Chapter 3: The global and Maela pneumococcal population structure

---

Declaration of work contributions:

MLST data for the Kilifi, Kenya and the Gambia was kindly provided by Dr Angela Brueggemann and Dr Sarah Burr respectively. Beth Sutton helped prepare MLST data for Southampton UK, the Gambia and Massachusetts USA. Bayesian clustering analysis was performed by Professor Jukka Corander. Variations in capsule biosynthesis loci of nontypable pneumococci was analysed by Susannah Salter. Parsimonious reconstruction of serotype switching events based on the phylogeny was conducted by Dr Simon Harris. Unless stated here, I was responsible for the analyses

3. The Maela and global pneumococcal population structure

3.1 Introduction and aims

Streptococcus pneumoniae can be detected worldwide, with a higher rate of carriage observed in resource-poor settings. Among these are refugees, a third of whom live in crowded camps, and are potentially at higher risk of developing respiratory infections (Bellos, Mulholland *et al.* 2010). A longitudinal carriage study was conducted in Maela, a refugee community located 5 km east of the Thailand-Myanmar border, to capture the frequency of pneumococcal carriage in this refugee community between 2007-2010 (Turner, Turner *et al.* 2012 & 2013). Whole-genome sequencing was performed on a random subset of isolates from this carriage study, generating 3,085 pneumococcal genomes. This allowed one to study the population structure of pneumococci circulating in this rural community in South East Asia, a region in which much less was known about genotype distribution prior to this study.

To build a comparative view, pneumococcal populations in contemporaneous carriage studies conducted in various locations including Kilifi, Kenya (Brueggemann, Muroki *et al.* 2013); Massachusetts, USA (Croucher, Finkelstein *et al.* 2013); Southampton, UK (Tocheva, Jefferies *et al.* 2013) were compared to Maela using multi-locus sequence typing (MLST) data. Variations in pneumococcal structure observed here might influence the success of clinical interventions including vaccines and choices of antibiotic treatments between countries.

A comparative view between countries also provides information on common sets of clones frequently detected in multiple locations. Most of these globally spread clones have been recognised and designated by the Pneumococcal Molecular Epidemiology Network (PMEN). Based on one of these clones, this chapter also captures multiple introductions of the globally circulating PMEN14 into one local community, using Maela refugee camp from Thailand as an example.

Together, the information on the population structure: both in breadth through MLST data of some global collections; and in depth through whole genome sequencing from

a local community like Maela refugee camp, might allow speculation on the differences or similarities in the population responses to clinical interventions exerted on each community.

This chapter is aimed at:
i) Determining the pneumococcal population structure in Maela, which is the main subject of this thesis, using data generated by whole genome sequences.

ii) Comparing the population structure observed in Maela to populations from other countries including UK, USA, Kenya and Gambia using data generated from multilocus sequence typing (MLST).

iii) Understanding the spread of one particular global clone (PMEN14) and its introduction into the Maela community.

3.2 Methods

3.2.1 Estimating the Maela population structure

Based on whole genome sequencing data, the population structure was estimated using both a minimum evolution tree (Price, Dehal *et al.* 2010) and Bayesian analysis with BAPS (Corander, Marttinen *et al.* 2008)

3.2.1.1 Minimum evolution tree

Based on the coarse mapping against the core genome of *S. pneumoniae* ATCC 700669 (see 2.5), an approximately-maximum likelihood phylogenetic tree was estimated by FastTree (Price, Dehal *et al.* 2010) using GTR + CAT (General Time Reversible with per-site rate CATegories) model of approximation for site rate variation. With 1,000 resamples, 80.6% and 32.6 % of the branches have over 0.700 and 1.000 bootstrap support respectively.

3.2.1.2 Bayesian Analysis of Population Structure (BAPS)

Population clustering was performed by Professor Jukka Corander. BAPS software v6.0 (Corander, Waldmann *et al.* 2003, Corander and Tang 2007, Corander, Marttinen *et al.* 2008, Tang, Hanage *et al.* 2009) was used to estimate the population structure based on the mapping against the core genome of *S. pneumoniae* ATCC700669 (see 2.5). The software searched for sufficiently similar nucleotide frequencies within each segment of the whole genome alignment and linked a particular group of bacterial isolates together on non-reversible stochastic optimisation. Due to the relatively large dataset, BAPS was performed in a hierarchical manner to resolve the population structure to a fine level of detail. First, the module for clustering individual strains was applied to obtain the posterior mode partition into primary clusters based on 5 runs of the estimation algorithm. Data from each of these primary clusters were then analysed again with BAPS in an identical manner to obtain secondary clustering within each primary cluster, thereby forming hierarchical clusters. The hierarchical approach was adopted when there are large genetic differences between the major lineages that may mask more subtle signals of divergence present within a lineage

(Willems, Top *et al.* 2012, Cheng, Connor *et al.* 2013). In this dataset, 33 primary and 183 secondary clusters were determined (**Appendix A**).

3.2.2 Estimating global pneumococcal population structure

Global population structure was estimated based on sequence type information. For the Maela population, the sequences of seven loci used for sequence typing were extracted from the genomes using ICORN (Otto, Sanders *et al.* 2010) to transfer the Illumina read mapping to the reference. The sequences were then analysed using www.pubmlst.org/spneumoniae/ and tabulated in **Appendix A**. Sequence types from other populations were extracted from the published data including Kilifi, Kenya (Brueggemann, Muroki *et al.* 2013); Massachusetts, USA (Croucher, Finkelstein *et al.* 2013); Southampton, UK (Tocheva, Jefferies *et al.* 2013). Together these were clustered into clonal complexes (CCs) using Phyloviz (Francisco, Vaz *et al.* 2012) with the following settings: Dataset type = Multi-locus Sequence Typing; Distance = eBURST Distance; and Level = SLV.

3.2.3 Serotype switches

States of changes in serotype were counted based on parsimony reconstruction of serotypes onto the phylogenetic tree constructed above. Changes in serotypes were reported as potential switches.

3.3 Results

3.3.1 Maela pneumococcal population structure

To explore the Maela population structure and diversity at a high resolution, whole genome sequencing was performed on 3,085 isolates collected from infants and mothers between 2007-2010. This section describes the Maela pneumococcal population structure captured by different methods, its dominant lineages and serotype switch events. The latter potentially play a role in population dynamics.

3.3.1.1 Determining closely related isolates

To estimate the population structure, reads from all 3,085 Maela pneumococcal samples were mapped onto a single core reference genome, *S. pneumoniae* ATCC700669 (Croucher, Walker *et al.* 2009) to generate a coarse alignment with an average of 82.3 % mapping coverage, which was sufficient for determining the overall structure. The coarse population structure was determined by two independent approaches – a Bayesian clustering (BAPS) and minimum evolution tree (FastTree) (see methods). Both methods provide rapid identification of population structure and are capable of handling large datasets. The BAPS approach partitions genomes into non-overlapping segments, each with a conditional probability of ancestry over the range of putative alternatives, given the heterogeneity of the genome. BAPS searches for sufficiently similar nucleotide frequencies within each segment and links a particular group of isolates together based on sequence similarity (Corander, Marttinen *et al.* 2008). Potential recombination in the population was incorporated into the model and separately treated as genetic admixture. The FastTree approach first uses a heuristic variant of neighbour joining to estimate the approximate topology, followed by an improvement of topology through different algorithms (Price, Dehal *et al.* 2009). Because of the size of the large-scale analysis, recombination in the population was not considered in the approximate phylogeny.

Both methods were first trialled on a smaller species-wide pneumococcal dataset of 127 isolates from Malawi (Everett, Cornick *et al.* 2012). The smaller sample size allowed the methods to be tested rapidly before their application on the larger Maela

pneumococcal sample set. Applications of the minimum evolution tree and Bayesian clustering yielded similar results on Malawian data. 87.5-100% of isolates in each BAPS cluster are found together on the same branch of the tree, suggesting that both methods seem to be robust and consequently both were used for analysing the Maela population structure.

3.3.1.1.1 Bayesian clustering

BAPS was applied to the whole genome alignment discussed above to investigate the population structure as described in (Corander, Marttinen *et al.* 2008, Tang, Hanage *et al.* 2009) except that the analysis was repeated within primarily defined clusters, giving a hierarchy of BAPS clustering with more detailed sub-population structure. The analysis was performed by Professor Jukka Corander, resulting in 33 primary clusters (BCs) with 183 secondary clusters (sBCs) sequestered within the major clusters (**Appendix A**). **Figure 3.1 a** (inner ring) presents the population partition based on the secondary BAPS clusters. These secondary clusters were mostly clonal, and were separated mostly by MLST clonal complex boundaries. However, a group of singletons was clustered together. Based on their positions on the tree (generated in 3.3.1.2) and ST profiles, this cluster appears to be of different lineages; as a consequence, no particular cluster could be further assigned due to their low levels of similarity. These mixed clusters were removed from cluster-focused analyses to be discussed in chapter 4 and 6.

3.3.1.1.2 Minimum evolution tree

Based on the same alignment used in 3.3.1.1, an approximate maximum-likelihood phylogenetic tree was constructed using FastTree (**Figure 3.1 a**). This provided an independent validation for the population structure estimated by Bayesian clustering described above. With 1,000 resamples, 80.6% and 32.6% of the branches had over 0.700 and 1.000 bootstrap support respectively. *Streptococcus mitis* was used as an out-group to re-root the tree. Each branch in the phylogeny represented a cluster of isolates sharing the same ST, and in most cases the same serotypes except for serotype switch events, which will be further discussed in 3.1.3. Although isolates sharing the same ST cluster together, the distance between individual isolates within each cluster varied from relatively close to more distant. Many of the latter were observed on the branches of NT isolates, which lack genes for capsule biosynthesis

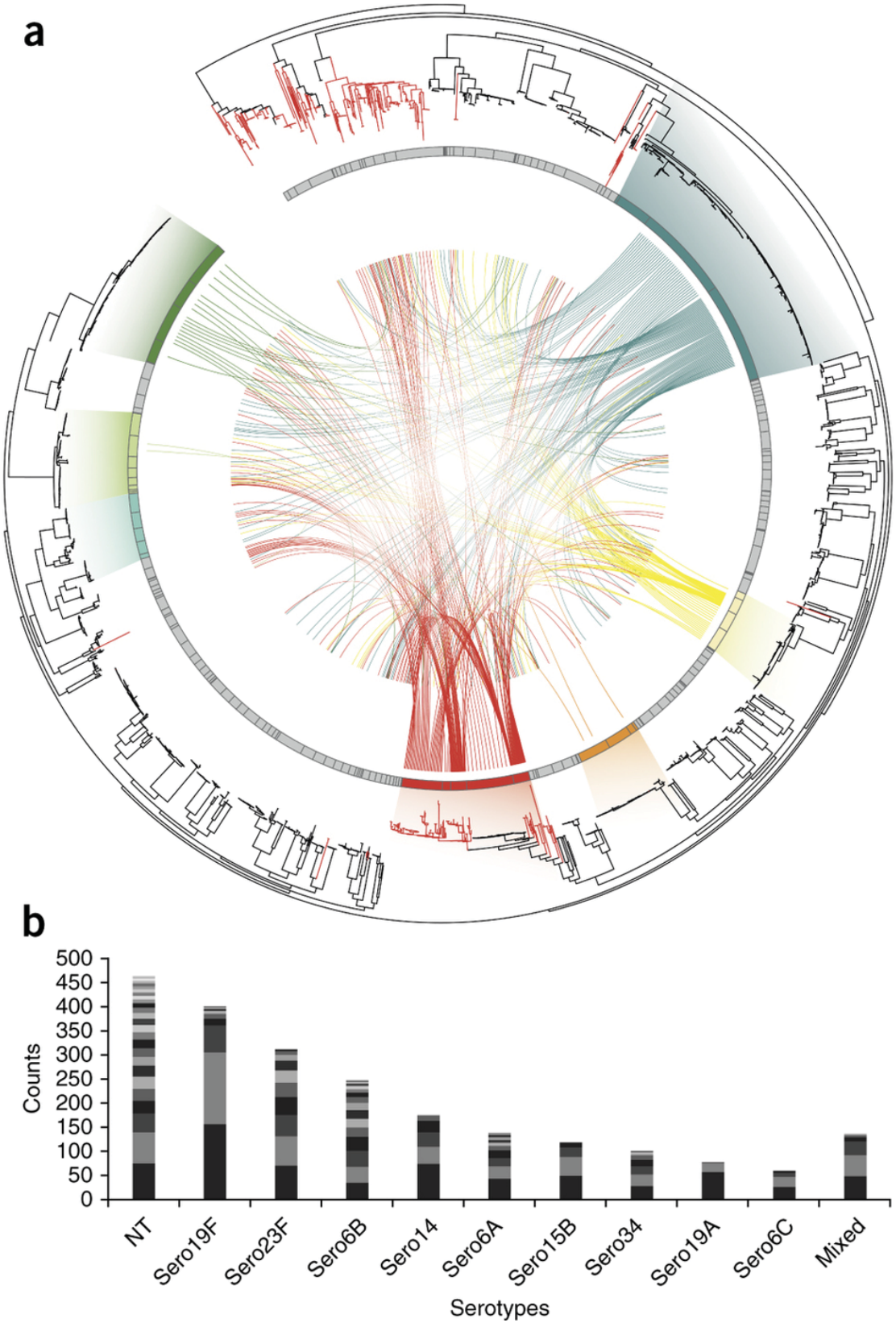
(highlighted in red in **Figure 3.1 a**). This observation could indicate that some lineages could have a higher evolutionary rates compared to others. Alternatively, this may suggest that the close clusters represent successful lineages that have undergone recent clonal expansion, whereas the distantly-related clades may represent rarer, less successful lineages.

3.3.1.1.3 Consistency between two methods

Although the minimum evolution tree did not consider genetic admixture from recombination fragments, the analysis of population structure from the Bayesian analysis and the phylogenetic approach largely agreed. Interestingly, a good agreement between these two approaches seems to suggest that recombination is largely observed between closely related lineages and is therefore less likely to affect overall population structure. On the other hand, if recombination across the species had dominated, more incongruence would have been seen between the two methods. Nevertheless, the consistent population structure resulting from both methods represents a result in which one can have more confidence.

Figure 3.1 Maela pneumococcal population structure

(a) SNP-based phylogeny of a pneumococcal population with connections between recipient and potential donors of recombination fragments. The outer ring shows a neighbour-joining tree built using SNPs from the whole population. Branches coloured in red are isolates classified as NT. The next ring represents population clusters based on secondary BAPS clustering. The seven most prevalent clusters are highlighted in different colours (clockwise): deep blue, BC1-19F; yellow, BC7-14; orange, BC6-15B and BC6-15C; red, BC3-NT; blue-green, BC5-23A and BC5-23F; pale green, BC4-6B; green, BC2-23F; gradients link these clusters to matched isolates on the phylogenetic tree. The centre shows connections between recombination recipients (from BC1-BC7 only; lines ending nearer the outside of the plot) and their potential donor clusters (lines ending nearer to the centre) (to be discussed in chapter 4). **(b)** Highly prevalent serotypes and their constituent population clusters by BAPS. The plot represents the ten most prevalent serotypes detected in the Maela population, each divided into separate population clusters on the basis of secondary BAPS clustering – serotype (number of clusters): NT (30), 19F (9), 23F (11), 6B (18), 14 (6), 6A (14), 15B (4), 34 (7), 19A (3) and 6C (4). Each cluster was shaded with different grey-scale to represent each genetic background, with NT showing the highest diversity. (**Figure is shown on the next page**)


a
b
Counts
500
450
400
350
300
250
200
150
100
50
0
NT
Sero19F
Sero23F
Sero6B
Sero14
Sero6A
Sero15B
Sero34
Sero19A
Sero6C
Mixed
Serotypes

3.3.1.2 Dominant lineages in Maela

The seven largest genotype groups determined by primary BAPS clusters (n>100 isolates) were denoted BC1 to BC7 and included the most common serotypes: BC1-19F, BC2-23F, BC3-NT, BC4-6B, BC5-23A/F, BC6-15B/C and BC7-14 (**Figure 3.1 a, Appendix A**), matching highly prevalent serotypes observed elsewhere in South East Asia (Jauneikaite, Jefferies *et al.* 2012). Isolates sharing the same capsule group may be part of different lineages, here represented by secondary BAPS clusters (**Figure 3.1 b**). NT pneumococci, which lack functional genes for capsule biosynthesis (*cps*) were the most prevalent capsule phenotype group in Maela; 512 isolates were classified in this group.

The Maela NT were made up of several genetically distinct groups, comprising 30 different secondary BAPS clusters. Five of these secondary BAPS clusters totalling 11 isolates (sBC 101-105) appeared to be more divergent from other NT pneumococci with their position being in close proximity to a *Streptococcus mitis* isolate on the phylogeny. Aside from these groups, other NT BAPS clusters, including one of the largest (BC3-NT), were distributed across multiple encapsulated pneumococcal clusters (**Figure 3.1 a**). Genetic variants within the *cps* loci of Maela NT were investigated by Susannah Salter, using sequence assembly and methods described in (Salter, Hinds *et al.* 2012). Partial deletion of the *cps* locus or disruption can lead to inactivation of capsule biosynthesis. Out of 512 isolates classified as NT, 42 appeared to have a partial or full deletion of the *cps* locus (**Table 3.1**). The remainder carried different sets of NT genes reported earlier (Salter, Hinds *et al.* 2012). Large sequence variations observed in the capsular loci of Maela NT implies that either: first, several independent events transformed the capsulated isolates circulating in Maela into NT; second, there were several introductions of NT into Maela from different origins; or a combination of both.

Table 3.1 Distribution of non-typable serotype (NT) in Maela

Categories		**Number of NT isolates**
Potentially non-pneumococcal streptococci (likely to be *Streptococcus mitis*)		11
Intact capsule locus but not expressing capsule (19A - 1 isolates, and serogroup 6 - like cluster - 3 isolates)		4
Group NT1: *cps* deletion/partial deletion		42
Group NT2: putative surface protein NspA		258
Group NT3: *aliB* genes	**Group NT3.1**: contain two *aliB* genes, *glf* pseudogene, *ntaAB* toxin/antitoxin	107
	Group NT3.2: contain *ISSpn10*, *aliB-1*, *aliB-2* pseudogene, *ntaAB*	51
	Group NT3.3: two *aliB* genes, *glf*	13
	Group NT3.4: *aliB-2, glf*	26
Total		512

3.3.1.3 Serotype switch events detected in the population

Some of the transitions from encapsulated to NT described above can be explained by serotype switch events. Serotype switches were determined by counting the states of changes in serotypes based on parsimony reconstruction of serotypes on the phylogenetic tree represented in **Figure 3.1 a**. Overall, 191 plausible capsule-switching events across the population were observed. 19 of these involved NT status. 9 events showed switches from encapsulated states to NT, whereas 9 events represented switches from NT to encapsulated states. One event had ambiguous direction. The results suggest that the conversion between the encapsulated and NT state is not uncommon in the Maela community and may play an important role in the population dynamics.

3.3.2 A snapshot of global pneumococcal population structure

3.3.2.1 Population structure and diversity

Carriage cohorts conducted between 2006-2010 from multiple geographical locations were selected for this study. In total 3,843 carriage isolates were recovered from children < 7 years from Southampton, UK (Tocheva, Jefferies *et al.* 2013); Massachusetts, USA (Croucher, Finkelstein *et al.* 2013); Kilifi, Kenya (Brueggemann, Muroki *et al.* 2013); The Gambia (unpublished data); and Maela refugee camp from Thailand (see materials and methods). While pneumococcal conjugate vaccines had already been introduced in UK (PCV7 in 2006, and PCV13 in 2010) and US (PCV7 in 2001), they had not been introduced in Kenya, The Gambia or Thailand during the sampling period; thereby exerting a large effect on differences in population structure detected in each population. STs detected from these data were clustered into clonal complexes (CCs), with members of each complex differing by single- (SLVs) or double-locus variants (DLVs). The clustering was performed using Phyloviz (Francisco, Vaz *et al.* 2012). The *ddl* locus was excluded because it is known to be linked to genes under stronger selective pressure and so could potentially bias the analysis (Enright and Spratt 1999).

There were 121 CCs (>2 STs) and 282 ST singletons identified among the entire collection of pneumococci. It is possible that the pneumococcal population detected in different countries may have different levels of diversity as a result of different carriage rates and vaccine administration. To test this hypothesis, the genotype diversity was compared by calculating the Simpson index of diversity for each population (Hunter and Gaston 1988), and 95% confidence intervals were calculated using the method of Grundmann (Grundmann, Hori *et al.* 2001). The index represents the probability that two consecutive isolates taken at random from a particular dataset will belong to different types. A higher index value is an indication of higher diversity. Discrimination indices of all studied locations fell in the range of 0.949 – 0.947 with largely overlapping 95% confidence intervals (0.937-0.976), showing no significant differences in overall population diversity either in different locations or with different vaccination programs (**Table 3.2**). The result is consistent with a previous carriage study from children < 2 years in Finland between 1995-1999

(Discrimination index: 0.981, 95% confidence intervals: 0.976 – 0.978) (Hanage, Kaijalainen *et al.* 2005). It is possible that there might be some unexplored diversity in the dataset due to the limited number of loci considered for this analysis. However, the similarity in magnitude of diversity observed here likely reflects the total capacity of nasopharyngeal colonisation that can be occupied by different pneumococcal lineages.

Table 3.2 Diversity captured through MLST in each sampling collection

Collections (Vaccinations)	Sampling size	Number of detected CCs	Number of detected ST singletons	Discrimination index (95% confidence interval)
Massachusetts, USA (PCV7)	280	51	12	0.949 (0.937-0.961)
Southampton, UK (PCV7/PCV13)	310	53	13	0.957 (0.946-0.967)
Kilifi, Kenya (No vaccination)	316	100	20	0.957 (0.949-0.965)
The Gambia (No vaccination)	445	63	4	0.968 (0.963-0.973)
Maela, Thailand (No vaccination)	2,492	170	29	0.974 (0.972-0.976)

3.3.2.2 More differences than similarities in population structure observed between countries

Although the pneumococcal population from each location appeared to have a similar degree of diversity, each population was comprised of largely different lineages. Indeed, 91.3% of the isolates described in the Kenyan carriage dataset, 75.7% in the Thai, 67.5% in the Gambian, 25.5% in the US and 21.0 % in the UK datasets were made up of unique CCs or ST singletons only detected in the local population but not elsewhere in this dataset (**Figure 3.2**). While the Kenyan, Gambian and Thai datasets encompassed a high proportion of distinct STs, there were fewer unique STs circulating in the UK and US. A pairwise comparison between these countries

revealed that the UK and US shared many common CCs or ST singletons with 74.4% of UK pneumococci matching the US, and 67.1% of US population matching the UK respectively (**Figure 3.3**). This high similarity between the UK and US populations could be due to relatively higher socioeconomic interactions between the two countries compared to any other studied locations. Alternatively, this could possibly be due to the impact of PCV7, which might have driven the post-vaccine populations in the same direction. The Kenyan carriage population appeared to be highly unique (91.7 % of its carriage population did not match elsewhere) with a small proportion co-detected in Gambian and Thai datasets. The Thai carriage population, which is the main subject of thesis, also consisted of a distinct population (75.7% did not match pneumococci observed elsewhere), many of which were contributed by NT isolates. A high proportion of NTs at this location appeared to shape the population behaviour and will be an important subject in this thesis.

In addition to the carriage cohorts described above, several invasive and carriage studies (1997 – 2010) conducted elsewhere have shown similar results with marked differences in population structure between developing countries including Nepal (Hanieh and Hamaluba *et al.* 2014), Nigeria (Adetifa and Antonio *et al.* 2012), and Ethiopia (Keenan and Klugman *et al.* 2014). Each of these countries displayed small number of overlapping STs with no close relatives. In contrast, population structures found in developed countries appear to share more similarity as observed in samples from Southampton, UK and Massachusetts, USA. This can be further supported by studies conducted in European countries including Oxford, UK (Brueggemann and Griffiths *et al.* 2003), Portugal (Simões and Pereira *et al.* 2011), Spain (Ercibengoa and Arostegi *et al.* 2012) and Norway (Vestrheim and Høiby *et al.* 2010) where a large proportion of common STs were reported. Overall, distinct carriage populations observed in developing countries, here represented by Kenya, Gambia, Thailand and other studies and their varied population structure likely suggest that the outcomes of clinical interventions identified from well-defined pneumococcal populations in developed countries like US, UK or other European countries may not be directly applied to developing countries due to variations in population structure.

Figure 3.2 Proportion of pneumococcal population commonly observed in multiple locations

For each location, a pie chart summarises proportions of pneumococcal population by number of isolates that were co-observed in other locations. Different colours denote co-detections of CCs or ST singletons that were common with other locations: shared with three other locations (red); shared with two other locations (orange); shared with one another location (cream) and uniquely observed in particular population (grey). These respectively represent 5.8%, 14%, 54.7% and 25.5% of US population; 11.8%, 13.8%, 53,4% and 21% of UK population; 2.9%, 8.8%, 20.8%, and 67.5% of Gambian population; 0%, 1.6%, 7.1% and 91.3% of Kenyan population; and 3.4%, 14.8%, 6.1% and 75.7% of Thai population. Note that the numbers are rounded up to one decimal place. The size each pie chart is correlated with the sample size of each study. (**Figure is shown on the next page**)

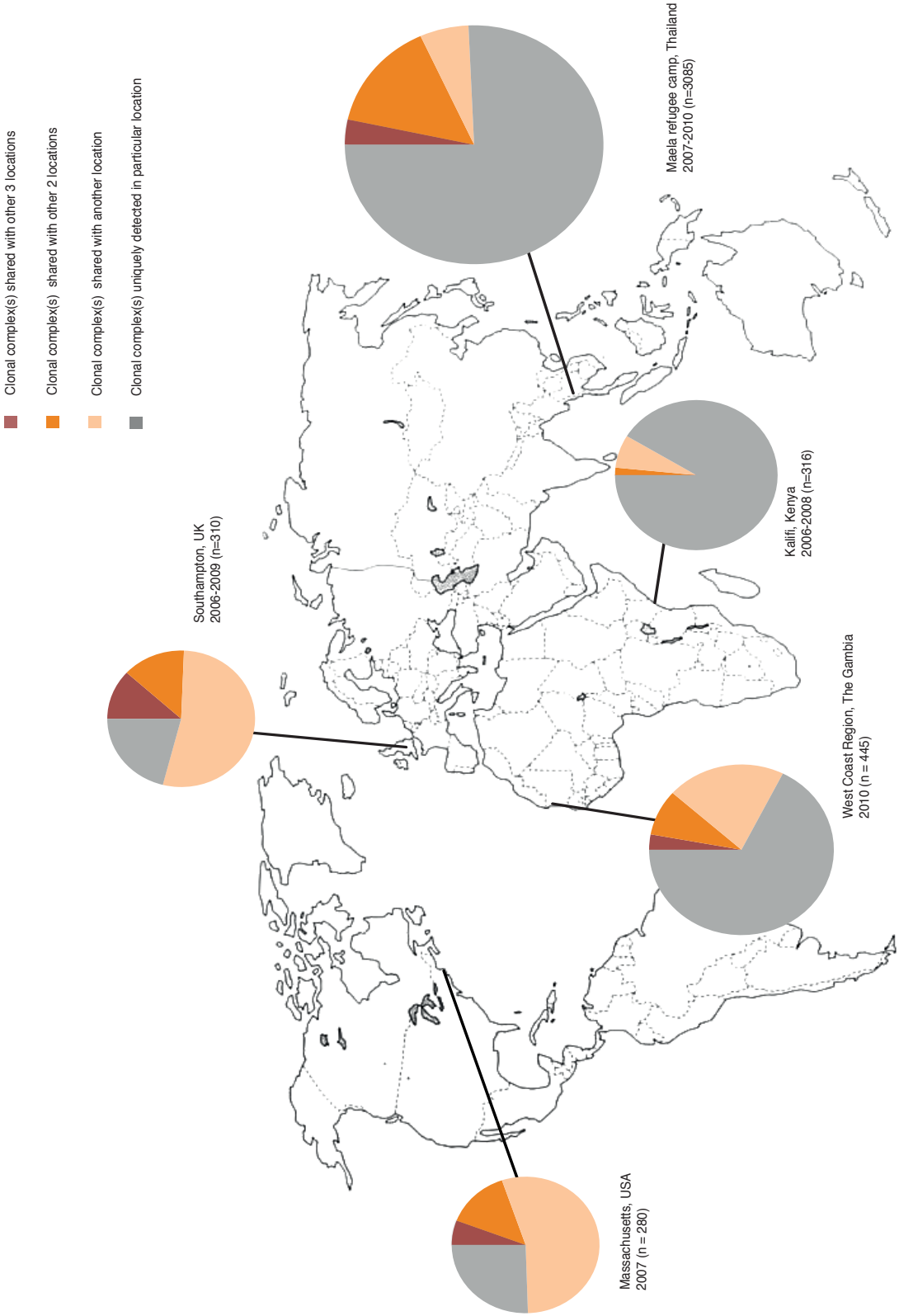
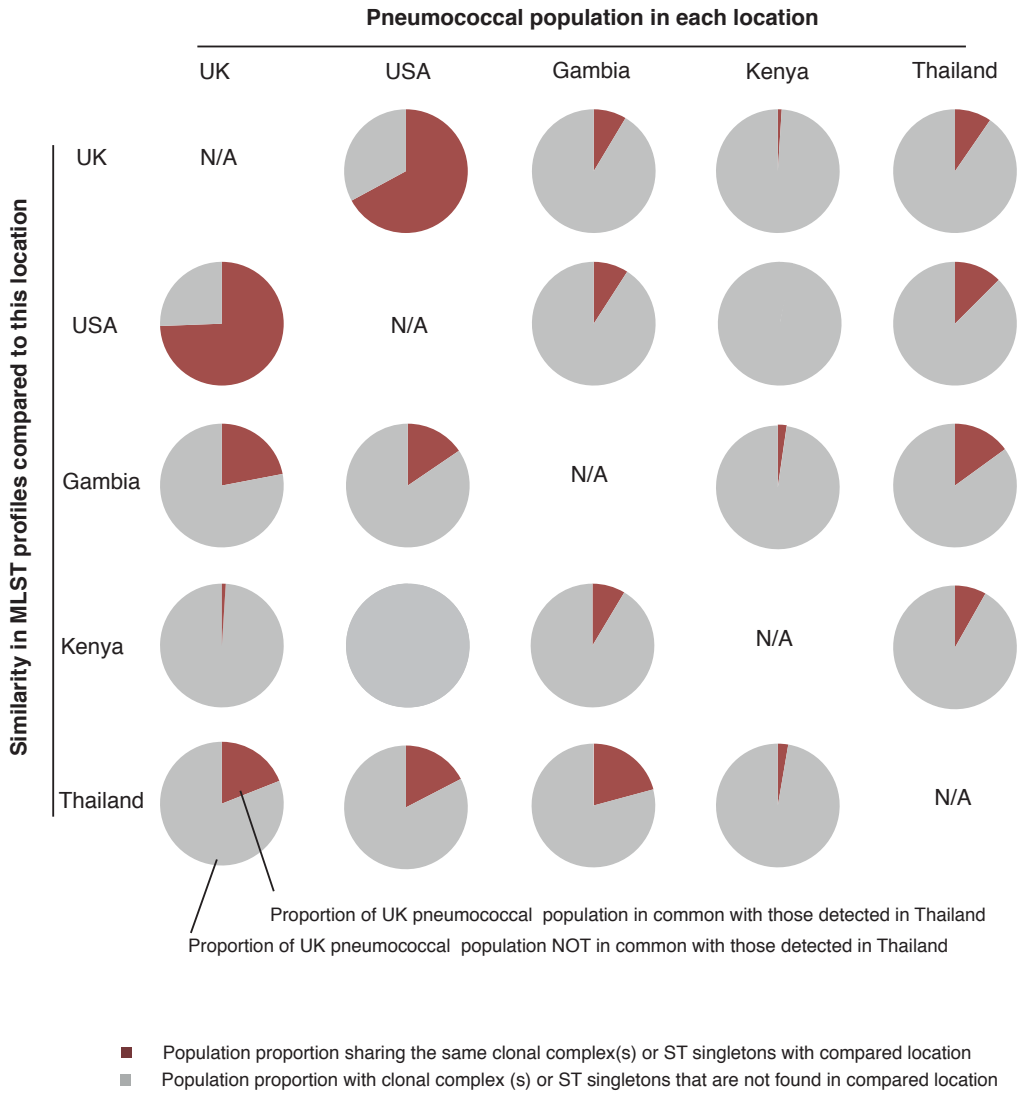

Clonal complex(s) shared with other 3 locations
Clonal complex(s) shared with other 2 locations
Clonal complex(s) shared with another location
Clonal complex(s) uniquely detected in particular location
Maela refugee camp, Thailand
2007-2010 (n=3085)
Kalifi, Kenya
2006-2008 (n=316)
Southampton, UK
2006-2009 (n=310)
West Coast Region, The Gambia
2010 (n = 445)
Massachusetts, USA
2007 (n = 280)

Figure 3.3 Pairwise comparisons of similarities and differences in pneumococci detected between different locations

Each pie chart summarises the proportion of the population by number of isolates from the location described in the column that were found in the population described in each row. Coloured in red is a proportion in particular location (column) that share the same CCs or ST singletons with another location (row). Grey represents a proportion in the population that does not share. The larger red proportion indicates the higher the population similarity between two locations.



3.3.2.3 Globally spread lineages

A small proportion of CCs or ST singletons were co-detected in multiple locations (**Figure 3.1**). In Kenya, Gambia, Thailand, US and UK: 1.6%, 11.7%, 18.2%, 19.8% and 25.6% of each population respectively matched populationd observed in at least 3 other locations. Many of the commonly detected lineages have been identified as PMEN clones, in part defined by their widespread nature (McGee, McDougal *et al.* 2001). Observed in these collections are PMEN3 (Spain^{9V}-3), PMEN14 (Taiwan^{19F}-14), PMEN25 (Sweden^{15A}-25), PMEN31 (Netherland^{3}-31), PMEN32 (Denmark^{14}-32), and PMEN43 (USA^{NT}-43). Consistent with these results, the public MLST database shows that these PMEN clones have been reported globally in Europe, North America, Latin America, Russia, Africa, Asia and Oceania (http://pubmlst.org/spneumoniae).

The prevalence of PMEN clones in multiple locations allowed different antigenic properties of the same clone to be compared to those from broader collections. Alternative serotypes, so-called serotype switches, were observed in many locations. This includes switches in PMEN3 (Spain^{9V}-3) from serotype 9V to 19A in the US data; in PMEN14 (Taiwan^{19F}-14) from serotype 19F to 19A in US; in PMEN31 (Netherland^{3}-31) from serotype 3 to 19F in the Gambia; in PMEN32 (Denmark^{14}-32) from serotype 14 to 19A in US and Thailand resulting in dominant 19A PMEN32 clones circulating in both areas; and in PMEN43 (USA^{NT}-43) from NT to 19F in the Gambia. These serotype switches, particularly from vaccine to non-vaccine serotype have been a great concern, as they contribute to serotype replacement and allow for the possibility of vaccine escape.

3.3.3 Multiple introductions of a globally spread lineage to a local community

One of the PMEN clones described earlier in 3.1.3, PMEN14 (Taiwan^{19F}-14) is highly prevalent in the Thai collection from the Maela refugee camp, representing one of the dominant clones detected in this community. The availability of whole genome sequences of PMEN14 isolates from both global (Croucher, Chewapreecha *et al.*

2014) and Maela collection enable one to understand how a clone circulating globally was introduced into a local area.

The first report of a PMEN14 isolate was from a Taiwanese hospital in 1997 (Shi, Enright *et al.* 1998) and was characterised as ST 236 with a serotype 19F capsule. Epidemiological surveillance subsequently detected isolates with a closely related genotype in Europe (http://pubmlst.org/spneumoniae), Africa (McGee, Klugman *et al.* 2001), and USA (Robinson, Edwards *et al.* 2001) with a large emphasis in South East Asia (Shi, Enright *et al.* 1998, Ip, Lyon *et al.* 2002). It was found to be among the highly multidrug resistant lineages found in carriage (Hanage, Bishop *et al.* 2011), a feature that potentially contributes to its success globally.

Whole genome sequencing was performed on 540 PMEN14 isolates (see methods), comprising 365 isolates from Maela refugee camp, Thailand and 175 isolates collected from twelve countries from the Middle East, Europe, USA and other South East Asian countries between 1997 – 2009 (Croucher, Chewapreecha *et al.* 2014). Forty isolates were associated with disease. The samples belong to CC320 comprising ST202, ST236, ST237, ST271, ST283, ST320, ST351, ST352, ST986, ST1583, ST1584, ST2116, ST2432, ST3259, ST3587 and ST4414. Reads from all samples were mapped onto a Taiwan19F-14 reference genome (accession number CP000921) following by variant calling (see methods). Recombined sequences were removed from the alignment using the method described in (Croucher, Harris *et al.* 2011). The algorithm removed 100,567 SNPs introduced by 892 recombination events from a total of 107,714 SNPs found in this dataset, allowing the maximum likelihood tree to be constructed using substitutions outside of recombination events. The most divergent isolate was found to be of ST1584 (a DLV of ST236) and was used to root the phylogeny.

The samples from Maela were polyphyletic with respect to the global samples with six distinct clades, indicating that the clone has been prevalent in South East Asia and entered the camp in at least six separate occasions (**Figure 3.4** left). Although Maela camp is remotely located with limited access controlled by the Thai authority, it is the largest refugee camp for Myanmarese in Thailand and has experienced several influxes of refugees. It is also considered a centre of studies for refugees with many

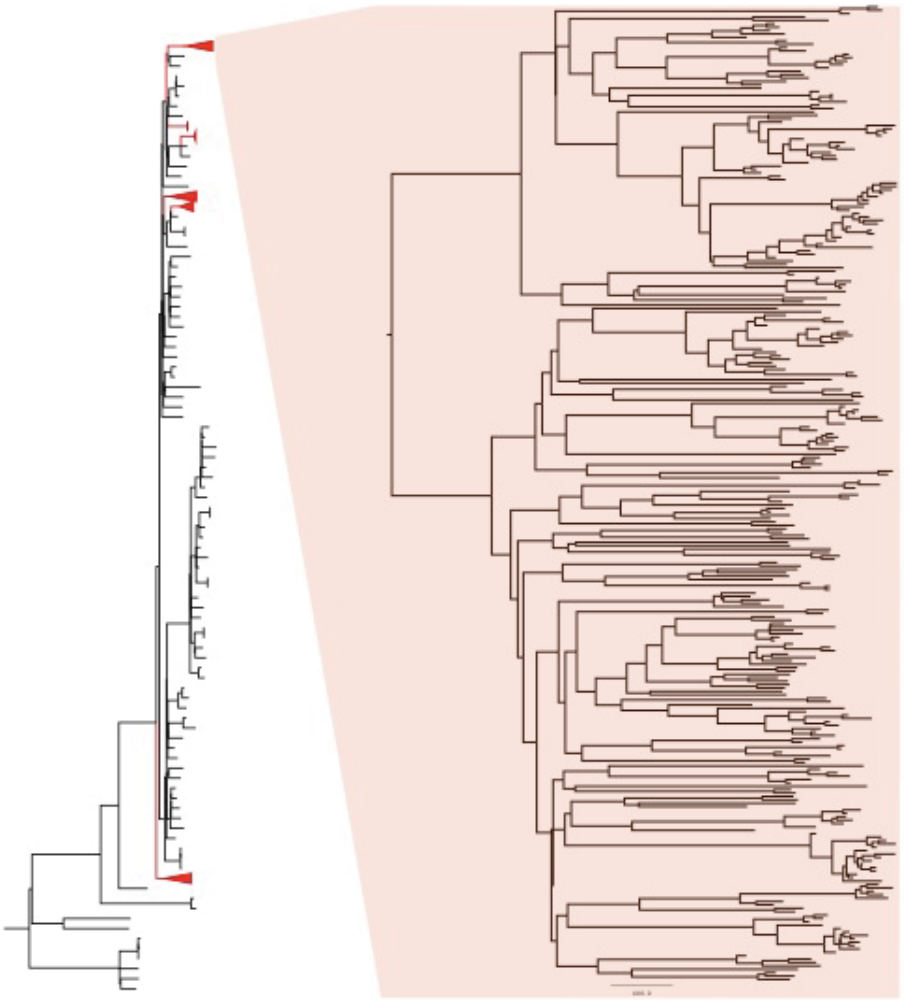
Myanmarese students coming to the camp for their education (The Border Consortium 2012). These regular visits from outside may explain the multiple introductions of a globally circulating lineage into this remote community.

To estimate the time of introduction, a Bayesian coalescent analysis was performed on the largest clade with a size of 288 isolates (**Figure 3.4** right). The clade accounted for 79% of PMEN14 isolates circulating in the camp and likely represented the most recent clonal expansion. Using BEAST software (see methods), the most recent common ancestor of this clade emerged around 1994 (95% confidence interval between 1988-1998), which postdated the official establishment of Maela camp in 1984. The estimated time roughly coincides with an increase in the Maela population size due to the closure of five other refugee camps: Mae-tawaw, Mae-salit, Mae-plu-so, Kler-kho and Kamaw-lay-kho in the north of Thailand in 1995. An increase in the camp population continued in 1997 and 1998 following the closure of Huai Bone and Shoklo camps. This mass influx of refugees from other camp sites into Maela during 1995-1998 approximately doubled its population size (The Border Consortium 2012). However, the large confidence interval on the date of origin of this clade make it difficult to conclide whether the population influx observed over this period had an impact on the introduction of this biggest PMEN14 clade into Maela.

Due to short sampling intervals and smaller sampling size, the most recent common ancestors of other clades could not be reliably estimated except for the one reported here.

Figure 3.4 Phylogenetic analysis of Maela pneumococci in comparison to global PMEN-14

(Left) Maximum likelihood phylogeny constructed using the vertically-inherited nucleotide substitution occurring in the sample taxa. The branch is coloured according to the location in which the strain was isolated: red presenting isolates from Maela refugee camp; and black from elsewhere in the world. The Maela branches were collapsed to reduce the figure size and facilitate graphical visualization. (Right) A temporal-phylogenetic reconstruction of the biggest PMEN-14 clade circulating in Maela using BEAST software.



3.4 Conclusion

This chapter gave a comparative view of the population structure based on available MLST data and demonstrated a marked difference in genotypic structure between countries. Therefore, the impact of clinical interventions and the potential for bacterial evolution in response to such selective pressure cannot be easily predicted based on the experience from a single country. Several globally spread clones were observed, most of which were identified as PMEN clones and have been under surveillance worldwide. The multiple introductions into Maela of one of these clones documented here demonstrates how a global clone can rapidly disseminate into a local area and produce an outbreak, even in a remote community like Maela refugee camp.

In addition to the global view, the chapter also gave a local view of a pneumococcal population at a high-resolution. A densely sampled collection of 3,085 complete genome sequences from Maela refugee camp over a period of 3 years allowed the population structure and the population-scale evolution to be studied at greater depth. A star-like phylogeny and a large number of distinct BAPS clusters showed a high diversity characteristic of multi-lineage population. Common serotypes detected in Maela matched those observed elsewhere in South East Asia, with NT being the most prevalent capsule group. A large number of NT observed in Maela and several conversions between encapsulated and NT states through serotype switching events detected here seem to suggest that it may not always be a disadvantage to lose the capsule. It is possible that the non-encapsulated state might confer some benefits to the pneumococci. This hypothesis will be tested in the next chapter.