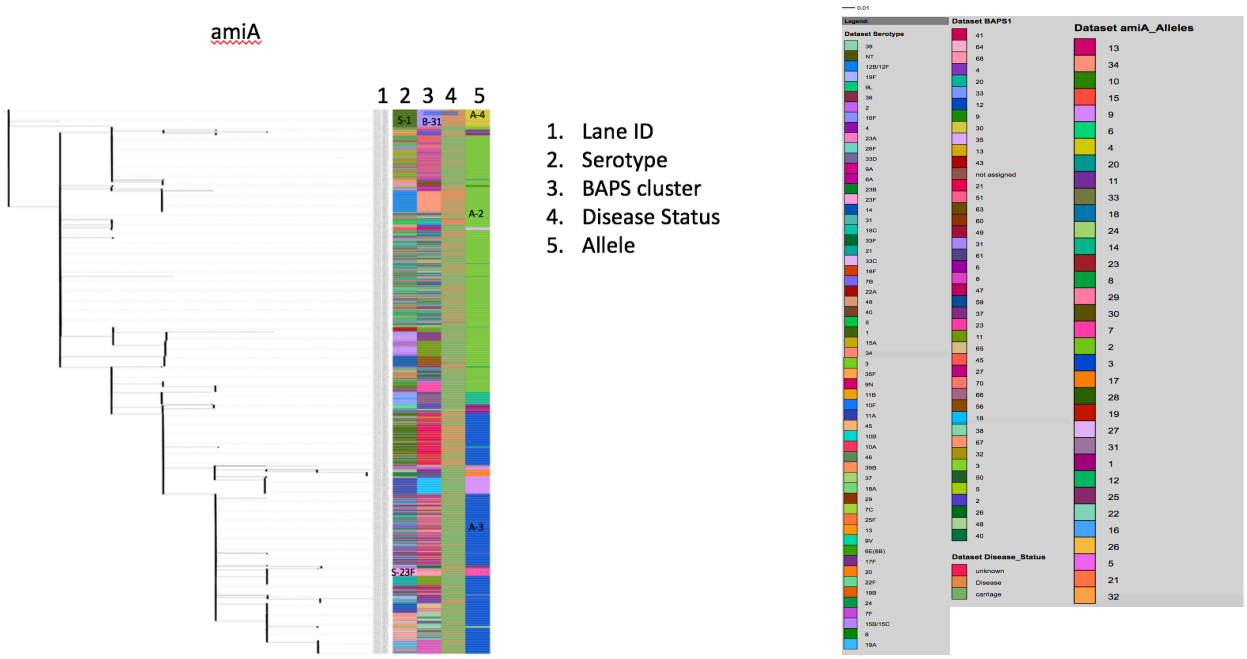
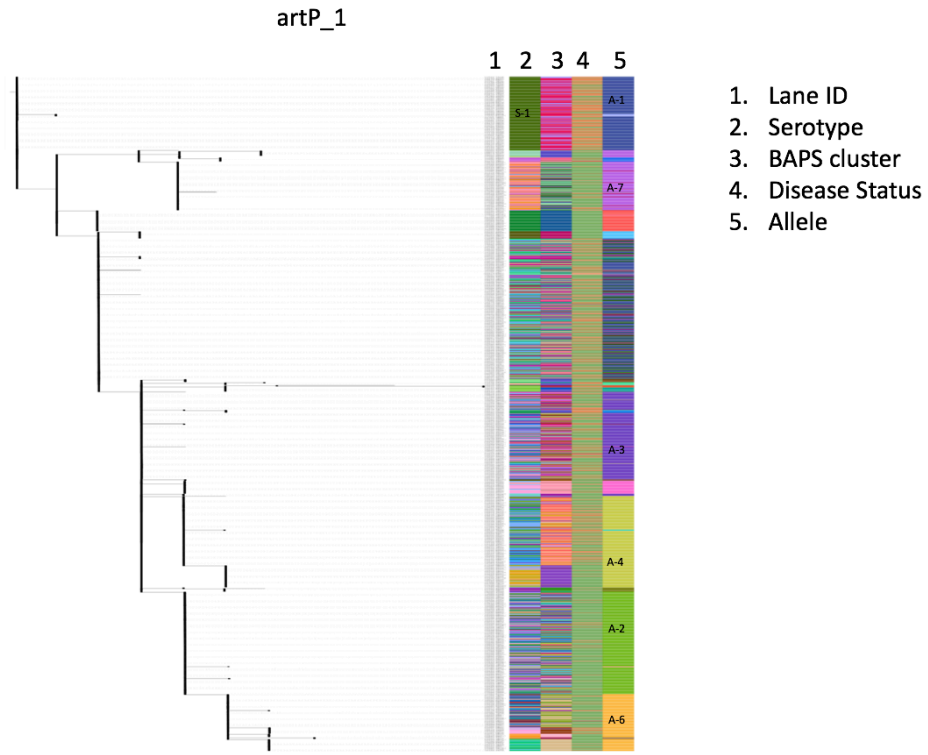


Figure 3.10 Phylogenetic gene tree of *amiA*.

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.



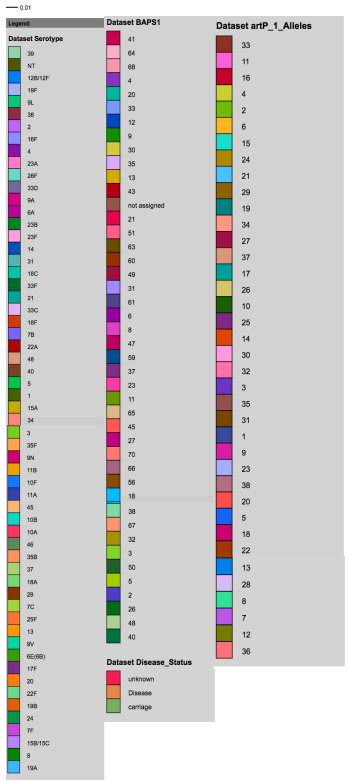


Figure 3.11 Phylogenetic gene tree of *artP_1*.

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

48 SNP sites used to reconstruct this gene tree.

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.

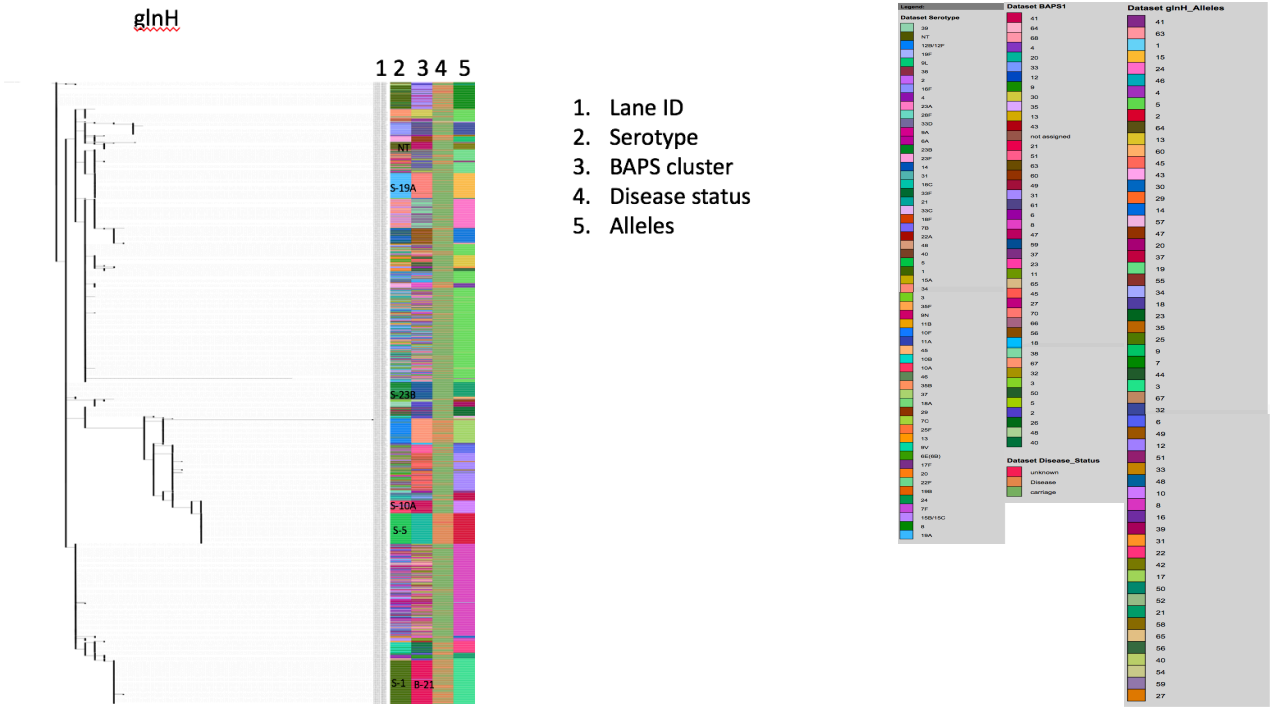


Figure 3.12 Phylogenetic gene tree of *glnH*.

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.

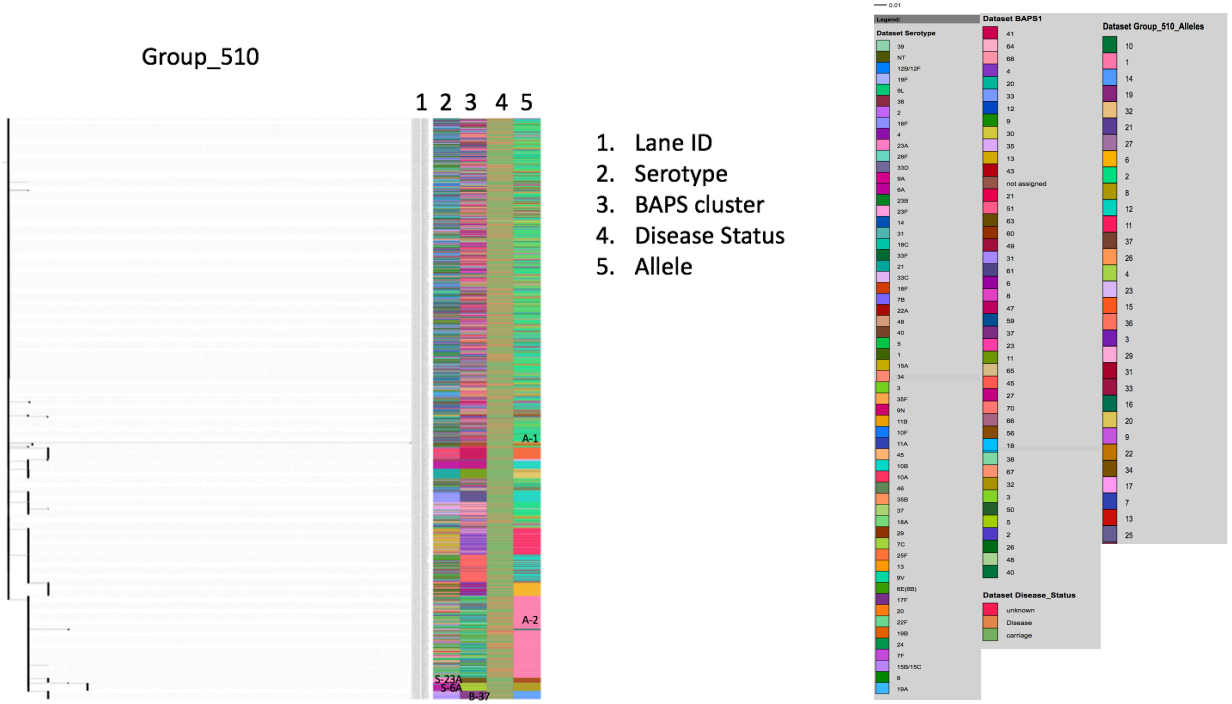
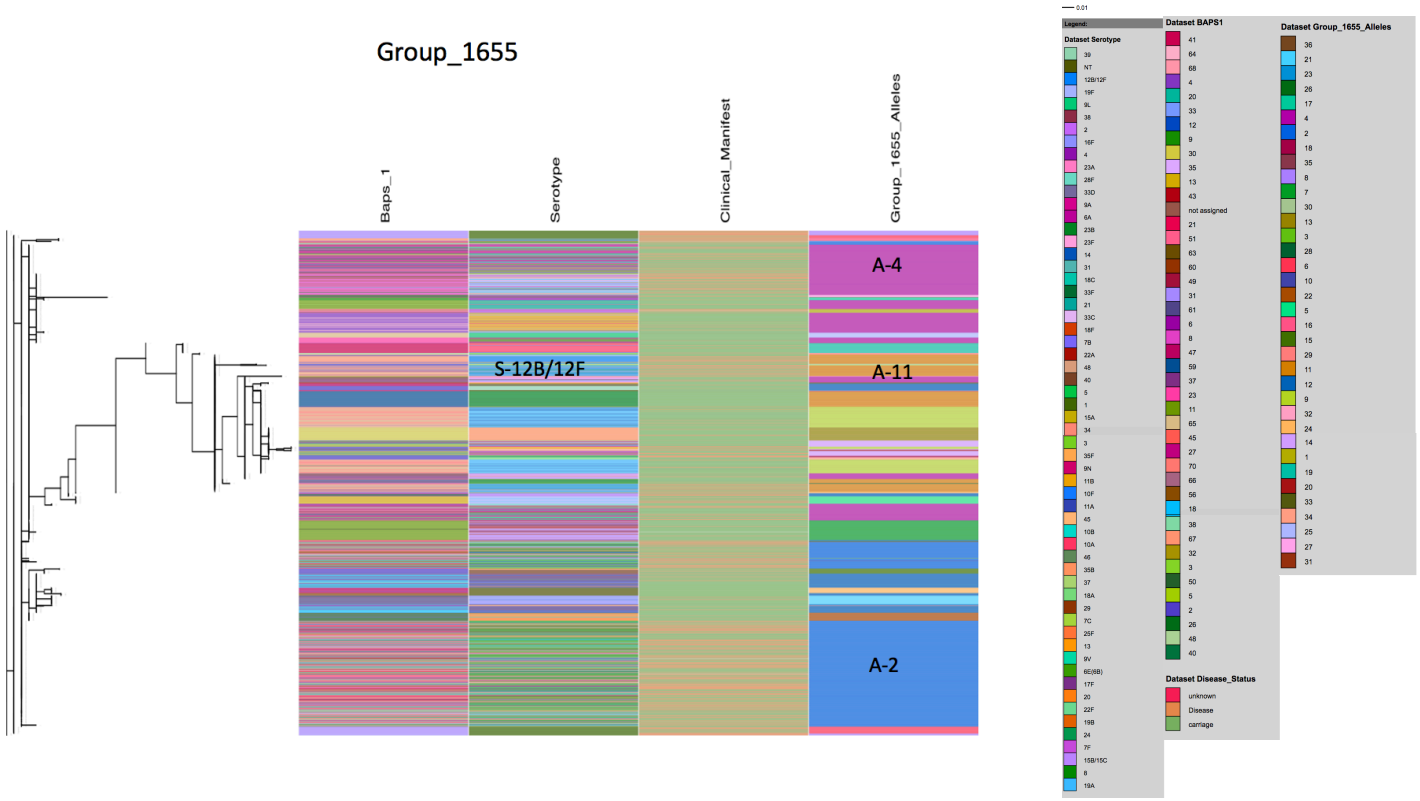


Figure 3.13 Phylogenetic gene tree of *Group_510*.

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

41 SNP sites used to reconstruct this gene tree.

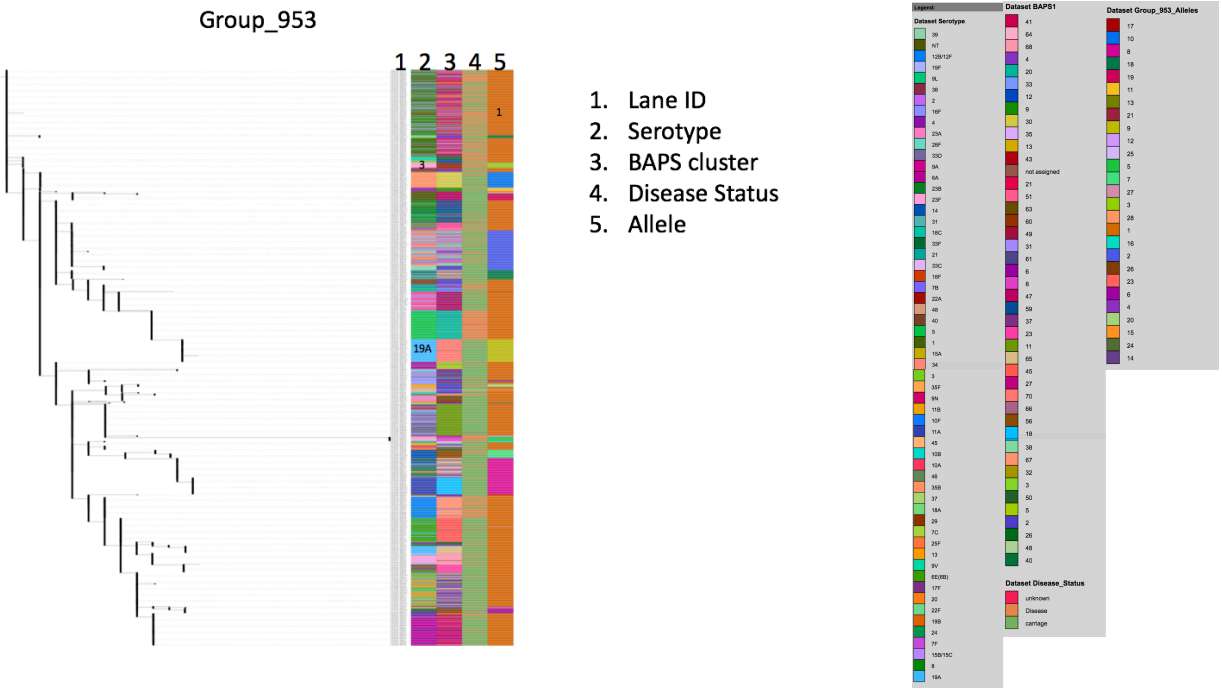
Group_1655 lipoproteins also have a short (165AA) sequence length. The most prominent alleles from the gene tree are alleles 2 and 4. Additionally, allele 11, which covers BAPS 67 of serotype 12B/12F is also important as this group includes many disease strains (Fig. 3.14).



55 SNP sites used to reconstruct this gene tree

Figure 3.14 Phylogenetic gene tree of *Group_1655*.
This gene tree shows the nucleotide relationship of the genes extracted from the genomes.

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.



74 SNP sites used to reconstruct this gene tree.

Figure 3.15 Phylogenetic gene tree of Group_953.

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

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.

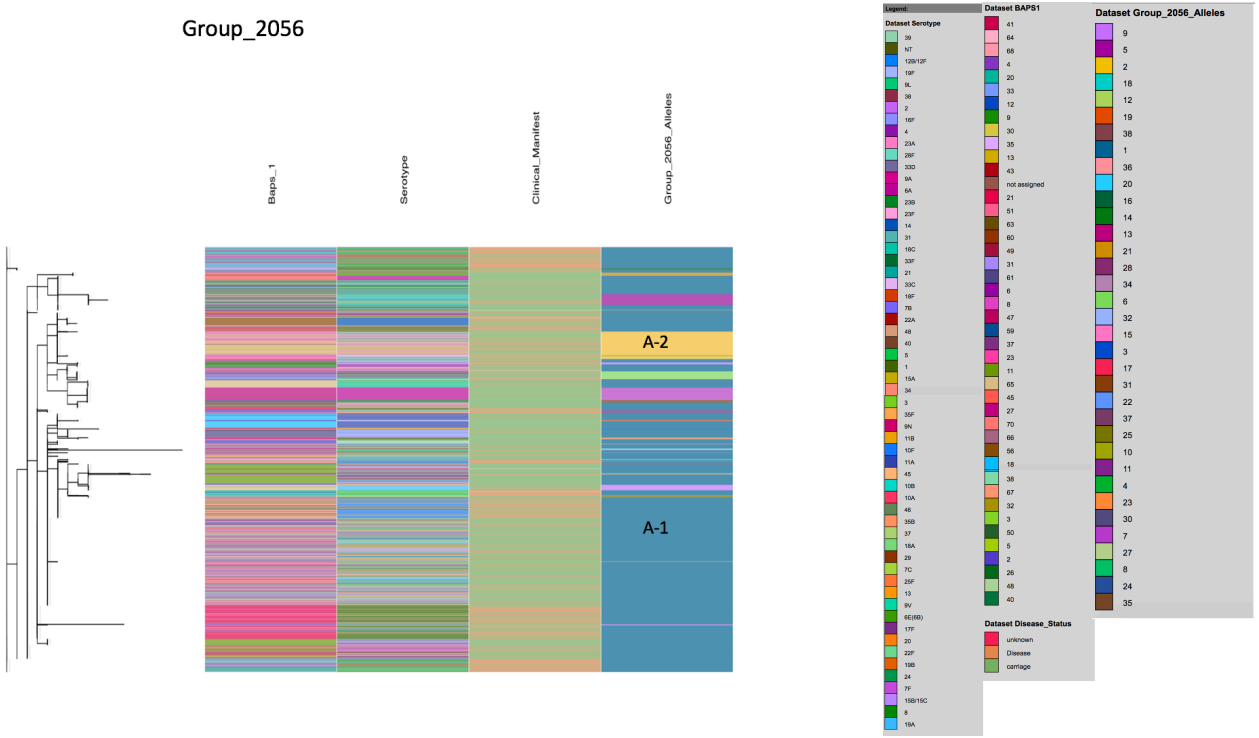


Figure 3.17 Phylogenetic gene tree of *Group_2056*.

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.

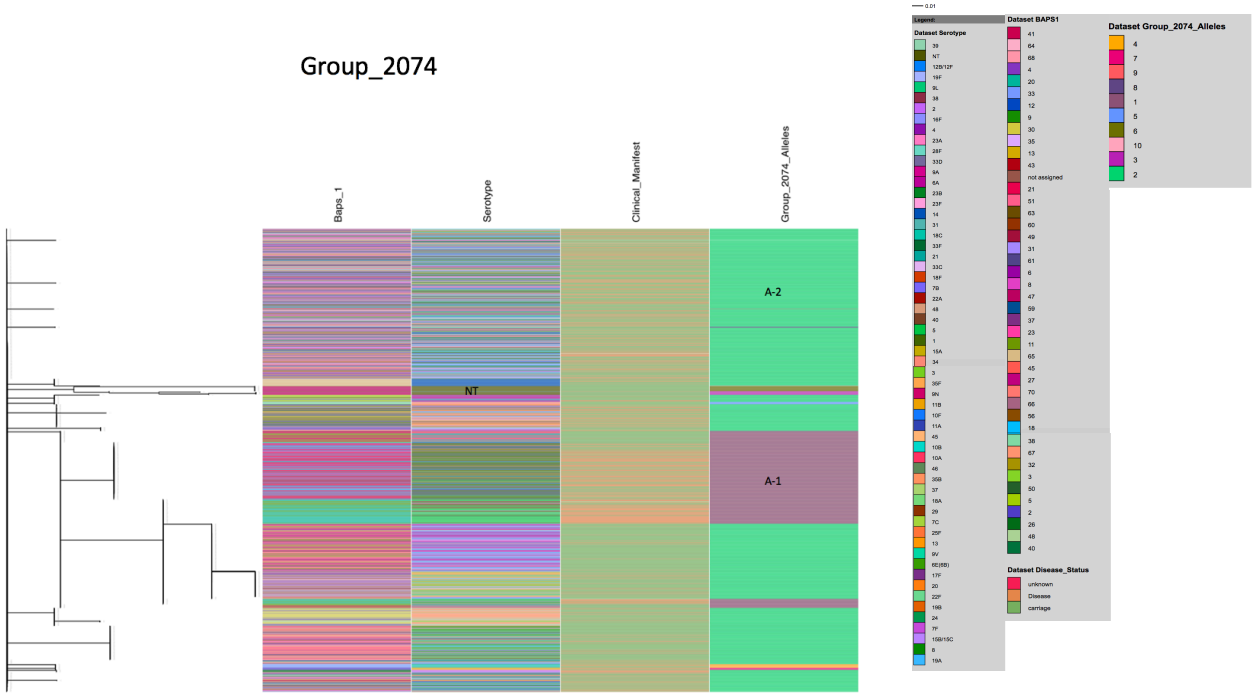


Figure 3.18 Phylogenetic gene tree of Group_2074.

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

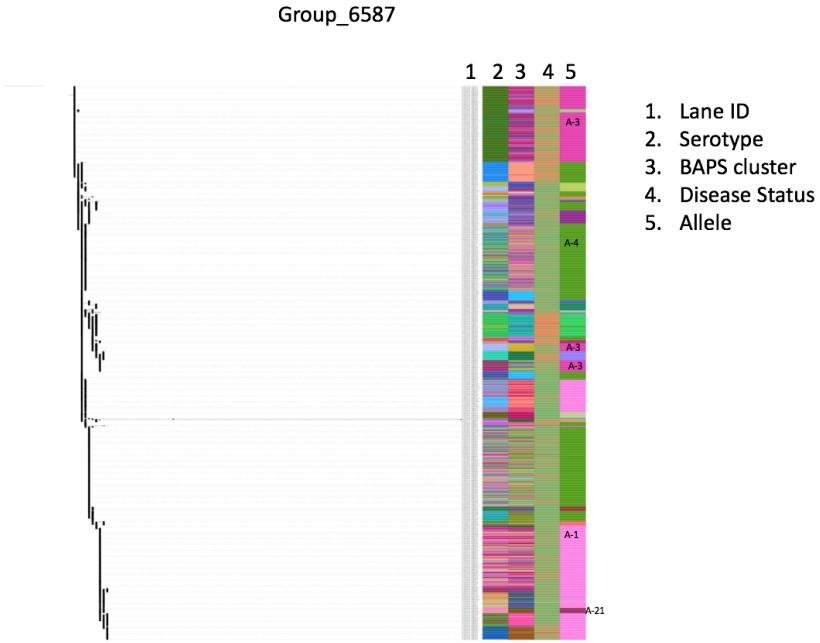
34 SNP sites used for the reconstruction of this gene tree.

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.

Figure 3.19 Phylogenetic gene tree of *Group_2298*.
This tree shows the nucleotide relationship of the genes extracted from the genomes.

32 SNP sites used for the reconstruction of this gene tree

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).



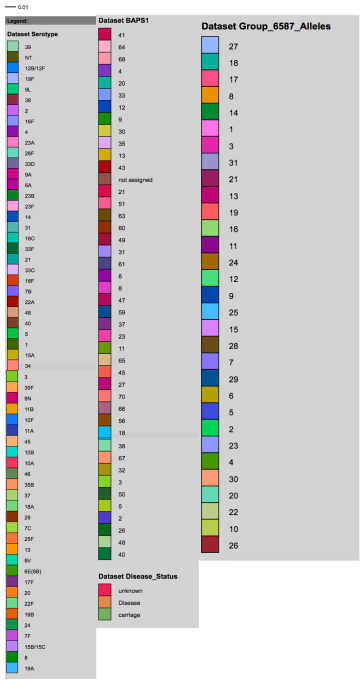


Figure 3.20 Phylogenetic gene tree of *Group_6587*.

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).

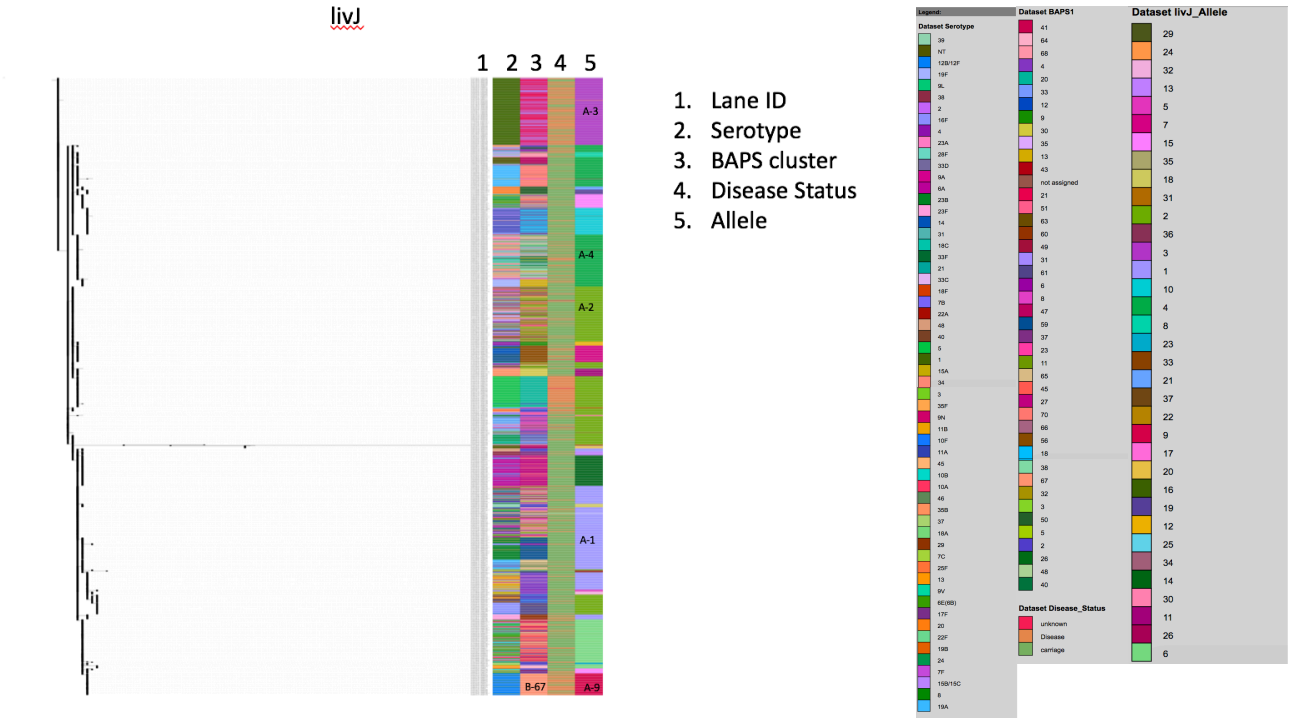


Figure 3.21 Phylogenetic gene tree of *livJ*.

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.

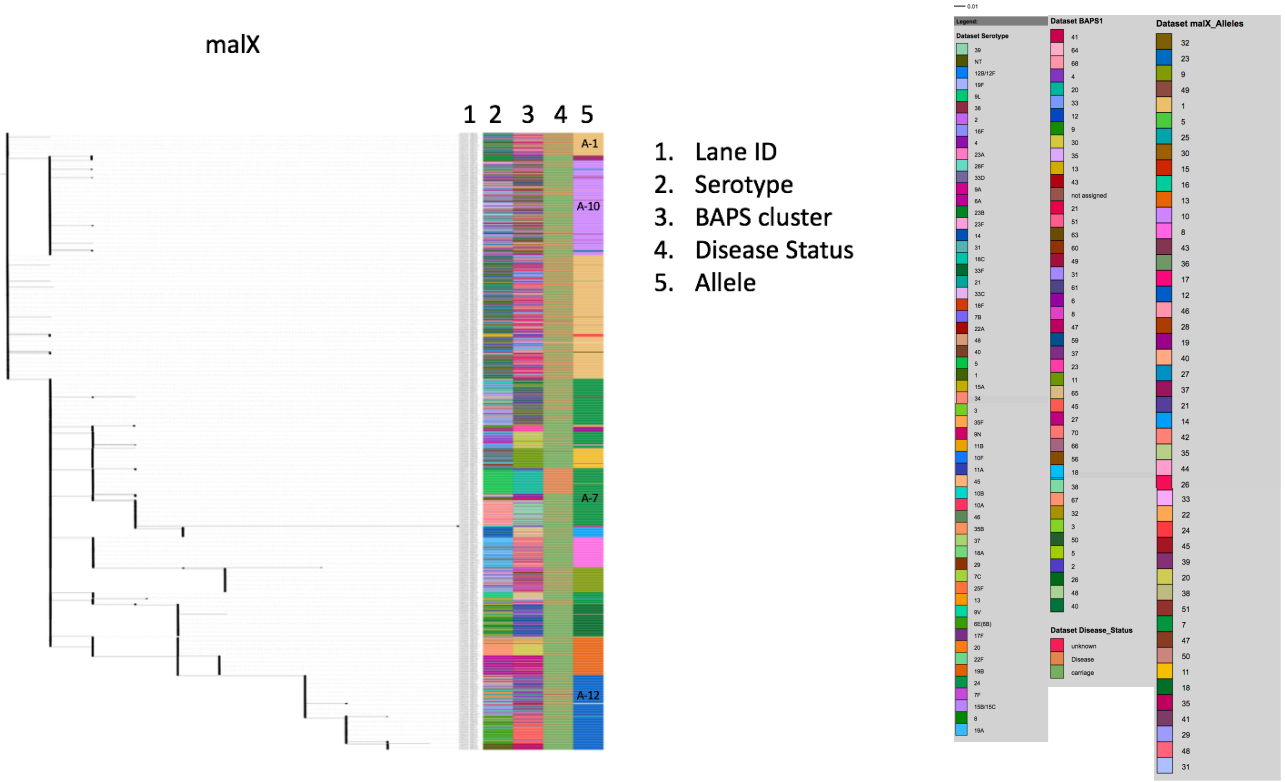


Figure 3.22 Phylogenetic gene tree of *malX*.

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.

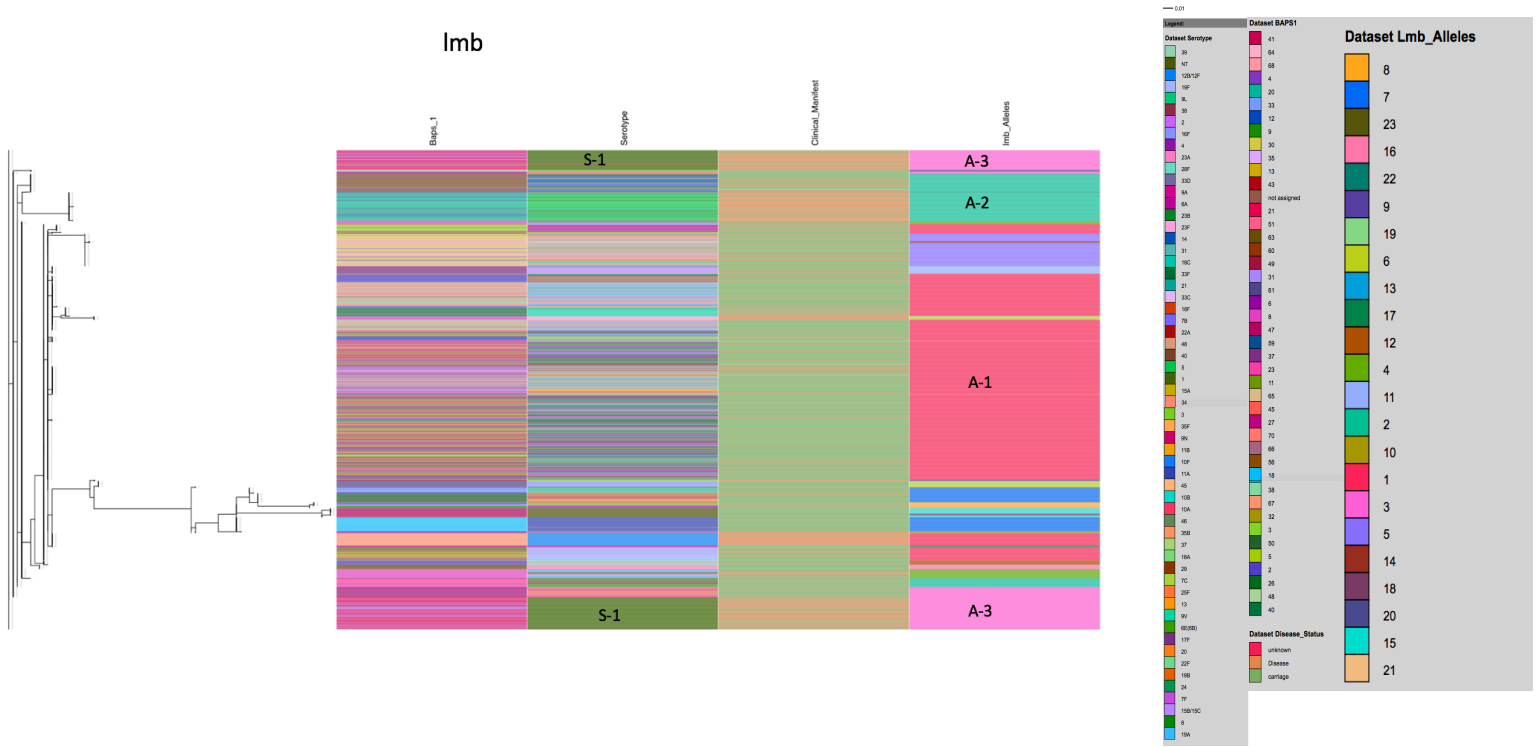
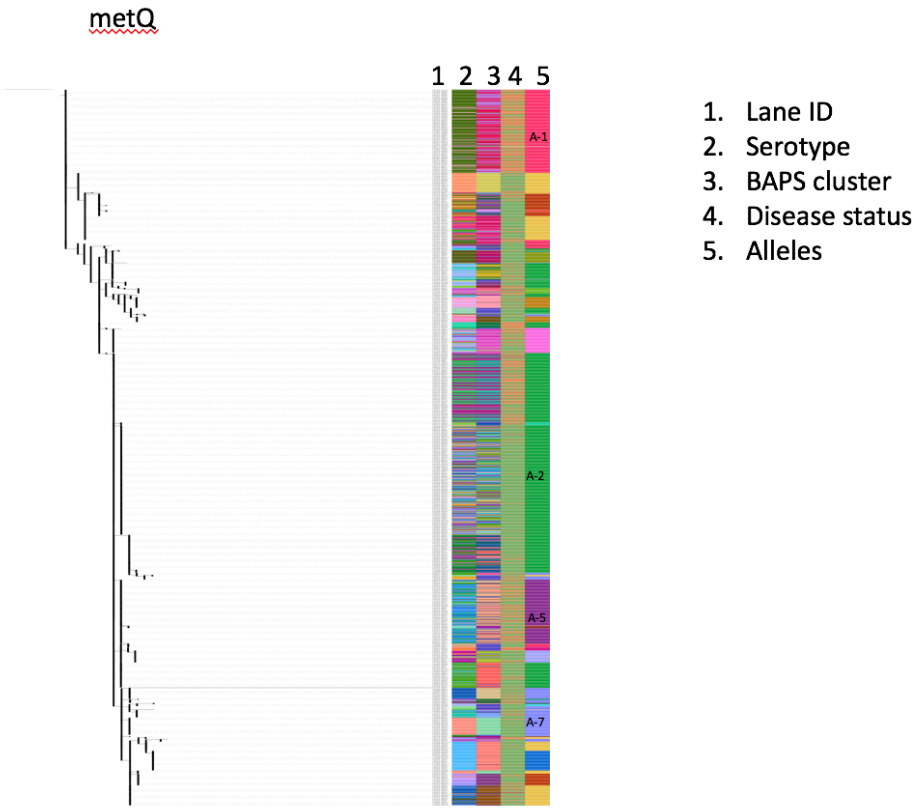


Figure 3.23 Phylogenetic gene tree of *lmb*.

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.



1. Lane ID
2. Serotype
3. BAPS cluster
4. Disease status
5. Alleles

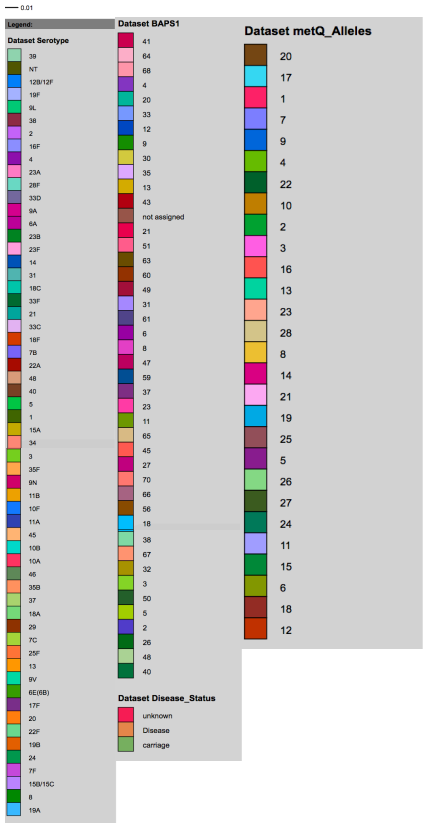


Figure 3.24 Phylogenetic gene tree of *metQ*.

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.

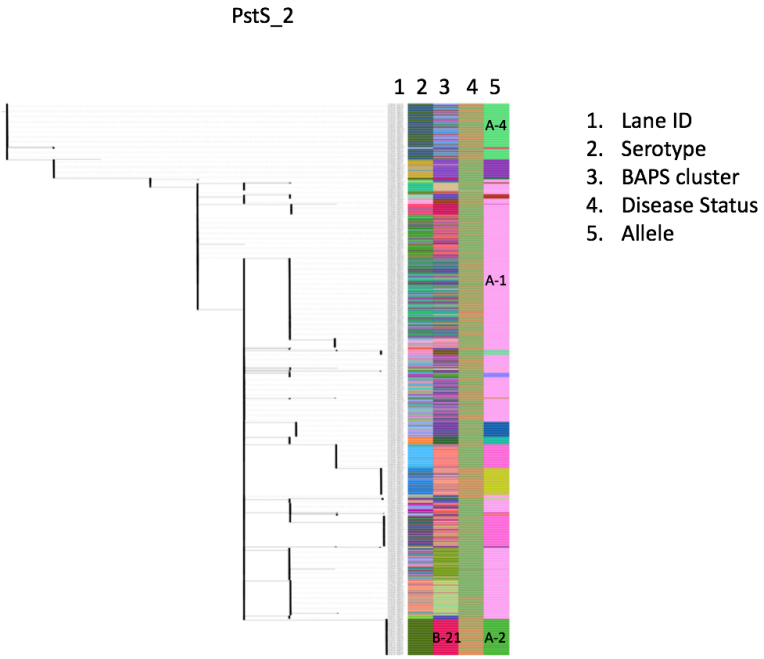
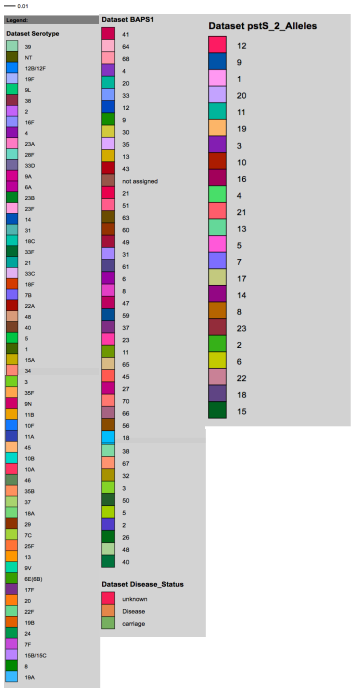


Figure 3.25 Phylogenetic gene tree of *pstS_2*.
This tree shows the nucleotide relationship of the genes extracted from the genomes.



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.

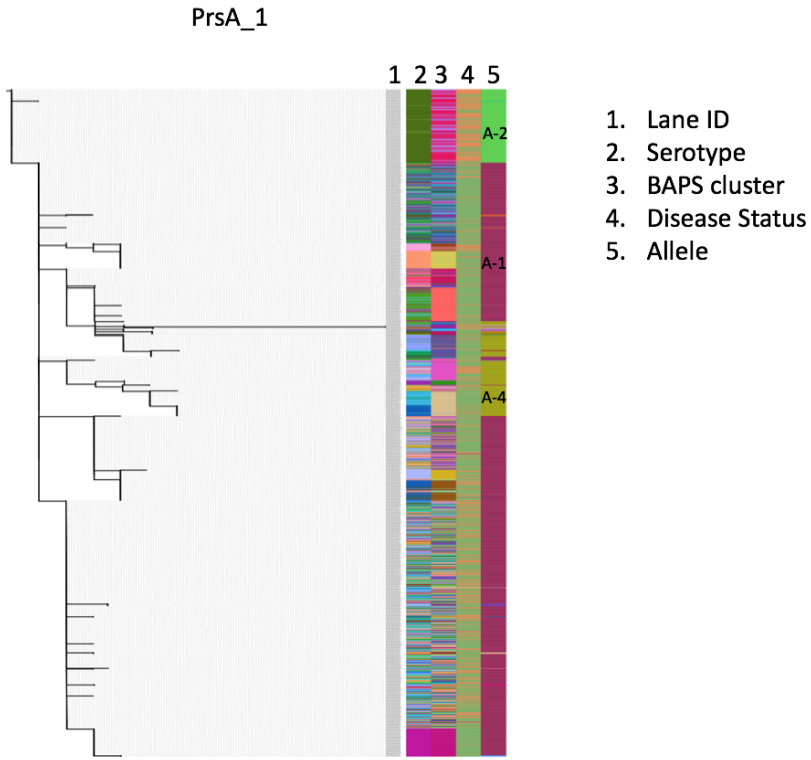
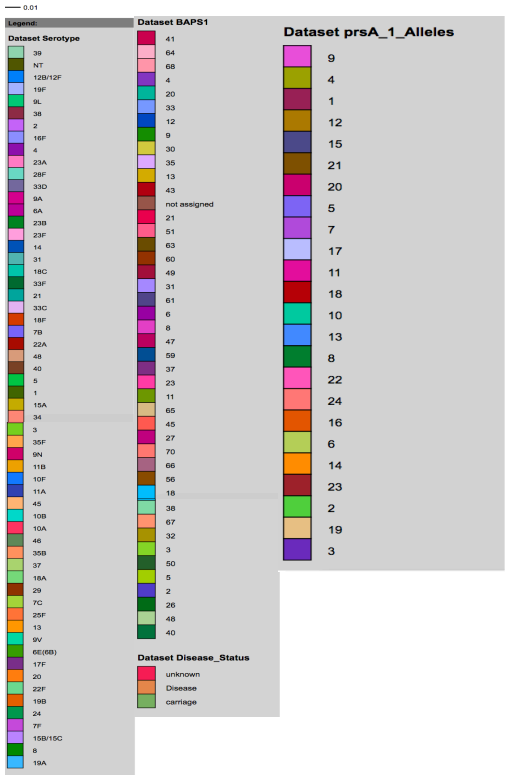


Figure 3.26 Phylogenetic gene tree *prsA*.

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



51 SNP sites used to reconstruct this gene tree

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.

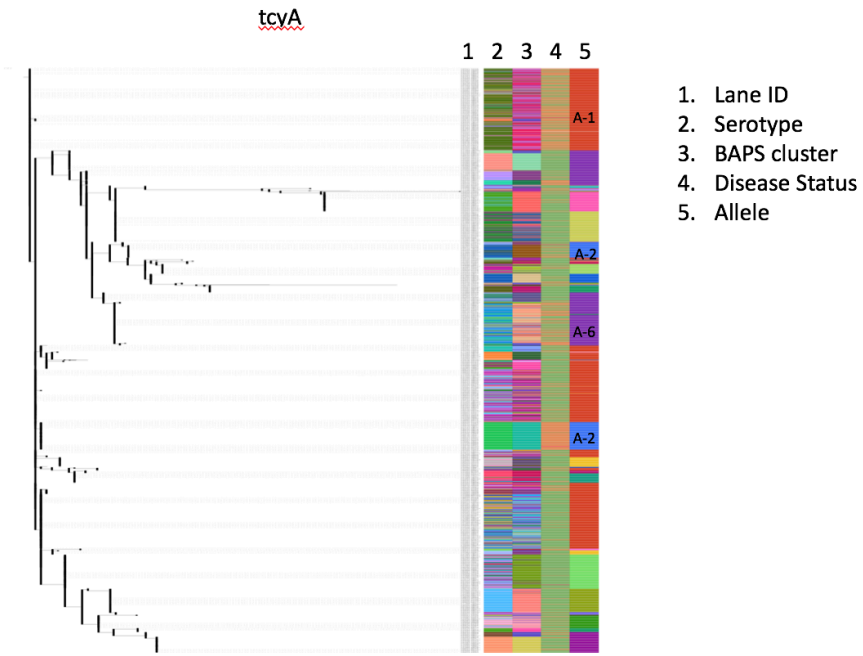
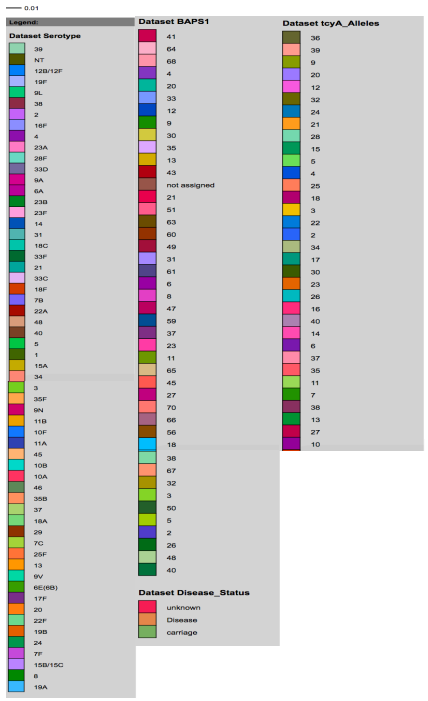


Figure 3.27 Phylogenetic gene tree of *tcyA*.

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.

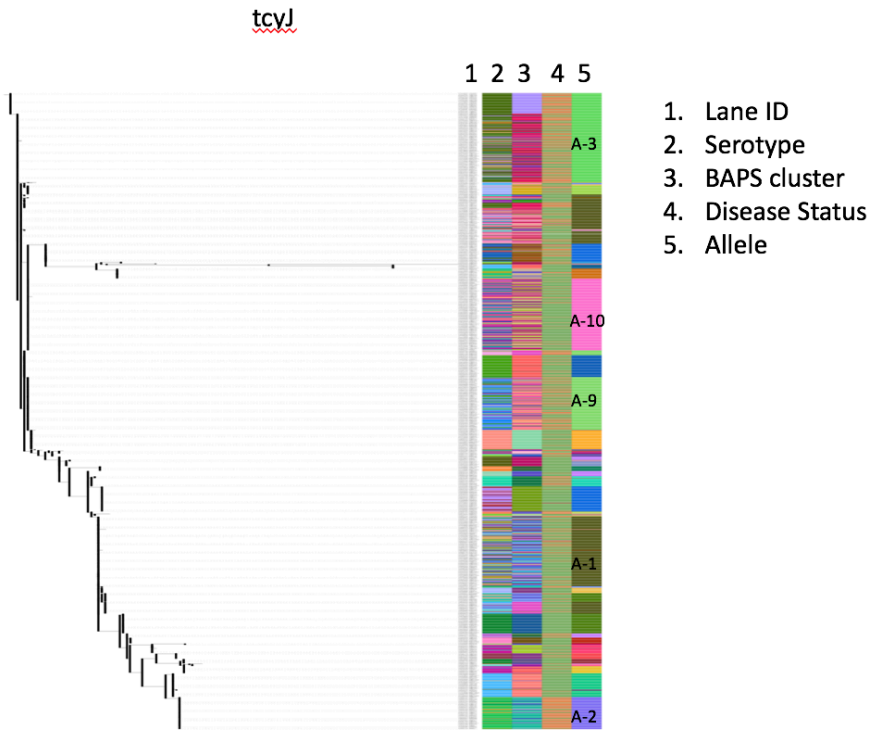
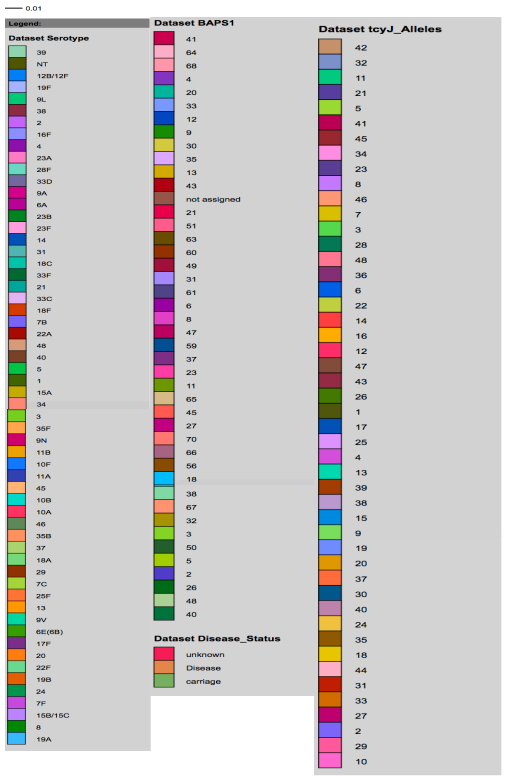


Figure 3.28 Phylogenetic gene tree of *tcyJ*.

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).

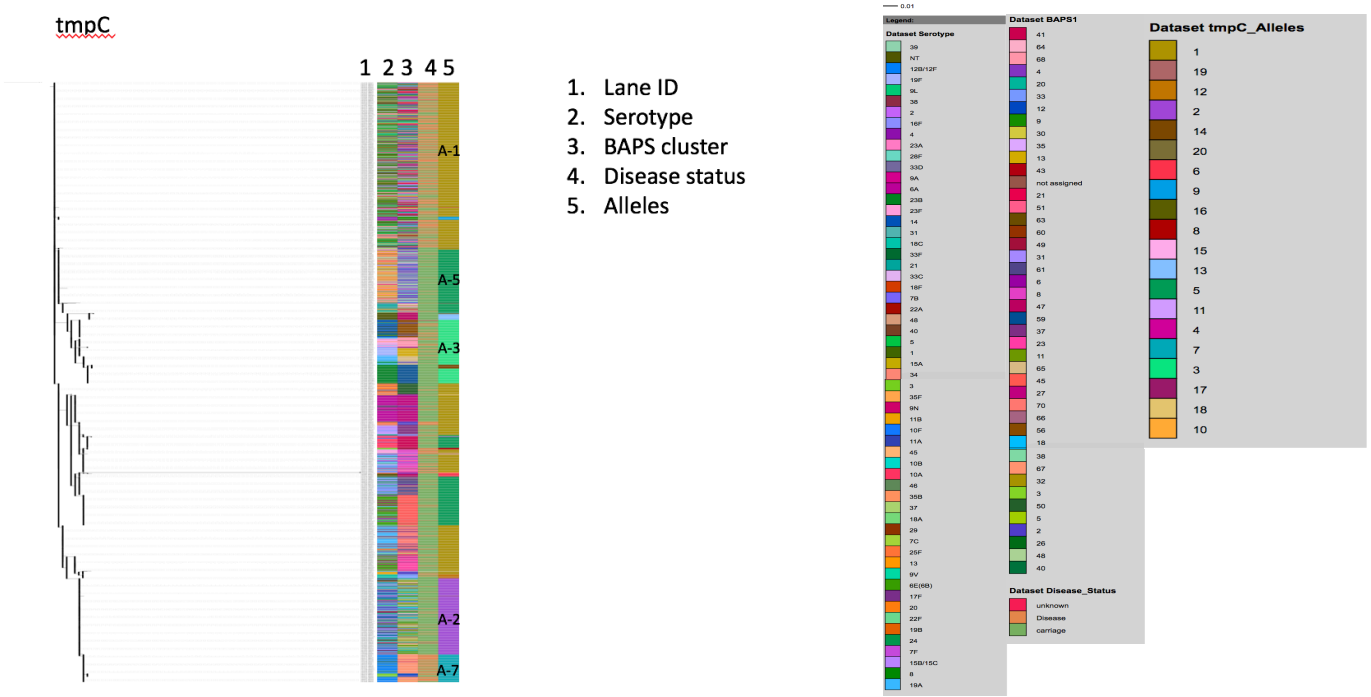
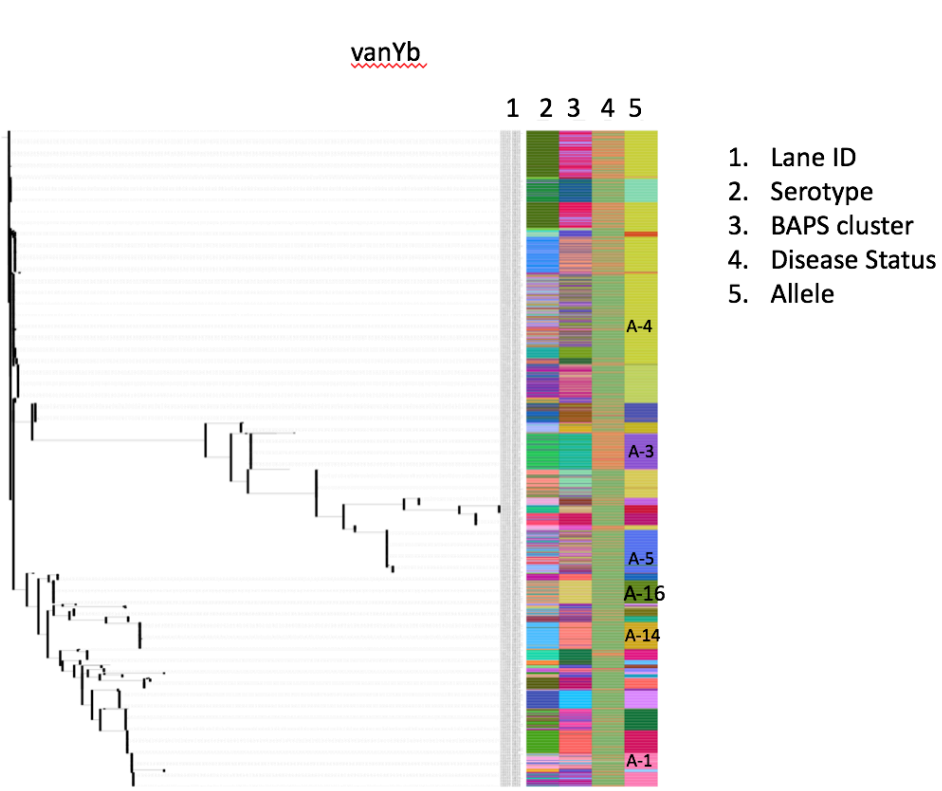


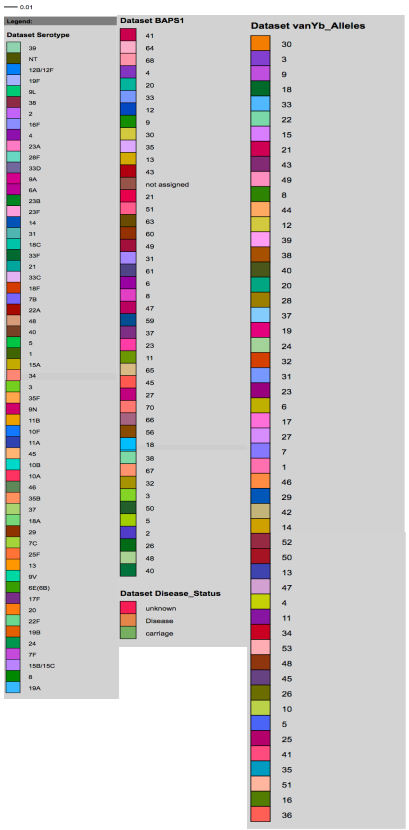
Figure 3.29 Phylogenetic gene tree of *tmpC*.

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.





256 SNP sites used to reconstruct this gene tree.

Figure 3.30 Phylogenetic gene tree of *vanYb*.

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

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.

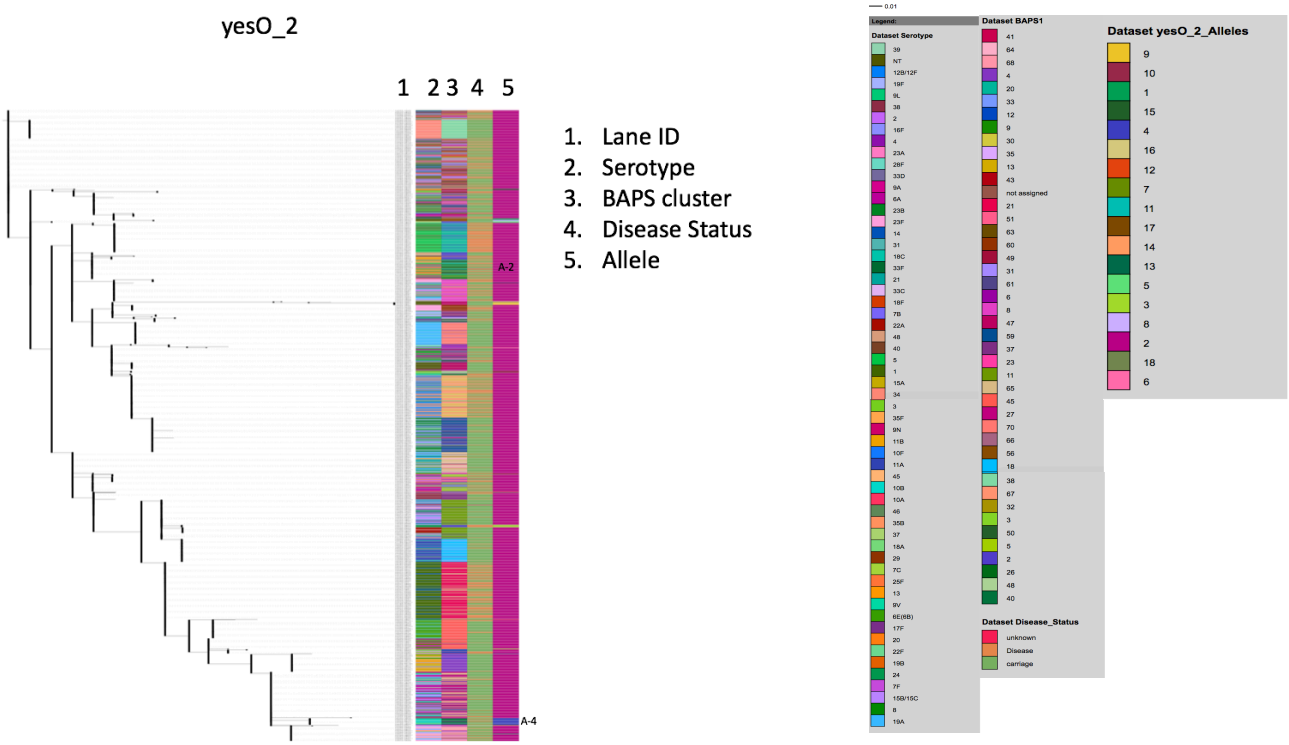
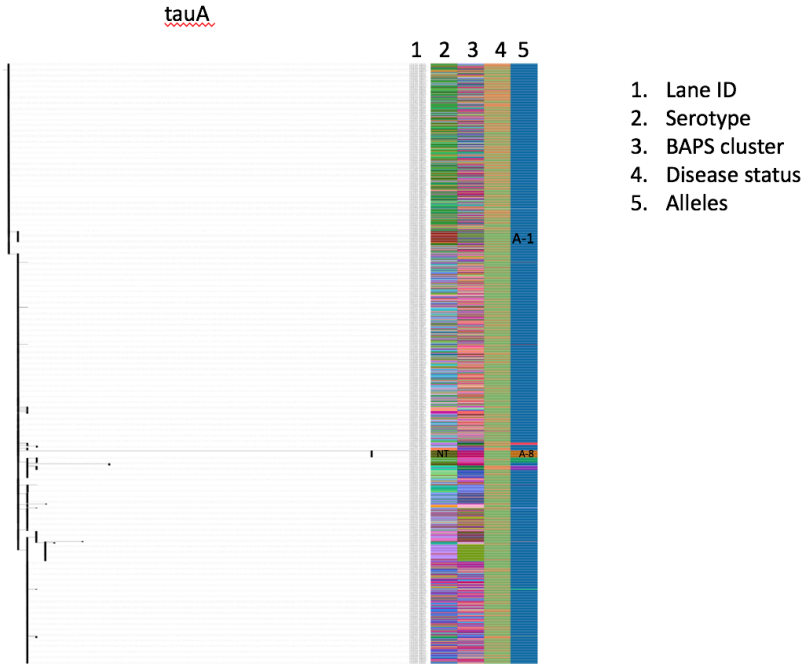


Figure 3.31 Phylogenetic gene tree of *yesO_2*.

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.



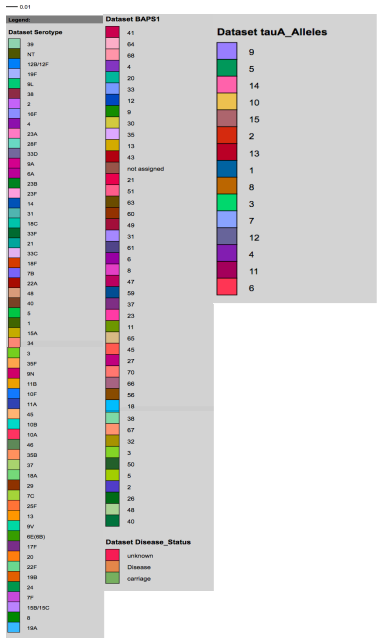


Figure 3.32 Phylogenetic gene tree of *tauA*.

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).

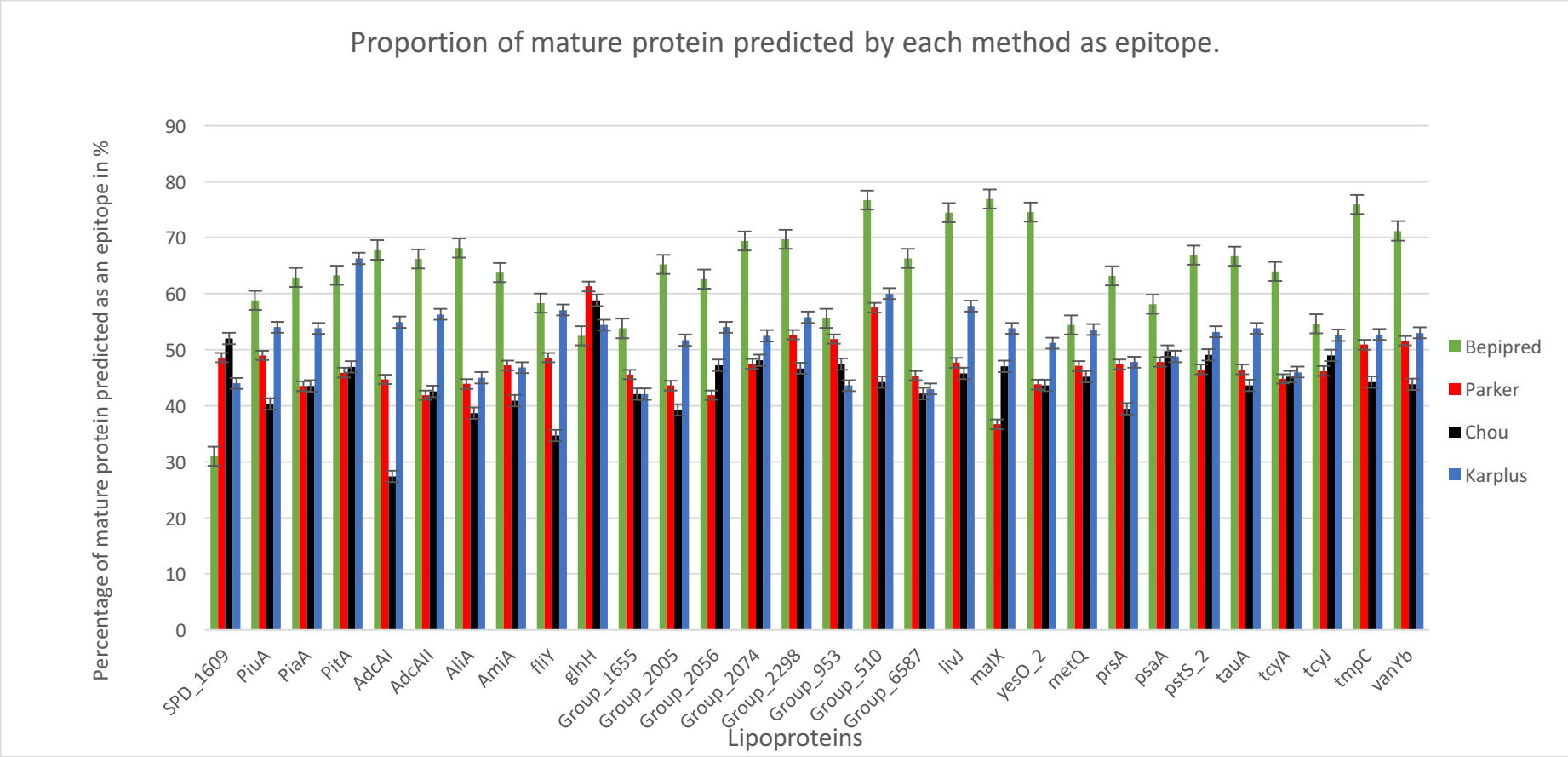


Figure 3.33 Percentage of mature protein predicted as an Epitope

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.

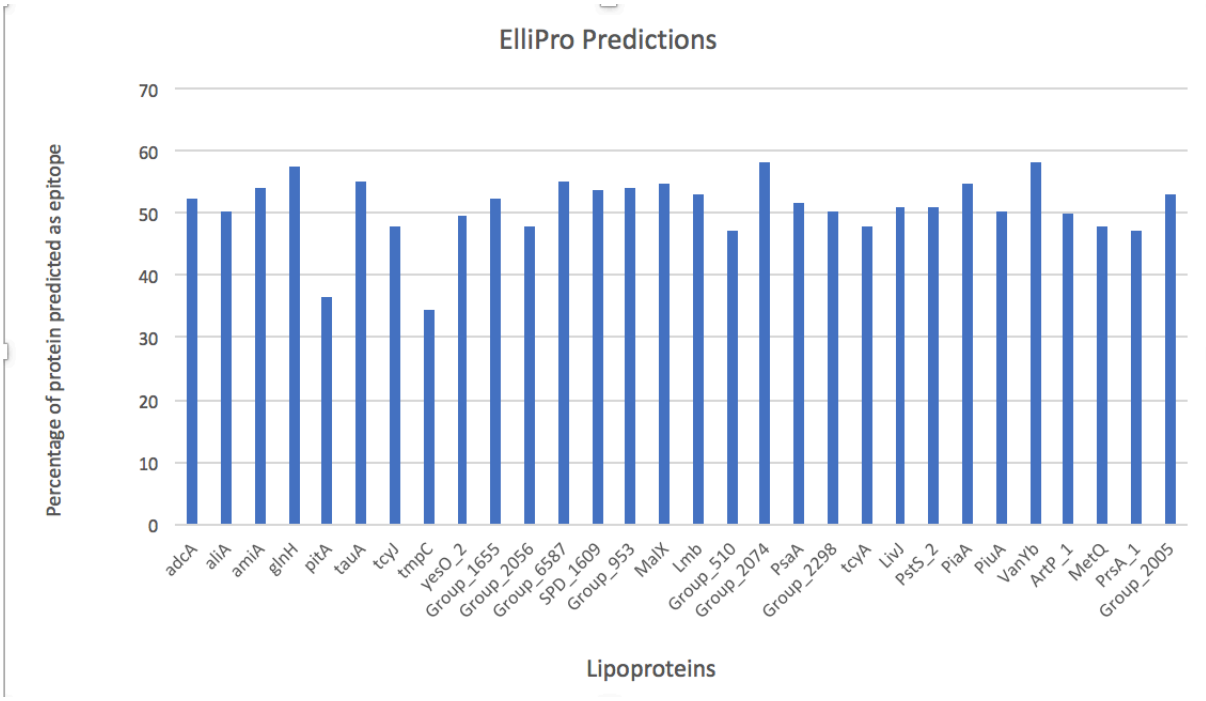


Figure 3.34 Percentage of proteins predicted as epitope by ElliPro
The horizontal axis is labelled with the protein names and the vertical axis has the percentages. The height of the bars for each protein represents the percentage of the protein that is predicted by ElliPro to be a part of an Epitope.

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 epitope (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).

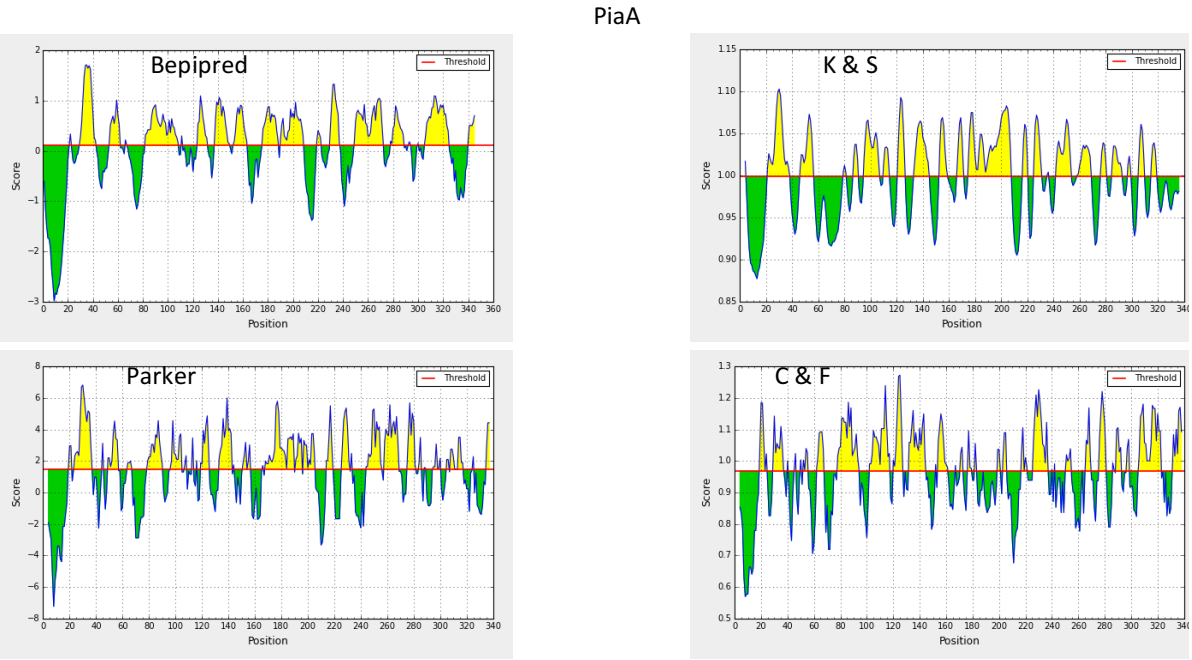


Figure 3.35 Linear epitope predictions of the PiaA protein.

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

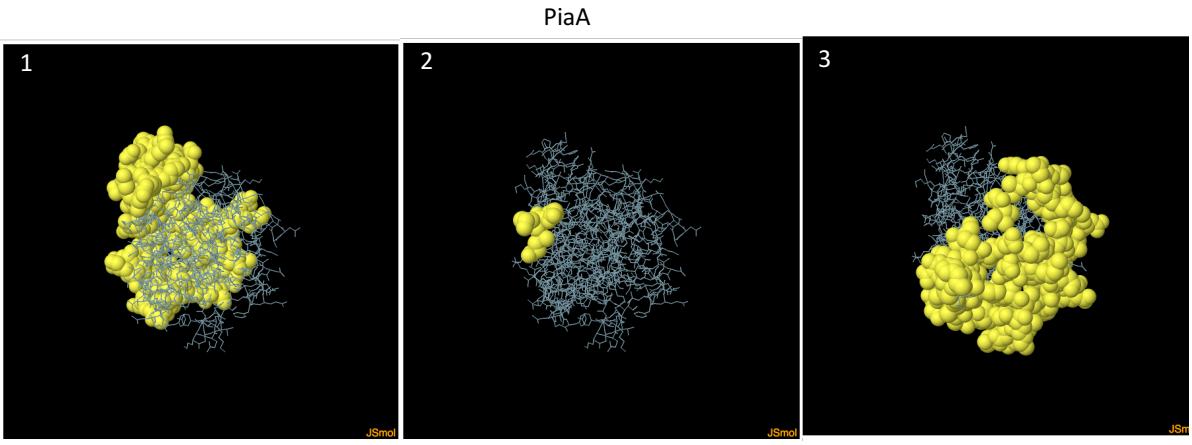


Figure 3.36 ElliPro predicted discontinuous epitopes for PiaA.

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.

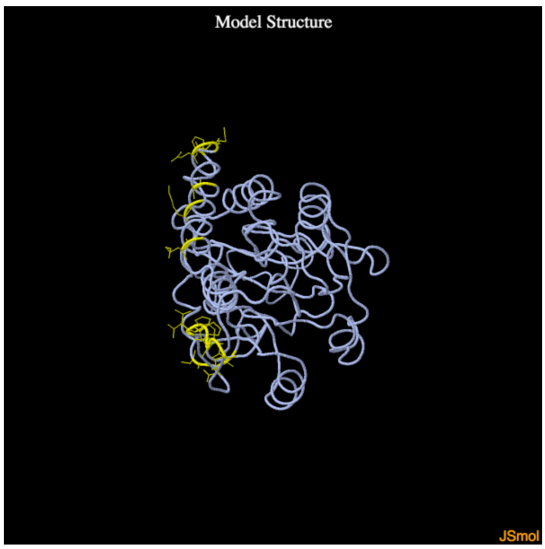


Figure 3.37 DiscoTope2 predicted discontinuous epitopes for PiaA.
The parts coloured yellow are the predicted epitopes.

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.

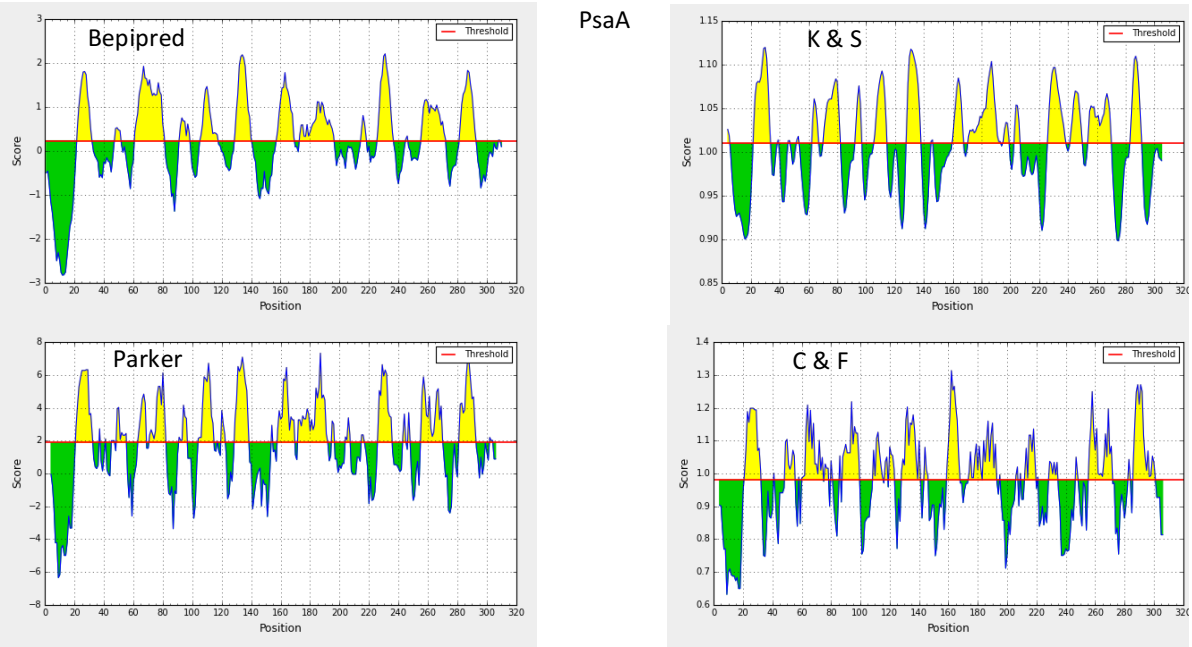


Figure 3.38 Linear epitope predictions of the PsaA protein.
This was predicted using Bepipred, Parker, Karplus and Schulz (K&S) and Chou and Fasman (C&F) methods.

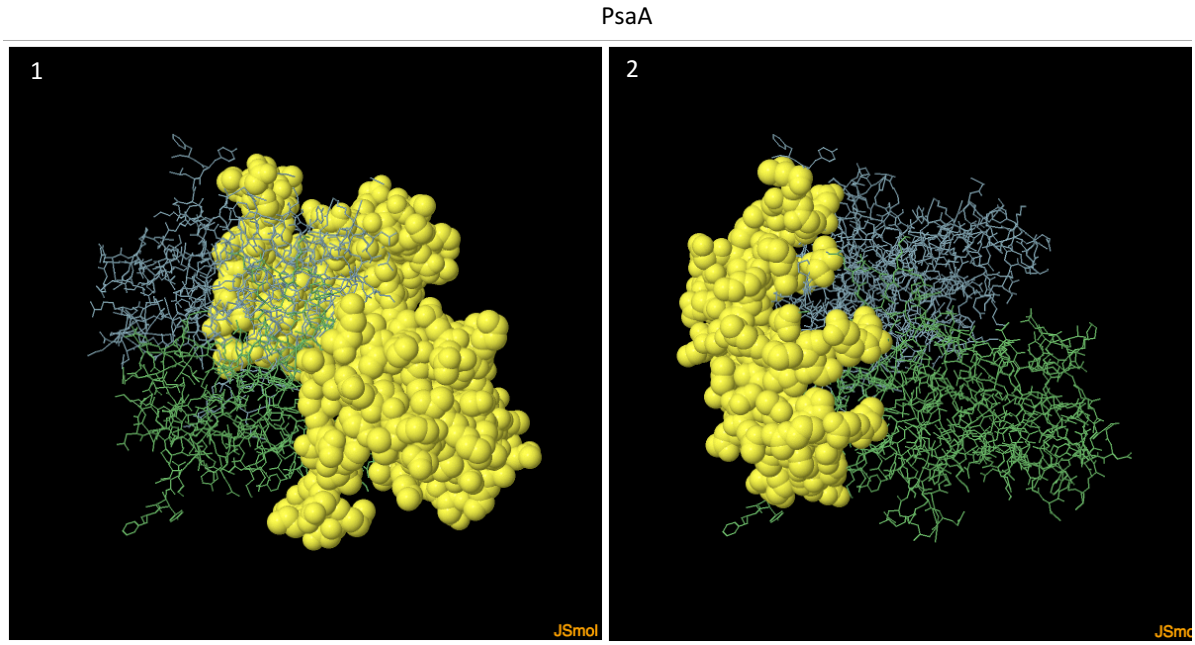


Figure 3.39 ElliPro predicted discontinuous epitopes for PsaA.

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.

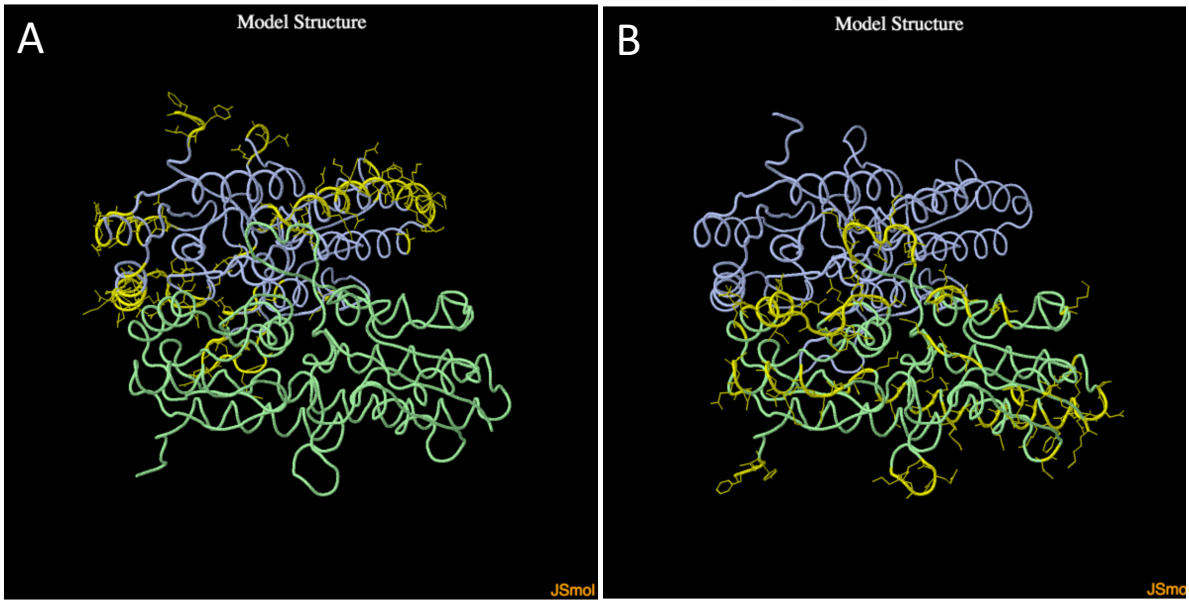


Figure 3.40 DiscoTope2 predicted discontinuous epitopes for PsaA.

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).

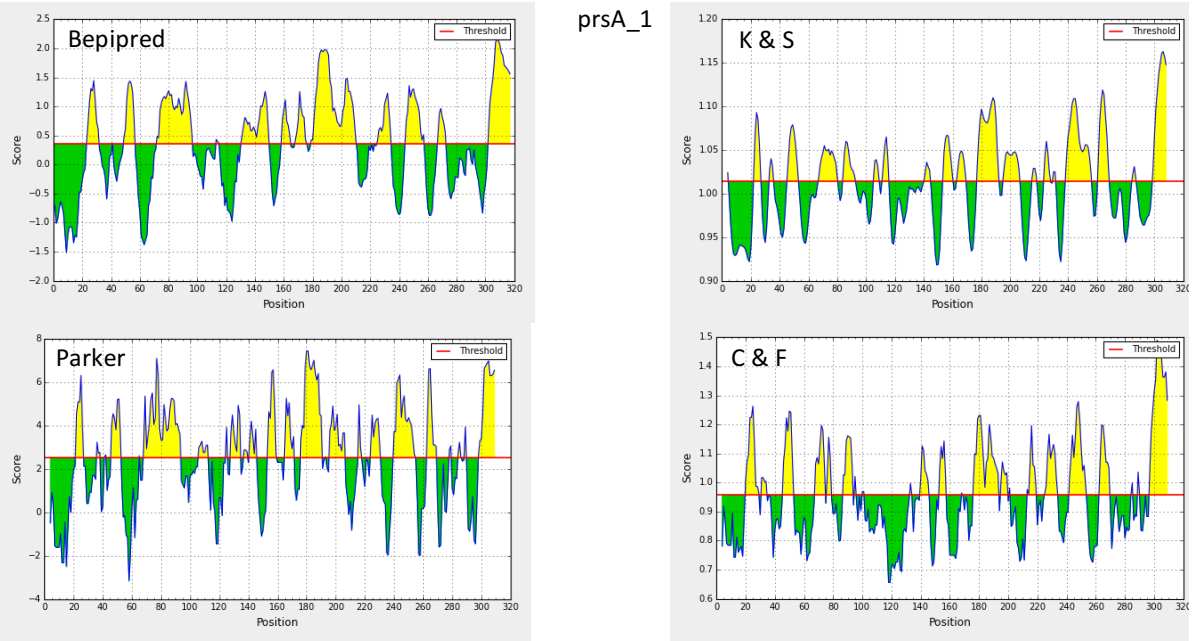


Figure 3.41 Linear epitope predictions of the PrsA_1 protein.

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

PrsA_1

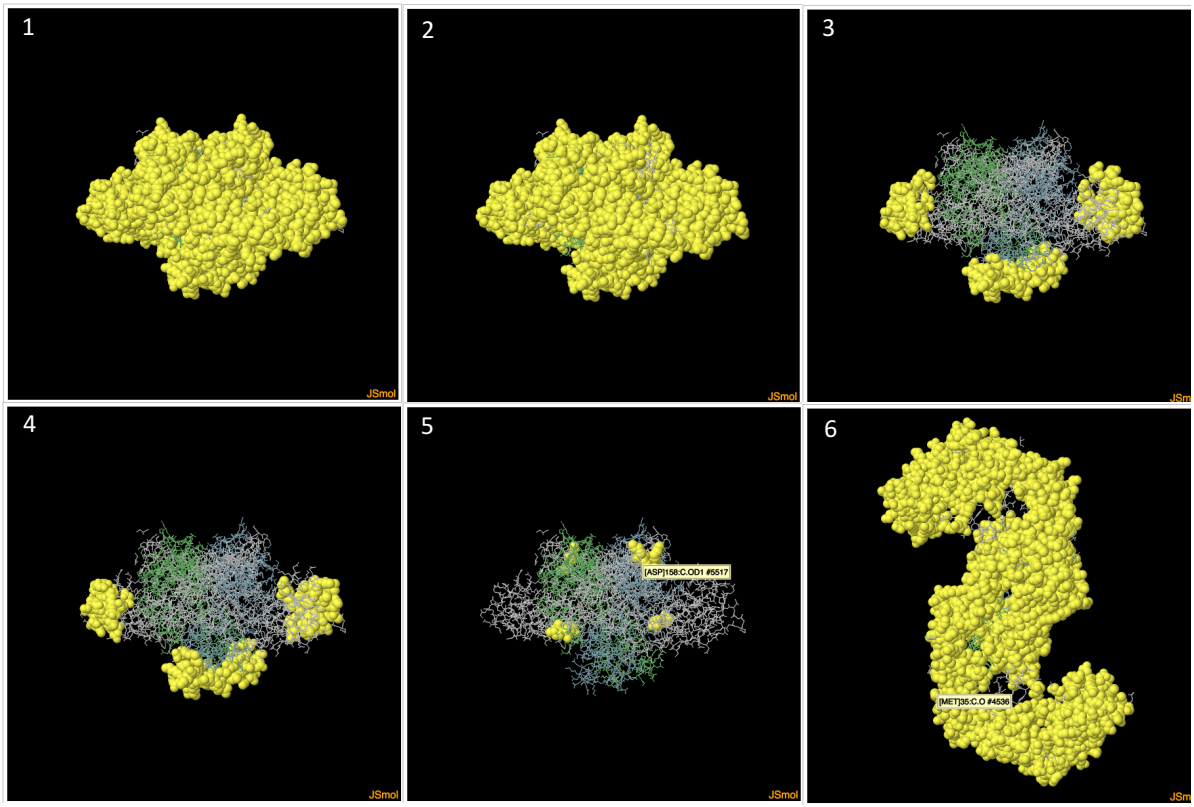


Figure 3.42 ElliPro predicted discontinuous epitopes for PrsA_1.

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.

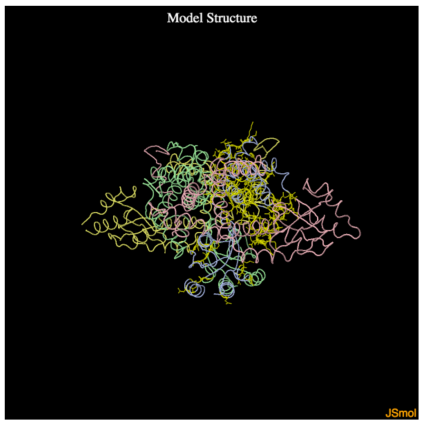


Figure 3.43 DiscoTope2 predicted discontinuous epitopes for prsA_1.
The parts coloured yellow are the predicted epitopes.

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%/