Evolution of *Streptococcus pneumoniae* during carriage

Kamolchanok Claire Chewapreecha

Magdalene College

University of Cambridge

August 2014

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

I hereby declare that this dissertation is my own work and contains nothing that is the outcome of work done in collaboration with others, except where specifically indicated at the beginning of each chapter.

All sampling, population survey and microbiology work, which contributed to the metadata of this study were performed by Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University in Thailand through a collaboration with Dr Paul and Claudia Turner. The sequence data used in this thesis was generated at the Wellcome Trust Sanger Institute by Research Development and Sequencing production teams.

None of the work presented here has previously submitted for the purpose of obtaining another degree. This dissertation does not exceed 60,000 words in length, as required by the School of Biological Sciences.

Kamolchanok Claire Chewapreecha

August 2014

Acknowledgements

I am very grateful for Julian Parkhill and Stephen Bentley for trusting and giving me an opportunity to perform this work. I would like to thank both Julian and Ste for their strong support throughout my PhD, offering insightful and invaluable advice, and keeping me going when times were tough. Also, I sincerely thank Sharon Peacock and Matthew Holden, who also help guide me from day one.

I would like to extend my gratitude to my collaborators Paul and Claudia Turner, physicians who spent their PhD setting up a clinic in a remote Maela refugee camp to provide medical services to local people. I am in debt of their hard works and the resources they collected, which constitute almost this entire thesis.

Much of my bioinformatics work would not be possible without guidance and patience of Simon Harris and Nicholas Croucher. I also thank Jukka Corrander, Pekka Martinen, and their team for welcoming the computational challenges and helping develop analysing tools implemented on this study; David Aanensen for his help with graphical visualisation; Alison Mather for her statistical advice; and Susannah Salters for her microbiological expertise. This work could not have been completed without the informatics, systems, sequencing and library making teams of the Sanger institute. Also, I am thankful for all members of team 81, especially E212 residents for their general supports, understanding, and continuous cake supplies.

I thank Joy, all my friends and tutors for being with me through good and bad times during these nine years at Cambridge and yet making it so memorable. A thank to P'Yod whom the burden of mentoring me during difficult times has fallen most heavily. A final thanks to my family mum, dad, grandad, and late grandma for being so supportive and understanding throughout my time studying abroad, and always welcoming me back home with Thai food and warm hugs.

My PhD was funded by a Royal Thai Government Scholarship and a Wellcome Trust PhD Studentship.

Abstract

Streptococcus pneumoniae is a commensal bacterium asymptomatically carried in the nasopharynx of healthy individuals. However, if the bacterium escapes from its natural habitat to other anatomical loci, it can cause a range of invasive pneumococcal diseases, which make it a killer of over one million children annually. Despite high casualties, both treatment and prevention through vaccines have become more difficult as the bacteria rapidly develop antibiotic resistance and vaccine escape serotypes. To understand how this happens, one needs to look at evolution during carriage, a phase where exchange of genetic determinants for antibiotic resistance, virulence, and vaccine escape occurs *via* the process called "recombination".

This thesis summarises findings from a collection of 3,085 genome sequences of pneumococcal isolates from a rural community in Thailand called "Maela". This highly dense sampling gave an opportunity to investigate patterns of recombination and gene flows within the population, as well as changes in evolutionary patterns according to changes in selection pressure, especially the use of antibiotics over time. The non-encapsulated isolates, which are less invasive and unaffected by currently licensed vaccines, have a higher rate of both acceptance and donation of DNA *via* homologous recombination than encapsulated pneumococci. Highly exchanged genes include those associated with antibiotic resistance, implying that the non-encapsulates may act as a reservoir of resistance that can be passed to pathogenic strains and thus enhance the threat posed by antibiotic resistance.

However, the view from the Maela community may not be directly applicable to the population elsewhere, as different population structures may result in a different capacity for adaption. I therefore compared pneumococcal lineages detected in Maela with other contemporaneous carriage collections from the USA, UK, Gambia and Kenya based on multilocus sequence typing. The results showed that while the USA and UK share a lot of common lineages, large proportions of pneumococci detected in Gambia, Kenya and Thailand are unique to each location. Therefore, the propensity for genetic exchange may vary geographically and temporally.

The next part of the thesis identifies genetic determinants of resistance to beta-

lactams, a group of antibiotics frequently prescribed for upper respiratory infections. Here I performed a genome-wide association study - a technique commonly used in human genetics but difficult in bacteria due to their clonal population structure. Nevertheless, the large sample size and highly recombinogenic nature of *S. pneumoniae* allowed me to identify potential sources of resistance with improved resolution from "mosaic" genes described in the literature to several discrete causative sites, some of which are novel. The non-uniform distribution of these alleles in both vaccine-targeted and non-vaccine targeted lineages also highlights the limitations of vaccine in the control of spread of antibiotic resistance.

Together, this snapshot of the evolution of pneumococci and their interactions during carriage highlights the speed at which *S. pneumoniae* can adapt to new challenges, including antibiotics, while informing limitations in current health control policy.

Table of Contents

List of Figures

List of Tables

Abbreviations

ANCOVA	Analysis of Covariance
BAPS	Bayesian Analysis of Population Structure
BC(s)	primary BAPS Cluster(s)
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BLAST	Basic Local Alignment Search Tool
BURST	Based Upon Related Sequence Types
BratNextGen	Bayesian recombination tracker Next Generation
CC(s)	Clonal complex(es)
CLSI	Clinical and Laboratory Standard Institute
cps	capsular polysaccharide synthesis locus
CSP	Competence stimulating peptide
DNA	Deoxyribonucleic acid
DVL(s)	double-locus variant(s)
ENA	European Nucleotide Archive
ICE	Integrative conjugated element
IPD	Invasive pneumococcal diseases
MLEE	Multilocus enzyme electrophoresis
MLST	Multilocus sequence typing
MMR	Mismatch repair system
NT(s)	Nontypable
PCV	Pneumococcal conjugate vaccine
PFGE	Pulse field gel electrophoresis
PMEN	Pneumococcal Molecular Epidemiology Network
sBC	secondary BAPS cluster
SMRU	Shoklo Malaria Research Unit
SNP(s)	Single nucleotide polymorphism(s)
ST(s)	sequence type(s)
SVL(s)	single-locus variant(s)
TVL(s)	triple-locus variant(s)
WGS	Whole genome sequencing

Chapter 1: Introduction

1.1 Identification and characterisation of *Streptococcus pneumoniae*

1.1.1 A brief history

1.1.2 How pneumococci are characterised ?

1.1.2.1 Capsular typing

1.1.2.2 Multi-locus typing

1.1.2.3 Whole genome sequencing

1.1.2.4 Typing methods are affected by recombination

1.2 The pneumococcus has a highly recombinogenic nature

1.2.1 Mechanism of recombination

1.2.2 Early observations of pneumococcal recombination

1.2.3 A higher resolution of pneumococcal recombination from whole genome sequencing

1.3 The pneumococcus in carriage

1.3.1 Prevalence and duration of carriage

1.3.2 Interactions between pneumococci and other bacterial species

1.3.2.1 Interactions between pneumococci

1.3.2.2 Interactions between pneumococcus and other species

1.4 The pneumococcus in disease

1.4.1 Morbidity and mortality

1.4.2 Bacterial progression from carriage to disease

1.4.3 Factors influencing the transformation to disease

1.4.4 Limited genetic interactions in disease compared to carriage

1.5 Natural and clinical mechanisms for pneumococcal elimination and how the pneumococcus evolves to evade them

1.5.1 Clearance through natural host immune systems

1.5.2 Clinical interventions

1.5.2.1 Vaccines

1.5.2.2 Antibiotics

1.6 Project aims and objectives

1. Introduction

Streptococcus pneumoniae (the pneumococcus) is a Gram-positive bacterium of phylum Firmicutes. It is not only an important cause of wide ranges of diseases, mostly in young children and the elderly; but also commonly found as a commensal of the human nasopharynx. There are at least 94 serotypes of pneumococcus, with each producing a unique polysaccharide, called the capsule. Apart from these encapsulated groups, there are non-encapsulated pneumococci circulating in the same population. This chapter introduces the biology of pneumococci, their behaviours, prevalence in carriage and diseases, and the aims of this thesis in understanding more about this organism.

1.1 Identification and characterisation of *Streptococcus pneumoniae*

1.1.1 A brief history

S. pneumoniae is a long known bacterial commensal and pathogen. The organism was first isolated in 1881 from two independent works of the U.S. Army physician George Sternberg (Sternberg 1881) and the French chemist Louis Pasteur (Pasteur 1881). Both recovered the isolates from rabbits infected with saliva from human carriers. The rabbits died from septicaemia following the infections, leading to the hypothesis that the newly discovered bacterium is a disease-causing agent. Soon after, the role of the pneumococcus in causing human diseases was revealed when Carl Friedlander identified the organisms from the lungs of patients dying from lobar pneumonia (Austrian 1960). Friedlander distinguished the pneumococcus from other bacteria that other researchers had recovered from the pneumonia specimens and concluded that the pneumococcus was one, among other species, that was capable of causing pneumonia. Subsequent studies also established the pneumococcus as the cause of meningitis, arthritis, endocarditis and otitis media. Despite its importance in the aetiology of these diseases, the pneumococcus was frequently found in the nasopharynx of healthy carriers, making it a harmless commensal, as well as a pathogen. Early studies described the organism they recovered from clinical specimens as “having smooth and rough morphology”, a characteristic of capsulated

and non-encapsulated strains respectively. These studies also noted that capsulated strains were generally virulent in mice and rabbits while non-encapsulated strains were not (Austrian 1960). These observations significantly shaped the field of pneumococcal studies and for many decades, pneumococcal research focused heavily on capsulated strains.

1.1.2 How pneumococci are characterised ?

1.1.2.1 Capsular typing

Much effort has been focused on capsule typing due to the strong links that have been made between capsule type and invasive disease potential (Hausdorff, Bryant *et al.* 2000, Brueggemann, Peto *et al.* 2004). Capsulated strains are grouped according to their "serotypes", based on the antigenic variability of the capsule. The next section discusses pneumococcal classification based on capsular typing, including the typable and nontypable. The latter is a major focus of this thesis and will be discussed frequently in subsequent chapters.

1.1.2.1.1 Typable capsule

To date, at least 94 serotypes have been described on the basis of the capsule antigenic and biochemical properties, as well as their genetic differences (Bentley, Aanensen *et al.* 2006, Song, Nahm *et al.* 2013). Capsular typing was initially done through serological tests with each serotype identified by the interactions between the capsule antigens and specific antibodies. The original test was invented by a German physician and bacteriologist Fred Neufeld in 1902 (Neufeld 1902). He called the test "Quellung", which is the German word for swelling. Under light microscope, Neufeld observed swollen cells as the result of the binding of capsular polysaccharide of pneumococci with type specific antibody contained in the typing antiserum. Although the classical Quellung test is regarded as the golden standard for pneumococcal serotyping (O'Brien, Nohynek *et al.* 2003), the method is labour-intensive and costly as each test requires antisera against 90 pneumococcal polysaccharide capsules.

Moreover, expertise in microscopy is essential to determine the “swollen” reactions and the method is limited to experienced laboratories.

Several alternative methods have been subsequently developed to make serotyping more efficient and affordable. One of the commonly used methods is the latex agglutination test where antibodies are coated onto the surface of latex particles. The reaction between capsule antigens and specific bead antibody results in visible clumping called agglutination. The application to beads allows multiple specific antibodies to be coated together, allowing a quick determination for pools of select individual serotypes (Singhal, Lalitha *et al.* 1996, Slotved, Kaltoft *et al.* 2004). This method is faster than the quelling reaction and quickly narrows down the typing to a smaller group of serotypes. However, subsequent Quellung reaction is still often needed to identify individual serotype within the narrowed down pool. Due to its high sensitivity (Leinonen 1980) and cost-effectiveness (Lalitha, Pai *et al.* 1996), latex agglutination in combination with Quellung test has become a common qualitative test. This method is also used in studies described in this thesis (see methods 2.1.1).

Capsular typing can also be done based on the nucleotide sequences of genes involved in capsule biosynthesis pathway. With the exception to serotypes 3 and 37 where a single synthase gene is responsible for the production of capsule, the capsule biosynthesis is mediated by the *wzx*/*wzy*-dependent pathway and encoded by genes at the *cps* (capsular polysaccharide synthesis) locus. The *cps* locus typically encodes multiple glycosyl transferases that link sugars to create a specific polysaccharide subunit, which is then polymerised and translocated across the membrane to the surface of the cell to form the complete surface polysaccharide capsule (Bentley, Aanensen *et al.* 2006, Moscoso and Garcia 2009). Each serotype harbours a distinct combination of *cps* genes or alleles, allowing identification of serotype through nucleotide sequence analysis.

An advantage of using nucleotide sequences in determining capsule type is that the method can be quantitative relative to cell count. In a specimen where multiple serotypes are present, a qualitative test alone may only determine the presence of each serotype. Given known nucleotide sequences of *cps* locus, quantitative PCR-based (polymerase chain reaction) methods have been developed for detection of multiple

serotypes within a sample (Brito, Ramirez *et al.* 2003, Pai, Gertz *et al.* 2006). A microarray technique for detecting multiple serotypes has also been implemented (Newton, Hinds *et al.* 2011, Turner, Hinds *et al.* 2011). Interestingly, quantitative multiple serotyping through whole genome sequencing has not yet been reported, however, the technique is theoretically plausible and with whole genome sequencing becoming more routine, this might be available in near future.

1.1.2.1.2 Nontypable capsule

Not all pneumococcal isolates can be classified by capsule serotype. Isolates that cannot be typed by the methods described above are collectively termed “nontypable” (NT). Although most of the NTs appear to be non-encapsulated, they may possess a capsule for which there are no typing antisera or recognised nucleotide sequences, or they may produce the capsule erratically.

Many NTs show complete or partial deletion in the *cps* region, which disrupts the function of genes in the locus. Some have their *cps* locus replaced by alternative genes which are unrelated to capsule production, some encode a conserved transporter *aliB* which shares high sequence similarity with its orthologue detected in other streptococci, some encode a highly variable surface protein *nspA*. Together, these abort capsule synthesis, rendering typable strains NT (Salter, Hinds *et al.* 2012). In addition, some NTs still have their *cps* locus intact, but show defects in genes participating in the pathway. In two separate studies, single point mutations in a transferase gene, *wchA*, were found in NTs which were genetically identical to serotype 7F (Melchiorre, Camilli *et al.* 2012) and serotype 8 (Park, Geno *et al.* 2014). Both reported intact *cps* locus. Their switch from typable to NT was attributed to truncated WchA protein, which is required for capsule assembly. Recent availability of NT strains with whole genome sequences allows the diversity within *cps* locus or its remnant to be explored, providing information on the transitions between capsulated and non-encapsulated states.

1.1.2.2 Multi-locus typing

Based on sequences reported in (Bentley, Aanensen *et al.* 2006), the *cps* locus only accounts for 0.5% (serotype 3) to 1.5 % (serotype 38) of the pneumococcal chromosomal DNA. This represents a small proportion of the genome so the resolution of typing based on *cps* region or antigenic variability of the capsule may be limited. Higher resolution can be achieved by analysing multiple loci by techniques such as **M**ulti**l**ocus **e**nzyme **e**lectrophoresis (MLEE), **P**ulse **f**ield **g**el **e**lectrophoresis (PFGE) and **M**ulti**l**ocus **s**equence **t**yping (MLST).

MLEE was originally developed for characterising polymorphism in human populations (Harris 1966). The method uses relative electrophoretic mobilities of intracellular enzymes to differentiate between different types of organisms (Selander, Caugant *et al.* 1986), and was applied to pneumococci (Coffey, Dowson *et al.* 1991). The technique relies on different enzymatic phenotypes caused by differences in amino acid sequences in the population. Although this improves on the method of capsular typing by considering multiple proteomic loci, resolution is still limited by three factors. First, the electrophoretic mobilities of the enzymes may be the same regardless of altered amino acid sequences; second, the altered nucleotide sequences may not result in a change at the amino acid level; and third, enzymes coded by same amino acid sequences but with altered protein modification will have different enzymatic properties. Despite these limitations, MLEE helped demonstrate that the multidrug resistant isolates comprising serotype 23F and serogroup 19 detected in Europe and USA, were part of the same lineage (Coffey, Dowson *et al.* 1991, Munoz, Coffey *et al.* 1991). The observation was later confirmed using methods of higher resolution (McGee, McDougal *et al.* 2001, Croucher, Harris *et al.* 2011), and the lineage was termed PMEN1 (**P**neumococcal **M**olecular **E**pidemiology **N**etwork 1).

Similar to MLEE, PFGE also employs electrophoretic mobilities to determine relatedness between types of organisms. The technique was first used for studying *Saccharaomyces cerevisiae* populations (Schwartz and Cantor 1984) and was applied to prokaryotes, including *S. pneumoniae,* in 1987 (McClelland, Jones *et al.* 1987). PFGE measures the mobilities of multiple DNA fragments following digestion of the genomic DNA with restriction enzymes. This method relies on variation in size of DNA fragments due to DNA polymorphisms that alter recognition sites at which restriction enzymes digest the DNA. PFGE was used as a guide for identifying related

isolates and helped track the spread of antibiotic resistant isolates in Europe and the USA (Barnes, Whittier *et al.* 1995, Figueiredo, Austrian *et al.* 1995). However, it is often not clear whether PFGE bands of same size are related pieces of DNA, so not all unrelated isolates can be discriminated based on this method.

Unlike MLEE and PFGE which use electrophoretic mobilities as a proxy for population diversity, MLST directly measures variations of DNA sequences from multiple loci, thereby distinguishing groups of bacteria based on their unique allelic profiles known as sequence type (ST). The technique was successfully applied to *Neisseria meningitidis* (Maiden, Bygraves *et al.* 1998) and *S. pneumoniae* in 1998 (Enright and Spratt 1998). Pneumococcal MLST is based on polymeric gene fragments of seven house keeping genes: *aroE* (shikimate dehydrogenase), *ddl* (D-alanine-D-alanine ligase), *gdh* (glucose-6-phosphate dehydrogenase), *gki* (glucose kinase), *recP* (transketolase), *spi* (signal peptidase I) and *xpt* (xanthine phosphoribosyltransferase). The choice of seven genes represents a balance between identification power, time and cost for strain typing. MLST has been used widely in pneumococcal epidemiology with 9,712 STs from 22,714 isolates reported to date (10^{th} June 2014, http://pubmlst.org/spneumoniae)

These genes were originally chosen on assumptions that they are under neutral selection and accumulate genetic variations slowly over time. However, the assumption is violated in *ddl* gene as it is linked to a gene under strong selection pressure. Enright and Spratt reported a hitchhiking effect where recombination exchanges at penicillin binding protein 2b gene (*pbp2*), which is selected by penicillin consumption. The exchanges were often extended to *ddl* gene located 783 bp downstream of *pbp2b* gene (Enright and Spratt 1999). As this linkage could bias the typing and any estimate of the population structure, the locus has been excluded in many studies (Hanage, Kaijalainen *et al.* 2005, Hanage, Fraser *et al.* 2009), and also in this thesis.

MLST can be organised into higher hierarchical order, thereby providing some evolutionary contexts to the studied population. STs are often grouped into clonal complexes (CC) where members of each complex share a number of loci in common with other members. A CC typically comprises a founding ST and its descendants

(Feil, Li *et al.* 2004). The founding population gradually diversifies overtime resulting in variations in one, two or three of the seven MLST loci termed single-locus variants (SLVs), double-locus variants (DVLs), and triple-locus variants (TLVs) and so on, resulting in a CC. CCs can be constructed through a web-implemented algorithm called BURST (Based Upon Related Sequence Types) based on STs, allowing the expansion or emergence of clones to be put in the context of their genetic background (Spratt, Hanage *et al.* 2004). The primary founder is defined as the ST that differs from the largest number of other STs at only a single locus. Although this is useful as a guide to identify outbreaks and expansions, different algorithms and more genomic information are required to determine more deep-rooted relationships within the population.

1.1.2.3 Whole genome sequencing

Whole genome sequencing (WGS) offers great resolution for characterising pneumococci. Several whole genome sequencing platforms are currently available. However, this thesis focuses exclusively on Illumina data and this technology will be explained in the methods section.

While methods described above only capture a subset of total variations either from genetic contents or proximate phenotypes, WGS can capture all changes at all positions in the genome from single nucleotide changes to large-scale insertions and deletions. Reads generated from next generation sequencing are commonly processed in two ways, either mapping to closely related reference genomes or *de novo* assembly, in which no reference genome is needed. Each results in different resolution with varied possible applications. Mapping generally enables rapid identification of polymorphic changes. However, not all reads necessarily map to the reference genome as some regions in the test genome might not be present in the reference. Also, some portions of the reference genome may not be called reliably because too few reads have been mapped or the positions contain ambiguous consensus nucleotides, which are often observed in repetitive regions of the genome. These features are marked by ambiguity code N rather than A, T, C, G in the mapped genome and sometimes are filtered out in the downstream analysis depending on

research questions (for example see 4.2.1.3). *De novo* assembly assembles short reads into longer contiguous sequences called contigs. This approach allows large genetic variants, such as insertions, deletions, mobile genetic elements and rearrangement, outside those observed in the reference genome, to be captured. However, the method is sensitive to repetitive regions, resulting in misassembled contigs. Both mapping and *de novo* assembly generate sequences from which genetic variations are called. In many studies including this thesis, both approaches are used to complement each other in order to capture all diversity in the population (Loman, Constantinidou *et al.* 2012, Wilson 2012).

A greater power in detecting variations provides a higher population resolution but also presents a computational challenge to the analysis. For each isolate, total allelic sequences used to determine MLST account for 0.145% of the whole genome sequences, indicating around 690 times increase in the amount of information processed in WGS compared to MLST. Several algorithms have been developed to subdivide population into groups of closely related strains while dealing with large diversity. These include phylogenetic reconstruction and Bayesian clustering approaches, which are employed for studies documented in this thesis and will be discussed in depth in the following chapters.

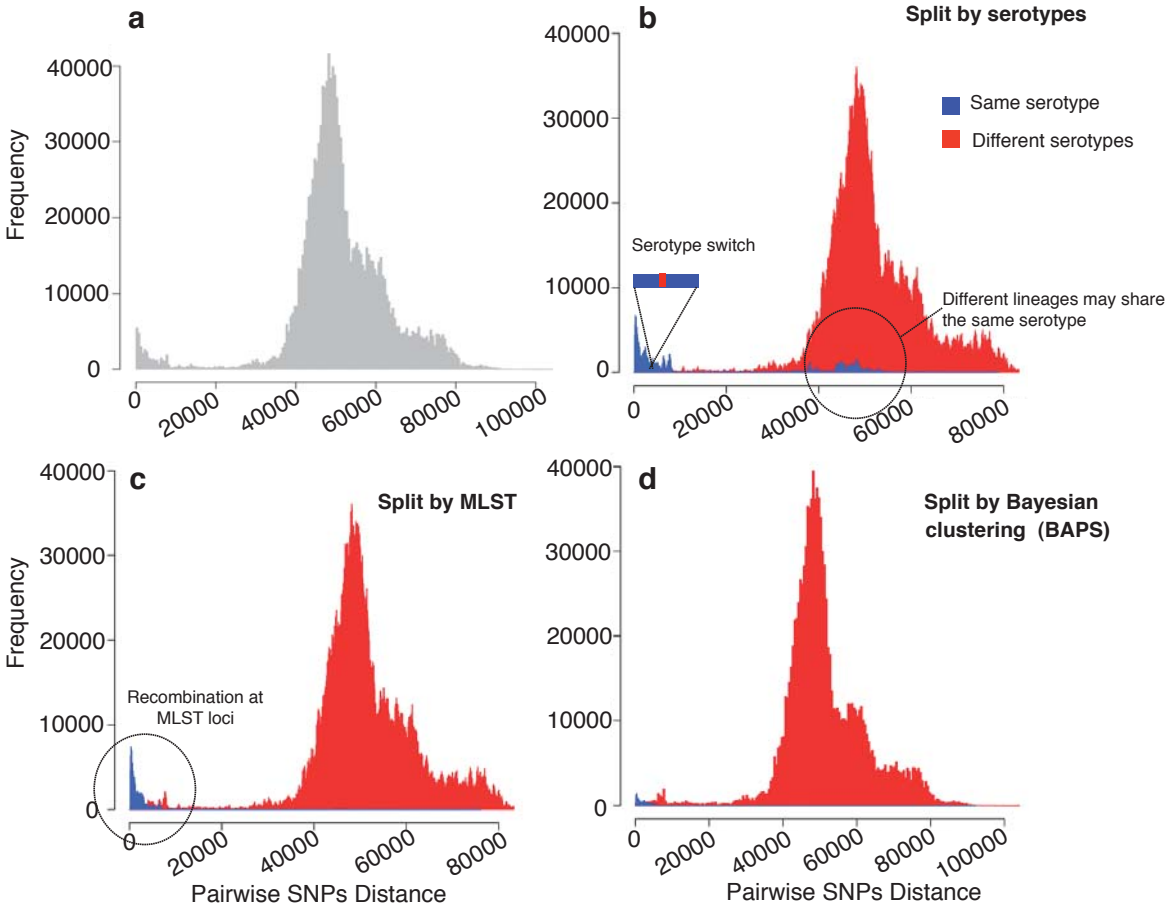
1.1.2.3 Typing methods are affected by recombination

Although various typing methods have been invented, they are all affected by the frequent recombination events observed in the pneumococcus. Recombination will be discussed in 1.2.1. In brief, the process allows an exchange between two similar DNA molecules. This creates mosaic genes in which different parts of the genome exhibit different evolutionary histories (Spratt 1988). If recombination occurs in genetic segments required for molecular typing, the isolates may be reported as unrelated despite having nearly identical background. This principle is demonstrated in **Figure 1.1** using pairwise Hamming distance. This represents differences in single nucleotide polymorphisms (SNPs) between any two randomly paired isolates in a subset of the species-wide pneumococcal population reported in this thesis (see 2.1.1). The distribution of all pairwise comparisons represents genetic relatedness of the studied

population. **Figure 1.1a** highlights two major peaks, with each showing a different pattern of evolutionary relationships. A smaller peak positioning toward the left (<10,000 SNPs) likely reflects the past evolution within a single lineage, while the larger peak positioning toward the right (>10,000 SNPs) likely represents pneumococcal evolution within the whole population. Applying information from different typing methods on this distribution reveals that neither capsular typing nor multiple loci typing (here presented by MLST) could completely differentiate isolate relatedness. **Figure 1.1b** demonstrates that isolates bearing the same serotype do not necessarily have the same evolutionary history. Likewise, isolates of the same genetic background do not necessarily have the same serotype. The latter is known as "serotype switch", many of which were mediated by recombination as shown in (Brueggemann, Pai *et al.* 2007, Pillai, Shahinas *et al.* 2009, Golubchik, Brueggemann *et al.* 2012, Wyres, Lambertsen *et al.* 2013) and have clear relevance to vaccine design. Recombination poses similar problems to MLST if one or more of the seven loci are affected. Isolates having similar genetic background can either be classified as same or different STs. An example of recombination event here includes an exchange of an *aroE* allele, which resulted in a change from ST802 to ST4413 (**Figure 1.1c**). Another method in characterising pneumococcal population is Bayesian clustering (BAPS – Bayesian analysis of population structure) which employs information generated from WGS to subdivide species-wide data into smaller clusters based on their sequence similarities (Corander, Marttinen *et al.* 2008). Here, BAPS clustering appears to be least affected by recombination and shows relatively clear separation between closely and distantly related isolates (**Figure 1.1d**). This is possibly due to noise from recombination being suppressed by true evolutionary signal from the most common history elsewhere in the genome; thereby providing a more accurate view of pneumococcal population structure. Nevertheless, pairwise distances classified under same serotype, ST and Bayesian clusters all show overlapping regions with those classified under different categories, suggesting that the typing or clustering boundary for pneumococci is blurred by recombination effects. This highly recombinogenic nature makes *S. pneumoniae* difficult yet interesting to study. The recombination process itself will be discussed in the next section.

Figure 1. 1 Effect of recombination on pneumococcal typing

a) A histogram of SNP pairwise distance between isolates of a species-wide pneumococcal population reveals distinctions between groups of closely related (pairwise differences < 10,000 SNPs) and distantly related isolates (pairwise differences > 10,000 SNPs). Panels b)-c) categorise distribution observed in a) based on serotyping, MLST and population clustering respectively. Pairwise distributions of the same serotype, ST or Bayesian cluster are highlighted in blue, while different categories are coloured as red. Overlapping region between the blue and red peaks can be observed in b)-d)



1.2 The pneumococci have a highly recombinogenic nature

More than sixty bacterial species from both Gram positive and negative phyla, have been described as “naturally transformable” (Johnsborg, Eldholm *et al.* 2007). This refers to an ability to take up, incorporate and express extracellular DNA found in the environment. The incorporation of exogenous DNA involves the process called recombination. This allows the maintenance of genetic diversity in bacterial population while counteracting the accumulation of deleterious DNA in a Muller’s ratchet fashion – the process by which the genomes of an asexual population accumulate deleterious mutations in an irreversible manner (Muller 1964, Moran 1996). Recombination also provides a mechanism for bacteria to make large evolutionary leaps which can be important for adaptions.

There are two main types of recombination described in bacteria. The first form is homologous recombination, which involves the replacement of DNA sequence with a significantly homologous i.e. highly similar but potentially distinct sequence. A second form of recombination known as illegitimate or non-homologous recombination involves the integration of a distinct DNA sequence (often a bacteriophage or transposon) into a largely non-homologous sequence of a recipient chromosome. Recombination can occur between sequences originating from within the same organism or from different organisms. In *S. pneumoniae*, the term recombination encompasses both homologous recombination mediated by the competence system, as well as site-specific integration of prophage and integrative conjugated element (ICE)(Lawrence and Retchless 2009, Didelot and Maiden 2010).

1.2.1 Mechanism of recombination

1.2.1.1 Transformation is induced by competence system

Pneumococci have a competence system - a dedicated machinery for promoting the uptake of exogenous DNA. Competence is a transient physiological state associated with the up-regulation of two sets of competence (*com*) genes which are expressed at

different states, the early genes involved in triggering an intercellular signalling; and the late genes includes those promoting DNA processing and genetic recombination (Campbell, Choi *et al.* 1998, Peterson, Cline *et al.* 2000). Competency is a tightly regulated process, influenced by different environmental signals. Under laboratory condition, competent state is induced in rich media during logarithmic growth phase when the cell density is high. However, such condition may not be easily obtained in the nasopharyngeal habitat due to limited nutrients available and killing from host defence system, which is triggered upon excessive bacterial load. Competence in nature is likely to be induced by stress and can be regarded as a fitness-enhancing strategy in response to stress conditions. Stress factors inducing competency include a change in external pH, temperature, and antibiotic-induced stress (Tomasz 1970, Claverys, Prudhomme *et al.* 2006, Prudhomme, Attaiech *et al.* 2006). Prudhomme *et al.* showed that sub-lethal concentrations of aminoglycoside antibiotics, which blocks ribosome function; fluoroquinolones, which inhibit topoisomerases; and the DNA-damaging agent mitocin C, all stimulate competence in *S. pneumoniae* (Prudhomme, Attaiech *et al.* 2006).

The signalling cascade for competency is mediated by *comAB* and *comCDE* operons. The cell-cell signalling is triggered by a 17 amino acid competence stimulating peptide (CSP) encoded by *comC*. The CSP is transported outside the cell via ComAB transporter and activates the membrane-bound histidine kinase receptor ComD of the neighbouring cells. This leads to ComD autophosphorylation followed by the transfer of the phosphoryl group to its response regulator ComE, resulting in ComE activation. The response regulator ComE then directly activates the expression of the other twenty earlier competence (*com*) genes by binding to the recognised motif in their promoter regions. This initiates a transcriptional cascade resulting in DNA uptake and recombination. The presence of this motif in *comAB* and *comCDE* operons suggests a binding of ComE to these operons creating a positive feedback of early *com* genes. Following signal amplification, this cascade signalling can be terminated through a two-component system encoded by *ciaHR*, allowing an effective control over the transformation process. (Tomasz 1970, Claverys and Havarstein 2002, Sebert, Patel *et al.* 2005, Claverys, Prudhomme *et al.* 2006, Claverys, Martin *et al.* 2009)

Through DNA microarray, the signalling cascade following CPS secretion was shown to alter expression profiles of 105-124 pneumococcal genes. Within this is the early expressed *comX*, coding for alternative σ factor. ComX controls the expression of around 60 late *com* genes. These include genes encoding the killing factor associated with fratricide, which promotes the release of DNA from neighbouring cells; as well as genes involved in DNA uptake and recombination (Claverys and Havarstein 2007).

1.2.1.2 Uptake and entry of exogenous DNA

An uptake of exogenous DNA is triggered by the binding of double-stranded DNA to the pseudopilus and progressed to the uptake pore complex (Campbell, Choi *et al.* 1998). One strand is degraded while the other is cleaved into smaller fragments as it enters the cell. Proteins coded by *recA* and *dprA* are loaded on to the single stranded DNA, forming a nucleoprotein complex in the cytosol. This complex is capable of invading the chromosomal regions with sequence similarity, allowing an integration of exogenous DNA into host chromosome (Chen, Christie *et al.* 2005).

1.2.1.3 A successful integration of exogenous DNA into host chromosome

Not all invasion from exogenous DNA discussed in 1.2.1.2 results in a successful transformation. Small genetic variations from imported DNA can be recognised and corrected by host mismatch repair system (MMR), resulting in no net variations being introduced. By estimating the frequency of polymorphism markers pre and post transformation, *in vitro* studies showed that small genetic variations including deletions of 3 bp or shorter and transition mutations, could be corrected efficiently by MMR. However, the efficiency of the repair system reduced when genetic variations between the host and imported DNA were larger as observed for deletions of 5 bp or longer (Lacks, Dunn *et al.* 1982, Gasc, Garcia *et al.* 1987). An increase in imported variations to 150 polymorphisms or more was shown to saturate and thus abolish the MMR system (Croucher, Harris *et al.* 2012). However, a degree of sequence similarity is still required for homologous recombination as it was shown that the frequency of recombination decreased upon an increase of sequence divergence

(Majewski, Zawadzki *et al.* 2000). This potentially acts as a barrier to genetic exchanges between the pneumococci and distantly related organisms.

Pneumococci were frequently observed as successful recipients for homologous DNA from their own species as well as from other related streptococci (Dowson, Coffey *et al.* 1993). This either suggests that homologous recombination is tolerant to ranges of sequence similarity, or indicates less sequence variations between pneumococci and other streptococci. Donati *et al.* compared sequence similarity between different *S. pneumoniae* and other groups of streptococci using a pan-genome approach (Donati, Hiller *et al.* 2010). They reported that within a group of pneumococci, 74% of sequences were conserved. The conservation dropped when pneumococci were compared against other streptococci. On average, 48%, 51%, 53% and 55% of *S. pneumoniae*, *S. mitis*, *S. oralis* and *S. infantis* genomes can be aligned against all other streptococcal sequences, indicating some sequence similarity that might allow homologous recombination between these species. Moreover, a streptococcal phylogeny generated from the same study placed *S. pneumoniae* closer to *S. mitis* than other species, suggesting a genetic similarity in support of frequent sequence transfers between *S. pneumoniae* and *S. mitis* (Dowson, Coffey *et al.* 1993, Chi, Nolte *et al.* 2007).

Together, these dedicated mechanisms in promoting the uptake of exogenous DNA, mediating the cell entry and integration of exogenous DNA into host chromosome make *S. pneumoniae* a highly recombinogenic organism in comparison to many other bacterial species (Thomas and Nielsen 2005).

1.2.2 Early observations of pneumococcal recombination

The transforming ability of *S. pneumoniae* was first recognised in 1928. This was also the first demonstration of “transforming principle” in any organism and subsequently led to a landmark discovery that established DNA as the hereditary material in 1944.

Frederick Griffith, a British bacteriologist, studied two *S. pneumoniae* strains of different surface morphologies, a smooth (type III-S) and rough (type II-R) strain. As

discussed in 1.1.1, the encapsulated smooth strain was capable of causing an infection in mice while the rough non-encapsulated strain was not. The difference in infectious outcome was due to the polysaccharide capsule, a virulence factor known to protect the cell from phagocytosis by host. Griffith showed that mice infected with heat-killed smooth strain alone, or live rough strain alone, were alive. However, infection with mixed heat-killed smooth strain and live rough strain resulted in the death of mice. Both live smooth strain and rough strain could be extracted from the blood of the dead host. He concluded that there was an hereditary material from the heat-killed smooth strain that transformed the rough strain changing their phenotype from non-virulent to virulent (Griffith 1928).

The hereditary substance was proved to be DNA through the experiments Avery, MacLeod and McCarty published in 1944. The experiments followed the design of Griffith's, however, the heat-killed smooth strain was additionally treated with deoxyribonucleopolymerase (to digest DNA), ribonuclease (to digest RNA), trypsin and chymotrypsin (to break down proteins) and enzyme hydrolysis of capsule (to destroy other cellular component). While a separate treatment of capsule hydrolysis, trypsin, chymotrypsin and ribonuclease did not affect the transformability between the smooth and rough strains; a treatment with deoxyribonucleopolymerase stopped the transforming ability. This led to an important conclusion that hereditary information was conferred by DNA, not RNA, proteins nor any cellular components (Avery, Macleod *et al.* 1944).

1.2.3 A higher resolution of pneumococcal recombination from whole genome sequencing

1.2.3.1 Recombination in a single isolate

The genetic codes that potentially transformed Griffith's non-virulent rough strain into virulent smooth strain were not known until 79 years later when whole genome sequences of the two strains were first available (Lanie, Ng *et al.* 2007). Lanie *et al.* compared the sequence of D39, a historically virulent serotype 2 strain used in the Griffith and Avery, MacLeod and McCarty experiments to a non-virulent non-encapsulated derivative of D39 called R6 strain. Although this investigation was

performed many decades later, the group reported unchanged phenotypes of D39 and R6 to the original studies, suggesting that their genotypes are likely to remain largely stable. Sequence comparison between the two strains revealed genetic variations that might differentiate their virulent phenotypes, some of which were possibly transformed in the original studies. Lanie *et al.* confirmed a large deletion of part of the capsule locus in R6 relative to D39, supporting a former hypothesis of the capsule acting as a virulent factor. Moreover, 71 nucleotide substitutions, six deletions and four insertions were found in R6 with respect to D39. Some of these variations matched known virulence determinants (Mitchell and Mitchell 2010), many others are in or affect the expression of genes that function in regulation and metabolism that indirectly result in different phenotypes between the two strains (Lanie, Ng *et al.* 2007).

1.2.3.1 Recombination in a population

More informative investigations into recombination have been possible using large scale sequencing data. Since the first publication of complete whole genome sequence of *S. pneumoniae* in 2001 (Tettelin, Nelson *et al.* 2001), there have been increasing numbers of large-scale whole genome studies, with most recent studies comprise n > 1,000 genomes (Croucher, Harris *et al.* 2011, Croucher, Harris *et al.* 2012, Everett, Cornick *et al.* 2012, Croucher, Finkelstein *et al.* 2013, Chewapreecha, Harris *et al.* 2014, Croucher, Chewapreecha *et al.* 2014, Croucher, Hanage *et al.* 2014). These studies not only permit the investigation of recombination at high resolution but also provide a genomic perspective on pneumococcal adaptation to environmental stresses through recombination.

An important study by Croucher *et al.* applied whole-genome sequencing to 240 isolates of a PMEN1 lineage from diverse geographical locations (Croucher, Harris *et al.* 2011). By comparing genomic sequences of numerous isolates derived from a common ancestor, the authors distinguished the regions of the genome that had undergone recombination, from nucleotide substitutions (see 4.1.1 for discussion on Croucher *et al.* methods). Impressively, 74% of the pneumococcal genome was affected by recombination in at least one isolate. The study reported a total of 57,736

SNPs, of which 50,720 SNPs (88%) were introduced by 702 recombination events. The polymorphisms introduced by recombination were 7.2 times greater than those generated by nucleotide substitutions, highlighting the magnitude of impact exerted from recombination on pneumococcal evolution. Recent application of mathematical models (Mostowy, Croucher *et al.* 2014) on this (Croucher, Harris *et al.* 2011) and another separate dataset (Croucher, Hanage *et al.* 2014) further described heterogeneity in the recombination process, comprising single, short, frequent replacements - termed micro-recombination, and rarer, multiple-fragment, saltational replacement termed macro-recombination. Mostowy *et al.* linked macro-recombination (>19 kb) to major phenotypic changes, including serotyping-switching events observed in Wyres *et al.* (Wyres, Lambertsen *et al.* 2013) and concluded that macro recombination might be a major driver of the diversification of *S. pneumoniae*.

Another study by Croucher *et al.* characterised the recombination mechanism through *in vitro* genome-wide transformation (Croucher, Harris *et al.* 2012). Croucher generated high-resolution snapshots of individual transformation events against known genetic backgrounds, allowing identification of the position and sizes of the recombinant fragments, along with physical properties of the DNA that affected the recombination process. The study showed that DNA was a key factor regulating transformation as increasing concentration of donor DNA by 100-fold caused a 38-fold increase in the number of transformants. In nature, this extracellular DNA is likely to be released from other pneumococci or other organisms that co-colonise the same host nasopharynx. The natural habitat of pneumococci in which this process occurs will be discussed in the next section.

1.3 The pneumococci in carriage

Pneumococci form part of nasopharyngeal microbiota that do not usually harm the host. However, the carriage state is known to be a reservoir where it can give rise to disease if the pneumococci extend to other areas of the respiratory tract or penetrate normally sterile body fluids (Austrian 1986). Since nasopharyngeal colonisation normally precedes invasive disease and potentially acts as reservoir for spread in the host population, asymptomatic carriage is regarded as a risk factor for the development of pneumococcal disease (Darboe, Fulford *et al.* 2010).

1.3.1 Prevalence and duration of carriage

The carriage of *S. pneumoniae* is widely prevalent in young children and disproportionately found at a higher rate in developing countries compared to developed countries (O'Brien, Wolfson *et al.* 2009). Pneumococcal acquisition rates were reported to be 2-6 times higher in children than adults (Mosser, Grant *et al.* 2014). Some children can be colonised within the first days after births, while 50% to 90% of children are found to be colonised several months later (Hill, Akisanya *et al.* 2006, Granat, Mia *et al.* 2007). The prevalence of colonisation peaks in the first three years of life with an estimate rate of colonisation of 20% or greater, then starts to decline at the age of ten (Bogaert, De Groot *et al.* 2004, Bogaert, van Belkum *et al.* 2004). A recent study explored the trend in 29 carriage studies, with a total of 20,391 individuals, and showed that, in all cases, nasopharyngeal carriage decreased with increasing host age (Le Polain de Waroux, Flasche *et al.* 2014). The same study offered predictions of carriage rates in adults and school aged children (5-17 years) based on their correlation with carriage rates in young children (<5 years old). A decrease in carriage rates over increasing age is likely due to the development of more mature immunological response. In addition, a change in hormonal control that leads to a shift in microbiota, may play a role in a decreased carriage rate when children enter puberty (Oh, Conlan *et al.* 2012). Although responses vary between individuals, the pneumococcal carriage generally remains low in elderly (> 65 years) with estimate rate of colonisation of 5 % or lower in the community (Flamaing, Peetermans *et al.* 2010, Palmu, Kaijalainen *et al.* 2012). While high carriage rates in young children

correlate with high risk of invasive pneumococcal diseases (IPD), high incidence of IPD detected in elderly cannot be directly correlated with low carriage rates observed in this age group. This complication is possibly due to heterogeneity in disease susceptibility among elderly individuals (Simell, Auranen *et al.* 2012).

Apart from the age group, several risk factors have been shown to promote acquisition of pneumococci. Evidences suggest that young children living close together, for instance in a day care centre or in a group of siblings, have increased level of pneumococcal carriage (Bogaert, van Belkum *et al.* 2004, Regev-Yochay, Raz *et al.* 2004). Different sugar condition from the child's diet can also influence the nasopharyngeal microbiota and the prevalence of pneumococci. A recent study showed that high intake of carbohydrate including sweet pastries and jam was associated increased risk of pneumococcal carriage (Tapiainen, Paalanne *et al.* 2014). An exposure to tobacco smoke in the household was also associated with pneumococcal carriage as well as promoting IPD (Nuorti, Butler *et al.* 2000, Coen, Tully *et al.* 2006, Greenberg, Givon-Lavi *et al.* 2006, Cardozo, Nascimento-Carvalho *et al.* 2008). There may be a role of host genetic factors that promote pneumococcal colonisation in certain ethnicity with higher carriage rates observed in Australian Aboriginal children than non-Aboriginal children (Watson, Carville *et al.* 2006), and in native American communities than non-native American (Millar, O'Brien *et al.* 2006). However, this is likely linked to socio-economic factors, which also affects the carriage rates.

Pneumococcal colonisation can persist from a few days to several months. Carriage duration is largely influenced by serotype of colonising strains (Sleeman, Griffiths *et al.* 2006, Lipsitch, Abdullahi *et al.* 2012) and the age of host (Turner, Turner *et al.* 2012), with carriage duration ranges from 28 days (type 20) to 124 days (type 6A). When carriage durations overlap, multiple colonisations by more than one strain can be observed.

1.3.2 Interactions between pneumococci and other bacterial species in carriage

1.3.2.1 Interactions between pneumococci

Nasopharyngeal carriage comprises multiple pneumococcal strains as well as multiple bacterial species. A greater sensitivity in detection tools allows multiple colonisation of pneumococci to be estimated (see 1.1.2.1). Microarray and PCR based studies showed that co-colonisation by more than one serotype, ST or a group of distinct evolutionary background is common in carriage. Consistent with carriage prevalence, rates of multiple colonisation appear to be higher in developing than developed countries. Based on the same microarray platform, the rates were reported to be 40% in Malawi (Kamngona, Hinds *et al.* 2014), 30% in Nepal (Kandasamy, Gurung *et al.* 2014) and 17% in UK (Slack, Hinds *et al.* 2014). The presence of multiple colonisation allows pneumococci to evolve through recombination, an important process driving pneumococcal evolution as discussed in 1.2. Hiller *et al.* detected homologous recombination among six isolates collected from a nasopharynx of a child suffering from chromic upper respiratory and middle ear infection (Hiller, Ahmed *et al.* 2010). Given sequence similarity and the time of isolation, the authors showed that one isolate served as a donor for separate recombination events detected in three generations of a clone, generating sequence variations in more than 7% of the genome between the parent, daughter and grand-daughter of this particular clone. This gives an example of interactions between pneumococci in a prolonged chronic infection.

1.3.2.2 Interactions between pneumococcus and other species

Nasopharyngeal habitats harbour a wide range of bacterial species including *Staphylococcus aureus*, *Haemophilus spp.*, *Neisseria spp.*, *Moraxella spp.* and *Streptococcus spp. In vitro* experiments demonstrated that hydrogen peroxide secreted by *S. pneumoniae* could inhibit the growth of *Staphylococcus aureus, Haemophilus influenzae, Neisseria meningitidis,* and *Moraxella catarrhalis* (Pericone, Overweg *et al.* 2000, Regev-Yochay, Trzcinski *et al.* 2006). Moreover, neuraminidase secreted by *S. pneumoniae* could remove sialic acid from the capsule of *H. influenzae* and *N. meningitidis*, thereby reducing protection from host immune system in these two species (Shakhnovich, King *et al.* 2002). On the other hand, *S. pneumoniae* clearance

is promoted by *H. influenzae* which stimulates neutrophil-mediated killing of pneumococci (Lysenko, Ratner *et al.* 2005). This cell killing does not only allow certain species to occupy limited resources, but also releases genomic DNA into the environment, making it available for uptake (discussed in 1.2.1.1 and 1.2.1.2). This availability of exogenous DNA contributes to a large pan-genome - the total numbers of the gene sets of all strains of a species - of most transformable species found in the human nasopharyngeal tract. The pan genome of *N. meningitidis* was predicted to be "open" with 1,337 genes forming the core genome, and the addition of at least 43 new genes from each new *N. meningitidis* genome (Schoen, Blom *et al.* 2008). A plethora of dispensable genes, especially those mediating host-pathogen interactions, were found in *H. influenzae* (Strouts, Power *et al.* 2012). Likewise, the pan-genome of *S. pneumoniae* was shown to be larger than the core genome (Hiller, Janto *et al.* 2007, Donati, Hiller *et al.* 2010, Muzzi and Donati 2011). For each species, variations of specific strains were observed between individual human hosts and likely reflect host specific selection pressure (Human Microbiome Project 2012).

Together, large inter-strain variations in pneumococcal carriage as a consequence of DNA exchanges within the group of pneumococci, and between pneumococci and other bacterial species, have provided a large platform for selection pressure to act on, resulting in the rapid evolution of this species.

1.4 The pneumococci in disease

1.4.1 Morbidity and mortality

Although the pneumococci are generally found as commensals, they are causative agents for several infectious diseases. Disseminations of pneumococci from the nasopharyngeal habitat to other respiratory tract loci or penetration to the sterile body fluid result in a range of mild to severe infections including: sinusitis (to sinuses), conjunctivitis (to conjunctiva), otitis media (to inner ear), pneumonia (to lung and alveoli), bacteraemia (to blood), and meningitis (to central nervous system). It is estimated that approximately 1.2 million children under the age of five died in 2011 from pneumonia with the number of casualties higher in the lower incomes countries where access to care and intervention that improved care are more limited (O'Brien,

Wolfson *et al.* 2009, Izadnegahdar, Cohen *et al.* 2013, WHO 2013). Apart from causing high casualties, the pneumococcus was reported to be an economic burden. Invasive pneumococcal diseases in children were estimated to cost $179-$260 million of the annual national income in Canada, $290-$435 million in Germany, and $277-$432 million in Mexico. Acute otitis media, which is a milder infection and associated with the lowest per-case costs, accounted for 45% - 88% of the national direct medical costs in Canada, Germany and Mexico (Talbird, Taylor *et al.* 2010, Welte, Torres *et al.* 2012)

1.4.2 Bacterial progression from carriage to disease

There is no clear evidence of phylogenetic distinction between carriage and disease strains, suggesting that both the carriage and disease strains evolve together (Donati, Hiller *et al.* 2010, Croucher, Harris *et al.* 2011). Instead, differences in carriage and pathogenic states are marked by strain differential expression profiles (Mahdi, Ogunniyi *et al.* 2008, Ogunniyi, Mahdi *et al.* 2012), some of which result in morphological changes.

Ogunniyi and Mahdi *et al.* applied genome-wide *in vivo* transcriptomic analysis to identify genes up-regulated in different host niches using a mice model. The studies reported 28 genes significantly up-regulated in the lungs relative to those in the nasopharynx, and 25 genes up-regulated in blood in relative to lungs, which reflect the transition between carriage and different pathogenic states. Genes up-regulated at this transition phase include those that function in transport machinery (*aliA*, *cbiO*, *piuA*), amino-acid synthesis (*ilvH*), surface adhesion proteins (*psr*, *cbpAE*), and capsule biosynthesis proteins (*cps4AB*). Mutants of the first four genes were attenuated for virulence in relative to wild type, indicating their roles in pathogenesis (Ogunniyi, Mahdi *et al.* 2012). The up-regulation of capsule biosynthesis proteins observed in these studies also supports the capsule role in disease pathogenesis in forming a physical barrier that limits access of antibodies and complement to the pneumococcal surface (Simell, Auranen *et al.* 2012).

Although direct links between altered gene expression and the phenotypic changes between carriage and pathogenic phase are still lacking, changes in the expression of genes encoding surface proteins might result in phase variation. Phase variation is recognised by a change of colony morphology between opaque and transparent phenotype and known to greatly affect the adherence and virulence of a given pneumococcal strain. The transparent colony morphology has a thinner capsule layer, thought to promote binding to hosts and aid nasopharyngeal colonisation in carriage. The opaque colony displays thicker capsule, which enables better resistance to opsonophagocytic killing in the blood stream (Simell, Auranen *et al.* 2012).

1.4.3 Factors influencing the transformation to diseases

1.4.3.1 Bacterial loads

The bacterial load has been shown to be associated with the disease outcomes. In both adults (Albrich, Madhi *et al.* 2012) and children under the age of five (The PERCH Study Group 2014), nasopharyngeal colonisation density of pneumococci appeared to be higher among patients with pneumococcal pneumonia than asymptomatic colonised controls, suggesting that the transition from asymptomatic carriage to disease may happen at a critical nasopharyngeal colonisation density. Moreover, severity of pneumococcal pneumonia was shown to be associated with bacterial load. As viral load indicates severity in viral infections, higher bacterial loads were shown to be associated with the likelihood of death, the risk of septic shock and the need for mechanical ventilation (Rello, Lisboa *et al.* 2009). However, it is also possible that patients with pneumococcal disease may transmit pneumococci carried in their nasopharynx more easily than those asymptomatic carriers, reflecting in higher bacterial loads as a result host illness (Simell, Auranen *et al.* 2012).

1.4.3.2 Synergism between virus and bacteria

Pneumococcal infection is often secondary to influenza infection, leading to a high mortality in seasonal flu and pandemic. Nearly a third of fatal cases from H1N1

pandemic in 2009 were reported to be bacterial-viral co-infection, most of which were caused by *S. pneumoniae* (Centers for Disease and Prevention 2009). Mathematical models have investigated the influence of influenza infection on invasive pneumococcal diseases while controlling for seasonal factors and bacterial colonisation density (influence of the latter was discussed earlier in 1.4.3.1). In two separate studies, Weinberger *et al.* showed that influenza activity was associated with significant increases in the incidence of invasive pneumonia in both children and adults, suggesting synergistic relationship (Weinberger, Grant *et al.* 2014, Weinberger, Harboe *et al.* 2014). Various mechanisms in which prior viral infection could facilitate subsequent pneumococcal infection have been proposed. First, primary influenza infection could cause epithelial damage, exposing epithelial cells for bacterial entry. Second, viral infection may increase pneumococcal adherence by up-regulating host receptors. Third, it can alter the host immune response, either dysregulate to reduce host defence against bacterial invasion, or amplify the inflammatory cascade (Brundage 2006, McCullers 2006).

1.4.3.3 Capsular types

Several lines of evidence suggest that capsular properties have a strong influence on propensity of invasiveness and lethality of infections. Weinberger *et al.* conducted meta-analysis to differentiate mortality rates in patients infected with different serotypes. The authors showed that strains with thicker capsules, including serotypes 3, 6A, 6B, 9N and 19F are generally associated with high mortality rate in patients (Weinberger, Trzcinski *et al.* 2009, Weinberger, Harboe *et al.* 2010). In mouse experiments, higher mortality rates were observed from infections with strains of thicker capsules than strains of the same serotypes but thinner capsules (Mac and Kraus 1950, Magee and Yother 2001, Bender, Cartee *et al.* 2003). This is likely due to capsule protection against host immune effectors, thereby allowing the bacterium to persist in the lungs and blood. While thick capsules give the bacterium an advantage to persist inside the body, it could obstruct the invasion process. Heavily encapsulated strains may be less likely to cross the epithelieum where direct transcytosis across epithelial cells or the induction of an inflammatory response to disrupt the epithelial barrier is required. Indeed, a separate meta-analysis reported strains with thicker

capsule including serotypes 3, 6A and serogroup 15 as less invasive (Brueggemann, Peto *et al.* 2004). Brueggemann *et al.* noted 60-fold reduction in invasiveness of these strains in comparison to highly invasive strains including serotypes 1, 5 and 7. Given different propensity in invasiveness and lethality, different serotypes may have different disease manifestations. Through another meta-analysis, Grabenstein amd Musey highlighted elevated risk of clinical outcomes in certain serotypes: empyema (serotype 1, 3, 5, 7F, 8, 19A), necrotizing pneumonia (serotype 3), septic shock (serotypes 3, 19A) and meningitis (serotypes 10A, 15B, 19F, 23F) (Grabenstein and Musey 2014).

Apart from factors discussed here, host factors such as very young age, old age, immunodeficiency, low socio-economic status, quality of healthcare, alcoholism and other underlying factors can influence the outcome of diseases (Simell, Auranen *et al.* 2012).

1.4.4 Limited genetic interactions in diseases compared to carriage

Genetic exchange through recombination requires the co-colonisation of donor and recipient strains. However, such scenario may be rare in invasive disease, where a single bacterial cell bottleneck was observed at the origin of infection (Gerlini, Colomba *et al.* 2014) Gerlini *et al.* challenged mice with mixture of pneumococci of three isogenic variants. The authors analysed sequential murine blood samples and revealed that bacteraemia episodes were mostly monoclonal, founded by a single bacteria cell. Given monoclonal infections, genetic interactions observed in carriage previously discussed in 1.3.2 would happen at less frequency in disease state. This suggests that a majority of pneumococcal evolution has taken place during carriage, not in disease. Hence a rationale behind this thesis is to focus on pneumococcal evolution in carriage not in disease state.

1.5 Natural and clinical mechanisms for pneumococcal elimination and how the pneumococcus evolves to evade them

1.5.1 Clearance through natural host immune systems

The human immune system is capable of clearing pneumococcal colonisation but the time required for clearance depends on host age and bacterial serotypes (Sleeman, Griffiths *et al.* 2006, Lipsitch, Abdullahi *et al.* 2012, Turner, Turner *et al.* 2012). Immune responses to colonisation are complex, comprising both innate and adaptive immunities with the latter playing a greater role after the first year of life.

The non-specific innate mechanisms are mediated by phagocytosis and complement system. Phagocytes bind to the bacterium via receptors that recognise a variety of pathogen-associated molecular patterns. This results in ingestion of the microbe into a phagosome, followed by digestion. The pneumococcal polysaccharide capsule inhibits phagocytosis, thereby allowing the bacterium to avoid this defensive mechanism (Hyams, Camberlein *et al.* 2010). The complement system represents an enzyme cascade activated via the binding of complement components to antigen-antibody complexes or directly to the pneumococcal surface. This initiates highly amplified response against the pathogen intrusion, which subsequently leads to an increase in vascular permeability, migration of immune cells to the site of infection, and marking the pathogen for further ingestion and destruction by phagocytes (known as opsonophagocytosis). Pneumococcal surface antigens PspA and PspC were shown to inhibit the complement activation, which elicits the cascade (Tu, Fulgham *et al.* 1999, Dave, Pangburn *et al.* 2004). Moreover, encapsulation decreases the level of the complement-mediated opsonisation, thereby reducing destruction by opsonophagocytosis (Hyams, Camberlein *et al.* 2010). Sequence analyses noted elevated recombination frequency in genomic regions coding for proteins in contact with host innate immunity including capsule biosynthesis locus, *pspA* and *pspC* genes (Croucher, Harris *et al.* 2011). Based on this observation, Croucher *et al.* proposed that sequence divergence introduced by recombination within these loci might provide selective advantage through diversifying selection.

However, non-specific host defences from innate immunity do not always succeed in clearing the pneumococcus. Defensive mechanisms also rely on adaptive immune systems, which are serotype/antigen specific. Colonisation-related serum immunoglobulin G (IgG) antibody showed response to capsular polysaccharide and several surface and virulence proteins. For some serotypes, there is evidence that naturally acquired anti-capsular antibody can protect against carriage of specific serotypes (Goldblatt, Hussain *et al.* 2005, Weinberger, Dagan *et al.* 2008). The antibody-independent mechanism is mediated by CD4+ T cell response, which acts via interleukin 17A secretion and neutrophil recruitment to enhance clearance of colonisation in older (more immune) hosts (Cohen, Khandavilli *et al.* 2011). Sequence diversity of some surface proteins that elicit specific responses have been described (Bergmann and Hammerschmidt 2006), reflecting ways in which pneumococci can further avoid detection and clearance by host adaptive immune systems.

1.5.2 Clinical interventions

Clinical interventions including vaccines and antibiotics have been respectively employed to reduce the rates of nasopharyngeal colonisations and stop the infections should the carriage progress into diseases. While these two interventions were shown to be successful in reducing the number of invasive diseases, vaccine escape serotypes and a rise in antibiotic resistance have been increasingly reported (WHO 2014). Together, these raise a concern over strategies to combat pneumococcal diseases.

1.5.2.1 Vaccines

1.5.2.1.1 Vaccine development

George Sternberg, one of the two researchers who first isolated the pneumococcus in 1881 (see 1.1.1) demonstrated that the rabbits inoculated with dead pneumococcus were immunised against the bacterium in subsequent injections (Sternberg 1882). This protection in animals set the foundations for the development of vaccine in humans. In the early 1990s, the first pneumococcal vaccine comprising dead pneumococci of unknown identity, was administered to 50,000 workers in South

African gold miners where prevalence of pneumococcal pneumonia were high (Austrian 1978). The vaccine was shown to reduce cases of pneumococcal diseases in four months post vaccination. However, the protection was lost over time (Wright 1914). There are several explanations for this loss in efficacy. One hypothesis refers to an increase in infections caused by non-vaccinated types, but it was difficult to trace the answer given limited knowledge of antigenic structure and variation at that time.

Human immune protection is not only triggered by the exposure to the whole pneumococcal cell, but also shows response to capsular polysaccharide and surface proteins (see 1.5.1). In the 1940s, Macleod *et al.* first showed that purified capsular polysaccharides could be used as active immunogens in adult humans with some protection efficacy (Macleod, Hodges *et al.* 1945). However, vaccination with capsular polysaccharides was not a popular option due to the availability of antibiotics around the same time (will be discussed in the next section). It was not until the rise of antibiotic resistant pneumococci in 1960s that prophylactic control through capsular polysaccharide vaccination was reconsidered. A 14-type mixture of the most prevalent serotypes were selected for vaccination in 1977 and increased to 23 types in 1983 (Robbins, Austrian *et al.* 1983). While the capsular polysaccharide vaccine was shown to be effective in older children (> 5 year olds) and adults, it appeared ineffective in infants due to poor immune response at the extreme age (Douglas, Paton *et al.* 1983). One way to elicit stronger capsular polysaccharide–reactive antibody response was to couple multiple capsular polysaccharides or their immune-determinant sugars to carrier proteins (Avery and Goebel 1929). This subsequently led to the development of conjugate capsular polysaccharide vaccines to protect infants. Indeed, the conjugate vaccine was shown to be first effective in infants protecting them against the capsulated *Haemophilus influenzae* type b (Schneerson, Barrera *et al.* 1980). *In S. pneumoniae*, the conjugate vaccines were also effective in preventing diseases in infants (Black, Shinefield *et al.* 2000). As the technology limits the number of serotypes that can be included, the vaccines have thus focused on serotypes that are more common, associate with invasive diseases or are highly resistant to antibiotics: the **7**-valent **P**neumococcal **C**onjugate **V**accine (PCV7) comprises serotypes 4, 6B, 9V, 18C, 19F, 23F; and the **13**-valent **P**neumococcal

Conjugate **V**accine (PCV13), developed later, contains the seven serotypes included in PCV7 plus additional serotypes 1, 3, 5, 6A, 7F, and 19A.

1.5.2.1.2 Serotype replacement and vaccine escape pneumococci

Although, PCVs have reduced the rates of nasopharyngeal colonisation and invasive diseases caused by the targeted serotypes, it has been shown that vaccine serotype can switch to, or be replaced by non-vaccine serotypes in the vaccine-induced community. This can lead to reduction of the overall efficacy of these vaccines.

An increase in prevalence of strains not covered by vaccine was shown to be a problem in some populations. Levy *et al.* reported unchanged prevalence of pneumococcal meningitis following the implementation of PCV7 in France (Levy, Varon *et al.* 2011). The authors noted an increased prevalence of infection by non-vaccine serotypes. It may be possible to include additional non-vaccine serotypes in conjugate mixtures to expand the protection. However, this may have transient value due to pneumococcal recombinogenic behaviours (see 1.2).

The pneumococci in their natural habitats have been shown to exchange their genetic contents rapidly through recombination (Croucher, Harris *et al.* 2011). One of the more frequently exchanged genomic regions is the capsule biosynthesis locus. This allows capsular biosynthesis genes of one serotype to be replaced by genes of different capsule types, leading to capsular switch. The switch can be within the same serogroup or between different serogroups. Two separate studies showed that recombination between capsular loci of vaccine-targeted serotype 6B and the serotype 6C strain had resulted in the novel serotype 6D strain, which is beyond the target of current conjugate vaccines (Bratcher, Park *et al.* 2011, Otsuka, Chang *et al.* 2013). In the USA, there has been a rise of serotype 19A pneumococci in carriage and, concomitantly, in disease following PCV7 introduction (Yildirim, Hanage *et al.* 2010). Sequence analyses showed that serotype 19A isolates had emerged from capsular switches in multiple lineages (Brueggemann, Pai *et al.* 2007, Croucher, Harris *et al.* 2011, Golubchik, Brueggemann *et al.* 2012). In one lineage called PMEN1, Croucher *et al.* identified recombination events at the capsule locus, where

an original 23F capsular locus was switched to type 19A, 19F, 3, 6A, 15A and 14. A switch to non-vaccine type 19A around 1992-1999 has facilitated the expansion of a 19A capsular type clone, replacing other PMEN1 serotypes following vaccine introduction into the USA in 2000 (Croucher, Harris *et al.* 2011). Capsular switch is not uncommon in the natural habitats. Wyres *et al.* investigated the frequency of capsular switch observed in 426 pneumococci collected from 1937 through 2007 (Wyres, Lambertsen *et al.* 2013). The authors reported 36 independent capsular switch events, and demonstrated that in some cases, the exchange may extend beyond the capsular locus to the nearby penicillin-binding protein genes *pbp2x* and *pbp1a*. This does not alter only the bacterial capsular antigenic properties, thereby allowing the vaccine escape; but also their antibiotic resistance profiles, which is a burden for invasive disease treatments.

1.5.2.2 Antibiotics

The antibiotic era began in 1928 with Alexander Fleming's observation that bacteria would not grow near colonies of the *Penicillium* mould. This discovery led to the development of the antibiotic penicillin, which has broad-spectrum activity and was effective against many serious infections caused by staphylococci and streptococci. Soon after penicillin, different classes of antibiotics including streptomycin (1943), tetracycline (1944), chloramphenicol (1946), erythromycin (1948), vancomycin (1953), rifampicin (1957), ciprofloxacin (1961), streptogramin B (1963) were developed and successfully used for treatment of various bacterial diseases such as tuberculosis and pneumonia (Lewis 2013). Together, they drastically reduced death rates associated with many infectious diseases.

However, Fleming already warned that bacteria could become resistant to the antibiotics in his Nobel prize speech in 1945 for his penicillin discovery. He claimed that it was not difficult to make microbes resistant to penicillin in the laboratory by exposing them to concentrations not sufficient to kill them, and the same had occasionally happened in the body (Fleming 1945). Indeed, resistance to many antibiotics were observed within a decade or less following the introduction of antibiotics including resistance to penicillin (1945), streptomycin (1946), tetracycline

(1950), chloramphenicol (1950), erythromycin (1955), vancomycin (1960), rifampicin (1962), ciprofloxacin (1968), streptogramin B (1966) (Lewis 2013). In *S. pneumoniae*, multidrug resistance was first observed in 1977 and has been widely spread since (Whitney, Farley *et al.* 2000). Rapid development of antibiotic resistance in *S. pneumoniae* has been a global concern and a growing numbers of reports have shown that bacterial pneumonia may not respond to available antibiotics in many settings (WHO 2014).

An association between recombination and antibiotic resistance was described by (Hanage, Fraser *et al.* 2009), where hyper-recombinant populations were significantly associated with resistance to penicillin, erythromycin, tetracycline, chloramphenicol and cefatoxime. In the case of penicillin resistance, which is mediated by penicillin binding proteins, resistant mosaic *pbp1a*, *pbp2b*, and *pbp2x* genes were shown to have developed in several different lineages and species before being acquired by *S. pneumoniae* through homologous recombination (Chi, Nolte *et al.* 2007, Hakenbeck, Bruckner *et al.* 2012). Together, this highlights a role of recombination in facilitating pneumococcal survival upon exposure to different clinical interventions.

1.5 Project aims and objectives

The overall aim of this project is to investigate different aspects of pneumococcal evolution during carriage, the state which is the prerequisite for development of invasive pneumococcal disease and also the phase that shapes the wider population structure.

The project takes advantage of a large dataset of whole genome sequencing data from a large longitudinal carriage cohort study conducted in the Maela refugee camp, which is located on the Thailand-Myanmar border during 2007-2010 (Turner, Turner *et al.* 2012). The Maela collection and its settings will be described fully in Material and Methods.

The first results chapter explores the pneumococcal population structure detected in Maela refugee camp based on whole genome sequence data. It also compares the

prevalence of different lineages detected in Maela to contemporaneous pneumococcal population detected at other geographical regions through MLST data. This comparative information allows one to predict how much observations made in the next following chapters, which exclusively focuses on Maela community, are applicable to elsewhere.

The second results chapter estimates evolutionary parameters, such as nucleotide substitution and recombination, and compares them between lineages. Genetic interactions through homologous recombination in Maela pneumococci are investigated. This chapter highlights a higher rate of both acceptance and donation of recombinant DNA in nontypable isolates, and proposes its role as a hub of genetic exchanges in Maela pneumococcal population.

The third results chapter explores the content of recombining genes described in the second results chapter, most of which were associated with antibiotic resistance and surface antigens. With the availability of clinical data on antibiotic consumption, predicted selection pressures, likely selected alleles, and the spread of these alleles through homologous recombination can be linked together.

The fourth results chapter identifies genetic determinants of resistance to beta-lactam antibiotic through a genome-wide association studies. The method is frequently used in human genetics but has been largely untried in bacteria due to the intrinsic clonal structure. The chapter discusses how this limitation might be less problematic in a highly recombinogenic bacteria like *S. pneumoniae* as well as documents genetic variations which might alter to beta-lactam non-susceptibility in pneumococci.

Together, I hope this thesis will make a small contribution towards better understanding of pneumococcal evolution and rapid development of antibiotic resistance observed in this species. Thank you so much and enjoy the thesis.

Chapter 2: Materials and methods

2. Materials and methods

Materials and methods documented in this chapter are commonly used in all of the following results chapters. Detailed methods specific to particular parts of the analyses will be described in the relevant chapter.

2.1 Pneumococcal collections

Genotypes and phenotypes of *Streptococcus pneumoniae* used in this thesis comprise the densely sampled carriage collection from Maela community and data from elsewhere to give a comparative view for comparison.

2.1.1 Maela whole genome sequencing collection

2.1.1.1 overviews

The Maela cohort study, the main subject of this thesis, was conducted by Drs Claudia and Paul Turner with the original aims of investigating the natural history and outcome of *S. pneumoniae* colonisation in infancy in the presence and absence of simultaneous colonisation with other bacteria and viruses. The study was based in Maela refugee camp, which is located in North West Thailand adjacent to the Myanmarese border. It is a densely populated area of 46,133 registered refugees (mostly of Karen ethnic group) with 12% of the population under the age of 5 years (The Border Consortium 2012). In 2005, infant mortality was estimated to be 21/1000 live births (similar to Thailand) and approximately 14% of deaths in children aged less than five years were due to pneumonia.

2.1.1.2 study population

The Maela cohort study included specimens collected from infants born at Maela camp as well as a quarter of their mothers. Pregnant women were recruited in the camp antenatal clinic and subsequently randomised into the routine follow-up cohort (n=750) or the immunology follow-up cohort (n=250). Monthly nasopharyngeal swabs were taken from infants in both routine and immunology cohorts from birth and terminated at 24 months. The immunology cohort also included additional

sampling for maternal colonisation. Overall, 8,386 swabs were taken and were part of the study described in (Turner, Turner *et al.* 2012).

2.1.1.3 Laboratory aspects

The swabs were collected and processed according to the WHO pneumococcal carriage detection protocol (O'Brien and Nohynek 2003). All isolates were serotyped and then tested for antibiotic susceptibilities. Serotyping was performed using latex-agglutination and Quellung reaction, while penicillin and co-trimoxazole susceptibilities were determined by disk diffusion following current **C**linical and **L**aboratory **S**tandard **I**nstitute (CLSI) guidelines. Microbiology work was done entirely at Shoklo Malaria Research Unit (SMRU) microbiology laboratory in Maesot, Thailand.

2.1.1.4 Sequenced isolates

A collection of over 3,000 single-colony isolates was randomly selected for whole genome sequencing in such a way that at least 100 isolates were recovered from each of the 30 consecutive months of studied period. DNA extraction for each isolate was performed using a RBC Bioscience MagCore HF16 platform. Our collection is tabulated in **Appendix A**.

2.1.2 PMEN14 whole genome sequencing collection

A collection of PMEN14 isolates (CC320) was also included in the study. The lineage represents a highly successful clone predominantly detected in South East Asia. The collection is fully tabulated in **Appendix B**, comprising 175 isolates, 40 of which were isolated from invasive disease. The collection came from multiple geographical regions including the Middle East, Europe, USA and other South East Asian countries between 1997 – 2009; thereby providing a view on multiple introductions of this particular global clone into a local community like Maela in chapter 3.

2.1.3 Other global MLST collections

Large pneumococcal carriage cohorts conducted elsewhere were included either to enable comparison of population structure from ranges of different geographical areas between 2006-2010 (chapter 3); or for cross-validation in the genome-wide association studies (chapter 6). Only studies with sampling size larger than 100 isolates were chosen to allow robust statistical analyse.

Table 2.1 Other pneumococcal carriage collections used in the studies

Type of study	Locations	Sample size (no. isolates)	Sampling	References
Comparison of global pneumococcal population structure from infants and children (<7 years) between 2006-2010 by MLST (chapter 3)	Massachusetts, USA	280	2007	(Croucher, Finkelstein *et al.* 2013)
	Southampton, UK	310	2006 - 2009	(Tocheva, Jefferies *et al.* 2013)
	Kilifi, Kenya	316	2006 - 2008	(Brueggemann, Muroki *et al.* 2013)
	The Gambia	445	2010	Unpublished*
	Maela, Thailand	2,492	2007-2010	Described in this thesis
Genome-wide association study (chapter 6)	Massachusetts, USA	616	2001-2007	(Croucher, Finkelstein *et al.* 2013)
	Maela, Thailand	3,085	2007-2010	Described in this thesis

*With kind permission from Dr Sarah Burr, Medical Research Council Unit, The Gambia for the data to be used in this thesis.

A global collection of PMEN14 (Taiwan^{19F}-14) used in Chapter 3 was fully list in (Croucher, Chewapreecha *et al.* 2014) as supplementary table 1.

2.2 Whole-genome sequencing

All processing and sequencing of genomic DNA for Maela pneumococcal collection was performed by the Wellcome Trust Sanger Institute's core sequencing teams. Illumina sequencing approach was employed to generate data used in this thesis. see (Bentley, Balasubramanian *et al.* 2008) for review, All samples were sequenced as multiplexed libraries using the Illumina HiSeq 2000 analyzers on 75bp paired end runs as described in (Croucher, Harris *et al.* 2011) giving a mean coverage of 276.67 reads per nucleotide. Short reads from this study have been deposited in the European Nucleotide Archive under study numbers ERP000435, ERP000483, ERP000485, ERP000487, ERP000598 and ERP000599.

2.3 Control for sample mix up through determination of serotype and sequence type

Serotype and MLST were derived from Illumina read data as described in (Croucher, Harris *et al.* 2011). Here, any short reads were mapped against pneumococcal *cps* loci reference sequences (Bentley, Aanensen *et al.* 2006) and alternative genes detected between *dexB* and *aliA* (Salter, Hinds *et al.* 2012). Any reference locus with the highest proportion of its length covered by mapped sequence reads was likely to encode the capsule. These *in silico* derived serotypes were compared to serotypes detected from latex-agglutination and Quellung reaction for quality control purposes. Any incompatibility between *in silico* and serologically derived serotypes were re-investigated. This allowed processing and human errors including label swaps and potential sample contaminations to be detected.

Of 3,157 isolates initially sequenced, any suspected contaminations were removed from the data, leaving 3,085 whole genome sequences used in this thesis.

For nontypable (NT) serotypes, methods as explained in Salter *et al.* (Salter, Hinds *et al.* 2012) were used to confirm the absence of a capsule locus or the presence of alternative genes detected between *dexB* and *aliA*. Diversity of the NT category detected in this population is discussed in 3.2.2.3.

2.4 Sequence assembly

De novo assembly was performed through a pipeline developed by the Wellcome Trust Sanger Institute's Pathogen Bioinformatics team (Dr Andrew Page). 3,085 strains were *de novo* assembled multiple times using Velvet (Zerbino and Birney 2008), where the kmer size was varied between 60% and 90% of the read length. The assembly with the best N50 was chosen. Contigs shorter than the insert size length were filtered out because they are most likely misassemblies. The sequencing data were then used to improve further the assembly. The contigs were iteratively scaffolded and extended 16 times using SSPACE (Boetzer, Henkel *et al.* 2011) beginning with the contigs where the greatest number of reads overlap. Gaps, denoted by 1 or more N's were targeted for closure by running 120 iterations of GapFiller (Boetzer and Pirovano 2012), cycling between BWA (Li and Durbin 2009) and Bowtie (Langmead, Trapnell *et al.* 2009), beginning where the greatest number of reads overlapped. A final QC step was performed on each assembly, with the reads mapped back to the assembly using SMALT 0.5.7. The assembly pipeline gave, on average, a total length of 2,161,240 bp from 111.279 contigs with average contig length of 33,191.4 bp and average N50 of 65,656.6. Where appropriate reference genomes were not available in the public databases, reference genomes (with contigs ordered and annotated) were created from *de novo* assembly. The assembly was created as described above and ordered relative to its closest references using ABACAS v2.5.1 (Assefa, Keane *et al.* 2009) and ACT (Carver, Rutherford *et al.* 2005). Annotations were directly transferred from *S. pneumoniae* ATCC 700669 followed by manual curation. These novel references were required for mapping and will be discussed in the next part.

2.5 Sequence mapping

Mapping was used in different parts of the analyses described in this thesis (see table 2.2 for a summary). Short reads from samples were mapped onto different reference genomes using SMALT 0.5.7. Bases were called and aligned using the method described in (Harris, Feil *et al.* 2010). In brief, reads aligning to the reference with a quality score of greater than 30 were considered. For each position, a base was only called if the Phred score (Ewing, Hillier *et al.* 1998) exceeds 50. This theoretically gives an accuracy of 99.999%. The call had to be supported by at least 4 reads, with at

least two on each strand. Any calls that failed the criteria were reported as unknown with character “N”.

For lineage-specific analysis (chapter 4), the final alignments also include short insertions and deletions (indels) using the pipeline developed by Dr Simon R. Harris.

Table 2.2 References used for mapping and mapping coverage generated for each dominant cluster

strain	reference	Accession number	serotype	ST	Use in the study	Genome size (bp)	% mapping
Spanish23F (ATCC700669)	Public database	FM211187	23F	81	Map against the Maela and Massachusetts collection to determine coarse population structure and capture variants (chapter 3, 4 and 6)	2221315	82.33
Taiwan19F-14	Public database	CP000921	19F	236	BC1-19F reference chapter 4	2112148	96.79
INV200	Public database	FQ312029	14	9	BC3-NT reference chapter 4	2093317	91.42
G54	Public database	CP001015	19F	63	BC7-14 reference chapter 4	2078953	96.22
SMRU 1949	Draft genome*	ERR057930	23F	802	BC2-23F reference chapter 4	1935768	96.53
SMRU 2513	Draft genome*	ERR064018	6B	315	BC4-6B reference chapter 4	1991123	95.92
SMRU 1861	Draft genome*	ERR057842	23F	2218	BC5-23A/F reference chapter 4	1896242	94.91
SMRU 1478	Draft genome*	ERR054427	15C	4209	BC6-15B/C reference chapter 4	1933435	96.73

Draft genome* in Table 2.2 indicates references generated from draft genome assemblies.

2.6 Visualisation of phylogenetic trees

Display and manipulation of phylogenetic trees was performed using the online tool Interactive Tree of Life (Letunic and Bork 2011) and the software package Circos (Krzywinski, Schein *et al.* 2009).

2.7 Statistical analyses

All statistical tests and associated diagrams were generated in R version 2.11.1(R Core Team 2014). Statistical analyses were discussed in relevant sections in the text.

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

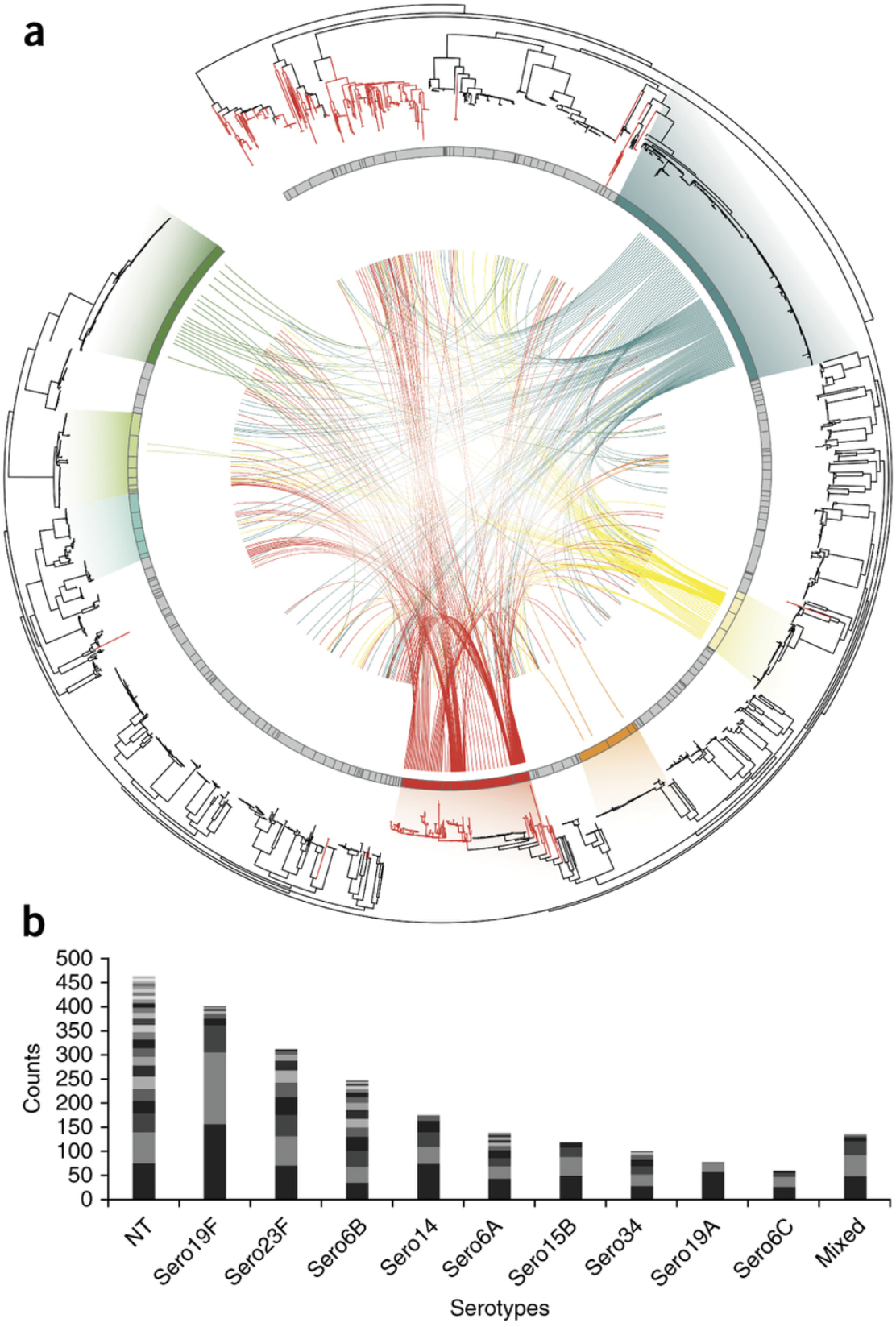
(highlighted in red in **Figure 3.1 a**). This observation could indicate that some lineages could have a higher evolutionary rates compared to others. Alternatively, this may suggest that the close clusters represent successful lineages that have undergone recent clonal expansion, whereas the distantly-related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

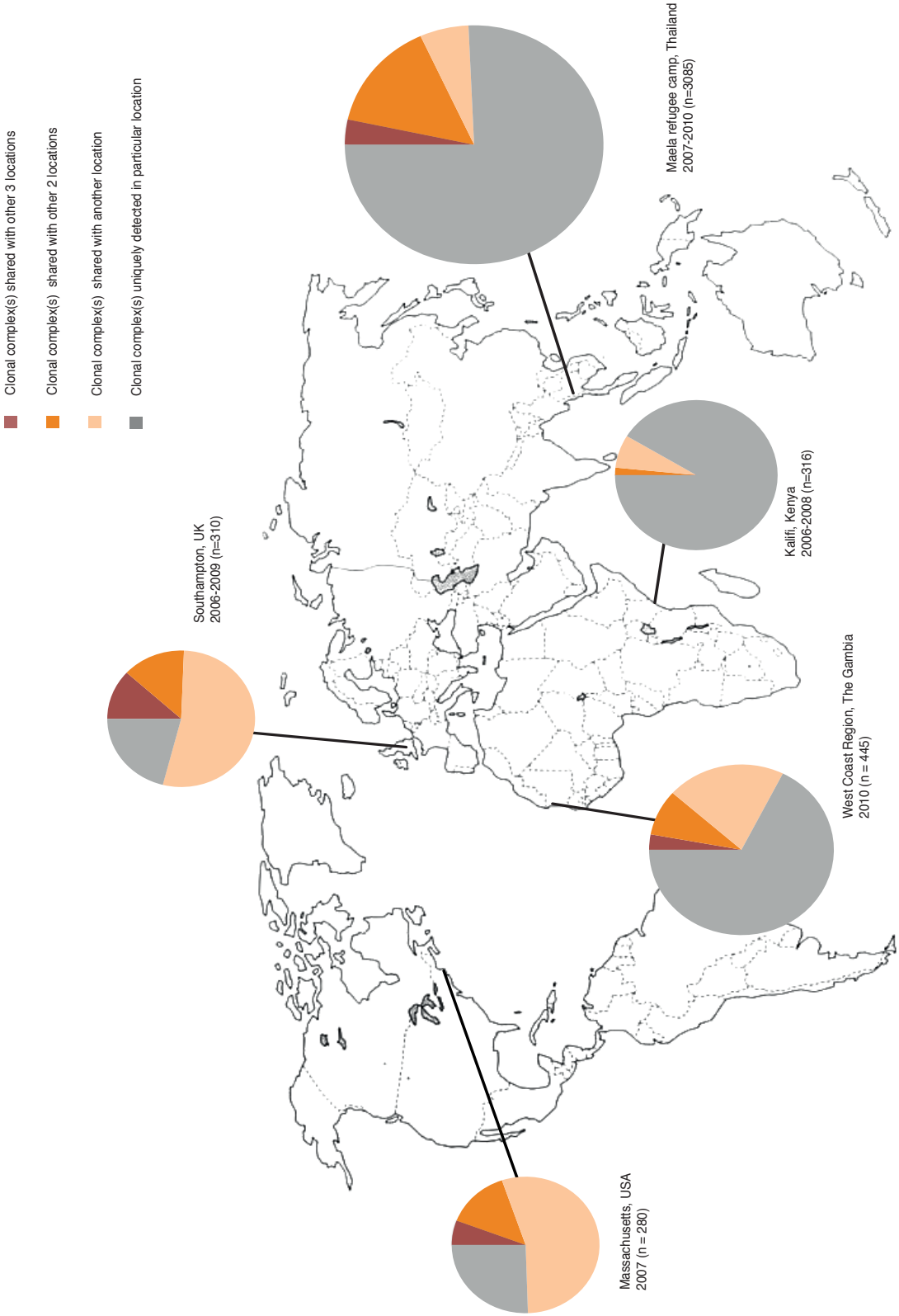
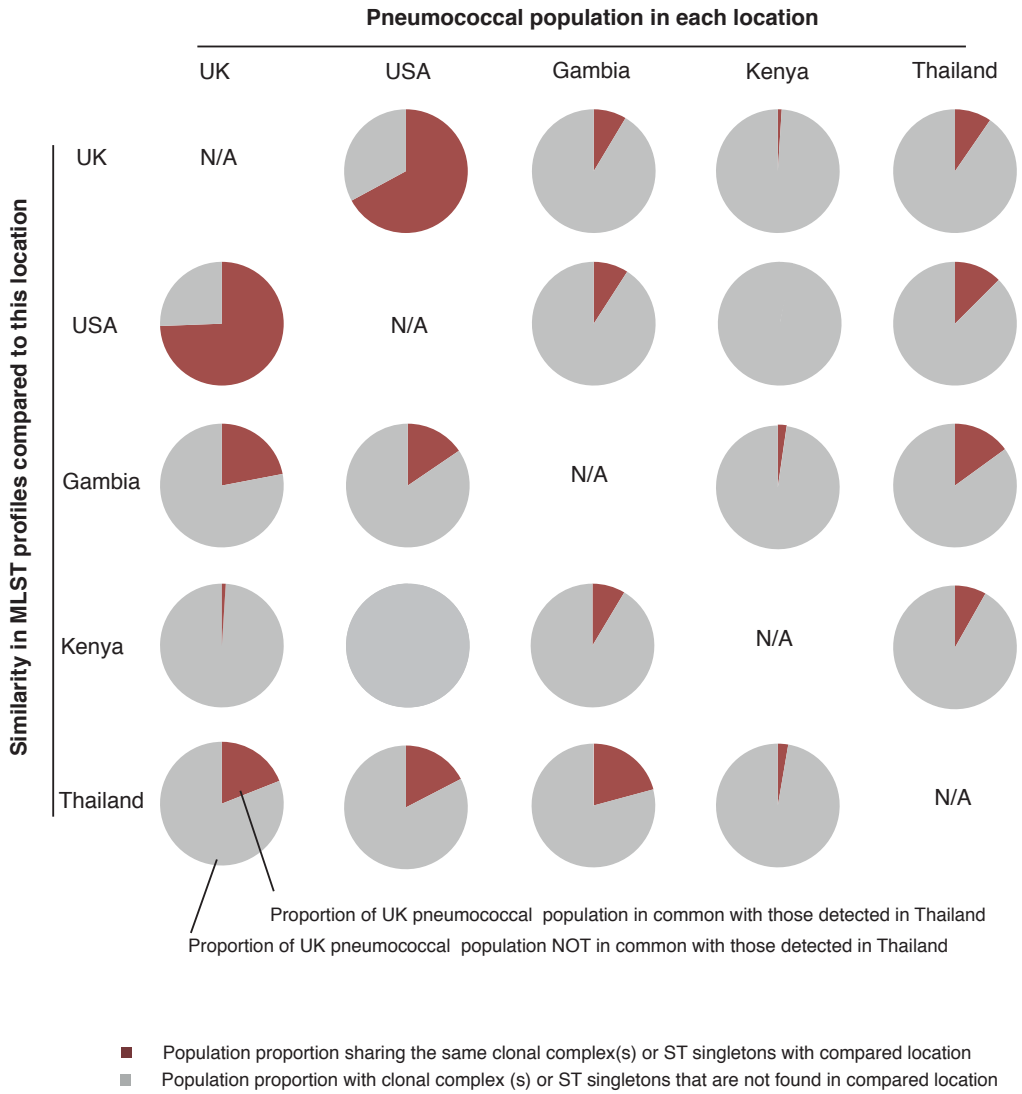

Clonal complex(s) shared with other 3 locations
Clonal complex(s) shared with other 2 locations
Clonal complex(s) shared with another location
Clonal complex(s) uniquely detected in particular location
Maela refugee camp, Thailand
2007-2010 (n=3085)
Kalifi, Kenya
2006-2008 (n=316)
Southampton, UK
2006-2009 (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

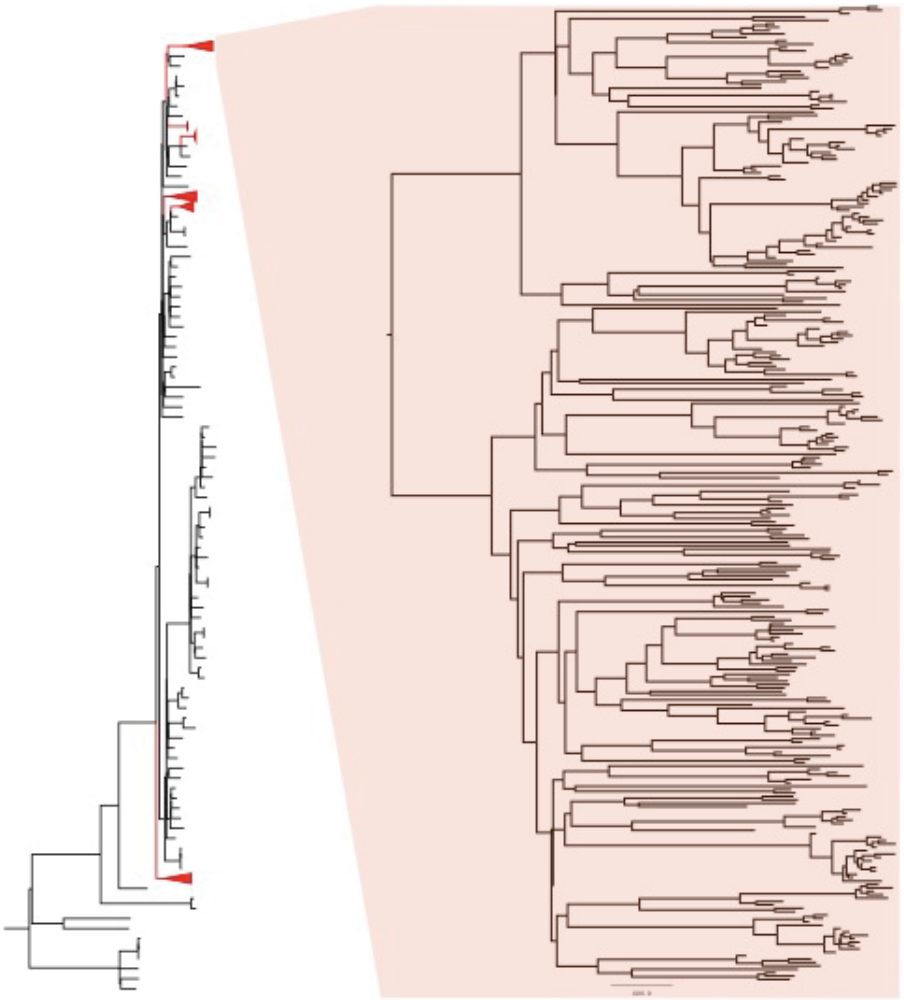
Myanmarese students coming to the camp for their education (The Border Consortium 2012). These regular visits from outside may explain the multiple introductions of a globally circulating lineage into this remote community.

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.

Chapter 4: Maela pneumococcal evolution and population-wide sequence exchange

---

Declaration of work contributions:

Analysis using BratNextGen discussed in 4.2.1.2 was performed by Dr Pekka Marttinen. Unless stated here, I was responsible for the analyses.

4. Maela pneumococcal evolution and population-wide sequence exchange

4.1 Introduction and aims

Understanding how pneumococcal populations evolve in carriage is essential for prevention and control of pneumococcal diseases. Prior to this study, evolutionary parameters, including rates of nucleotide substitution and homologous recombination, have been captured in a multidrug resistant lineage called PMEN-1 (Croucher, Harris *et al.* 2011). Diversity brought into the PMEN-1 genome by horizontal gene transfer, including homologous recombination, is frequently associated with genes conferring antibiotic resistance and antigenic variation. This introduction of genetic variation was proposed as a key contributor to the success of this lineage in evading clinical interventions and spreading globally. However, one can see from last chapter that the globally spread clones only make up a small proportion of the whole pneumococcal population at each locality, suggesting that observations from a single clone are unlikely to represent the whole pneumococcal community. Less was known about how other lineages in the population evolve and interact with each other.

This chapter explores the evolution and genetic interactions in a species-wide sampling of a pneumococcal population, from the densely sampled Maela collection.

This chapter aimed at:

i) Estimating lineage-specific evolutionary parameters including nucleotide substitution and recombination.

ii) Identifying the sources and sinks of recent recombination events.

4.2 Methods

4.2.1 Estimating lineage-specific evolutionary parameters

4.2.1.1 Criteria for selection of dominant clusters

Major lineages representing the whole population were selected for estimating and comparing lineage-specific evolutionary parameters. As currently available tools for detecting recombination are designed for a single lineage population rather multiple lineages (Croucher, Harris *et al.* 2011), the major clusters were chosen in such a way that the genetic diversity within the lineage must not exceed the limitations of recombination detection tool and the sample size of each cluster are is large enough to enable provide robust statistical power. Limitations of the recombination detection method was fully discussed in Croucher *et al.* 2013 (Croucher, Finkelstein *et al.* 2013). Out of eleven large primary clusters that comprised more than 100 isolates, seven primary clusters appear to have members that either share the same serotype/serogroup or differ by a few MLST locus variants, suggesting that isolates within these clusters are not too distant to exceed the tool limitations. Also, a phylogeny represents members of these main clusters on monophyletic branches. As a result, BC1-19F (n = 365 isolates), BC2-23F (n = 213 isolates), BC3-NT (n = 202 isolates), BC4-6B (n = 126 isolates), BC5-23A/F (n = 106 isolates), BC6-15B/C (n = 102 isolates) and BC7-14 (n = 102 isolates) (See **Appendix A**) were selected.

4.2.1.2 Recombination

Recombination fragments were identified using a phylogenetic-based algorithm as explained in (Croucher, Harris *et al.* 2011) and BratNextGen software (Marttinen, Hanage *et al.* 2012). Brief summaries of both algorithms are given below.

4.2.1.2.1 Croucher *et al.* method

Recombination fragments were identified using phylogenetic-based algorithms as explained in Croucher *et al.* 2011 (Croucher, Harris *et al.* 2011). In brief, the method

reconstructed a phylogeny based on the original alignments using RAxML v7.0.4 (Stamatakis 2006) followed by a reconstruction of nucleotide polymorphic sites on the phylogeny using PAML (Yang 2007). Regions on each phylogenetic branch where SNPs occur in clusters of significantly higher density than expected by chance (recombination events) were called and iteratively removed from the alignment. Remaining variants were then used to reconstruct the phylogeny, and recombination fragment detection repeated, iteratively, until there were no further changes in the tree topology. Recombination fragments were predicted from the SNPs dense regions in the whole genome alignment and were used to estimate lineage-specific rates of recombination.

4.2.1.2.2 BratNextGen method

In addition to above method, BratNextGen (**B**ayesian **R**ecombin**a**tion **T**racker) software (Marttinen, Hanage *et al.* 2012) was employed to identify recombination fragments. BratNextGen analysis was performed by Dr Pekka Marttinen, generating the predictions used in combination with results from Croucher *et al.* method described above. In brief, BratNextGen is a derivative of BAPS; the software exploits the data from different ancestral sources in each isolate predicted by BAPS (regarded as admixture events in the population) and expands the admixture model further by probabilistically characterising the origin of any particular site. This allows sites of recombination fragments with distinct origins to be identified, differentiating them from the rest of the genome. The software was applied on the whole genome alignment with the same default setting as in (Marttinen, Hanage *et al.* 2012). Significance of each putative recombinant segment (p-value <0.05) was determined through a bootstrap test with 100 replicates. As shown in (Marttinen, Hanage *et al.* 2012), the approach yielded highly similar recombination with the analysis of PMEN1 data (Croucher, Harris *et al.* 2011). Results generated from BratNextGen were thus used for cross-validation against results produced by Croucher *et al.* algorithm. Comparison between results generated by BratNextGen and Croucher *et al*. method will be discussed in this chapter.

4.2.1.2.3 Comparing level of recombination between different clusters

Following an identification of recombination, recombination per nucleotide substitution ratio (r/m) was calculated given the number of polymorphic sites produced by Croucher *et al.* algorithm, excluding any signals localised in the mobile genetic elements. The ratio r/m was calculated using two different approaches.

The first method modelled the relationship between recombination events and mutations as a linear regression under non-parametric distribution. Ranked recombination events and ranked number of SNPs were used as the outcome (y-axis) and the predictor variable (x-axis) respectively, with the slope representing r/m. ANCOVA tests were used to determine the significant difference in recombination rates between different clusters when statistical assumptions were met.

The second method used the arithmetic mean of r/m for each cluster, representing an average r/m from each branch within the cluster. The Kruskal-Wallis test was used to test for significant differences in r/m calculated by arithmetic mean. A consistency between two methods will give greater confidence over the results.

4.2.1.3 Mutation

4.2.1.3.1 BEAST

Mutation SNPs were separated from recombination SNPs using method described previously above (Croucher, Harris *et al.* 2011). However, there was difficulty in correlating the overall accumulation of SNPs through time from the whole cluster owing to a narrow sampling time frame. Therefore, correlations were performed within subclades of dominant clusters instead of using the whole cluster to capture the signals. Temporal signals and clock-like-ness in each subclade of the phylogenies were screened with Path-O-Gen v1.3 (Path-O-Gen 2010). Where positive correlations were observed (p-value <0.05), the mutation rates were then calculated with BEAST (**B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees) (Drummond and Rambaut 2007) using the skyline population size prior and a relaxed lognormal clock model.

4.2.1.3.2 Comparing the rates between different clusters

The Kruskal-Wallis test was used to test for significant differences in mutation rates estimated from different clusters.

4.2.2 Tracing genetic exchanges through homologous recombination

Potential donor isolates for homologous recombination were identified from recombinant fragments detected in the recipients. Recombination fragments co-predicted by the methods described above were used as query sequences for nucleotide blast searches for potential donor blocks from 3,085 draft assembled genomes. Using blastall v 2.2.15 (Altschul, Gish *et al.* 1990), hits that have an exact match to the query sequence were likely to be potential donors. Several criteria and filters were applied onto these analyses to reduce detection of false positives. The rationale for the criteria will be discussed in full in the chapter.

Probabilities of a single isolate, as well as of each BAPS cluster acting as a donor for a recipient were then calculated. For each recipient isolate, “n” potential donor isolates were identified, and each donor isolate was assigned a probability of “1/n” of having been the donor. Isolates showing no particular hit for a particular search were given a probability of 0. The total frequency of each isolate in being the donor was represented by the sum of the above probabilities from all potential donation events. Isolates were grouped into lineages based on their population cluster from BAPS. The relationship between cluster size, cluster diversity and probability of being a donor were estimated using Spearmann’s ranking correlation.

4.3 Results

4.3.1 Estimating evolutionary rates within the population

Two evolutionary parameters, nucleotide substitution and homologous recombination were investigated here. This section documents how the two evolutionary signals were separated, identifying comparable nucleotide substitution yet heterogeneous recombination rates within the pneumococcal population and the potential biological implications.

4.3.1.1 Separating recombination signals from single nucleotide substitutions

To estimate the evolutionary rates, a phylogenetic network would be required to calibrate observed genetic changes with time. However, a conventional phylogeny typically assumes a single history from the whole genome, which is not true when recombination events have occurred. Different regions in the alignment affected by recombination may have different underlying phylogenies, resulting in an abrupt change in tree topology (Posada and Crandall 2002, Ruths and Nakhleh 2005). Here, signals from recombination were distinguished from nucleotide substitutions using the method described in Croucher *et al.* (Croucher, Harris *et al.* 2011). Recombination is generally identified as SNP dense regions in the genome. The algorithm searches the whole genome for regions with high SNP density and iteratively removes these suspected recombined regions until there is no further change in tree topology.

The method has been successfully applied to a single lineage study of 240 PMEN-1 isolates (Croucher, Harris *et al.* 2011). However, it was not feasible to run the algorithm on the entire species-wide dataset of 3,085 genomes. A preliminary test on a smaller species-wide pneumococcal dataset of 127 isolates from Malawi (Everett, Cornick *et al.* 2012) showed that the method failed after reaching a high diversity threshold. A starting tree prior to the removal of recombination displayed a typical long-branch phylogeny, which is a result of the combined SNPs from both nucleotide substitutions and recombination. The removal of recombination in this diverse genetic background proved to be difficult. With each lineage having different evolutionary history, their recombination patterns are distinct in each isolate. This resulted in a

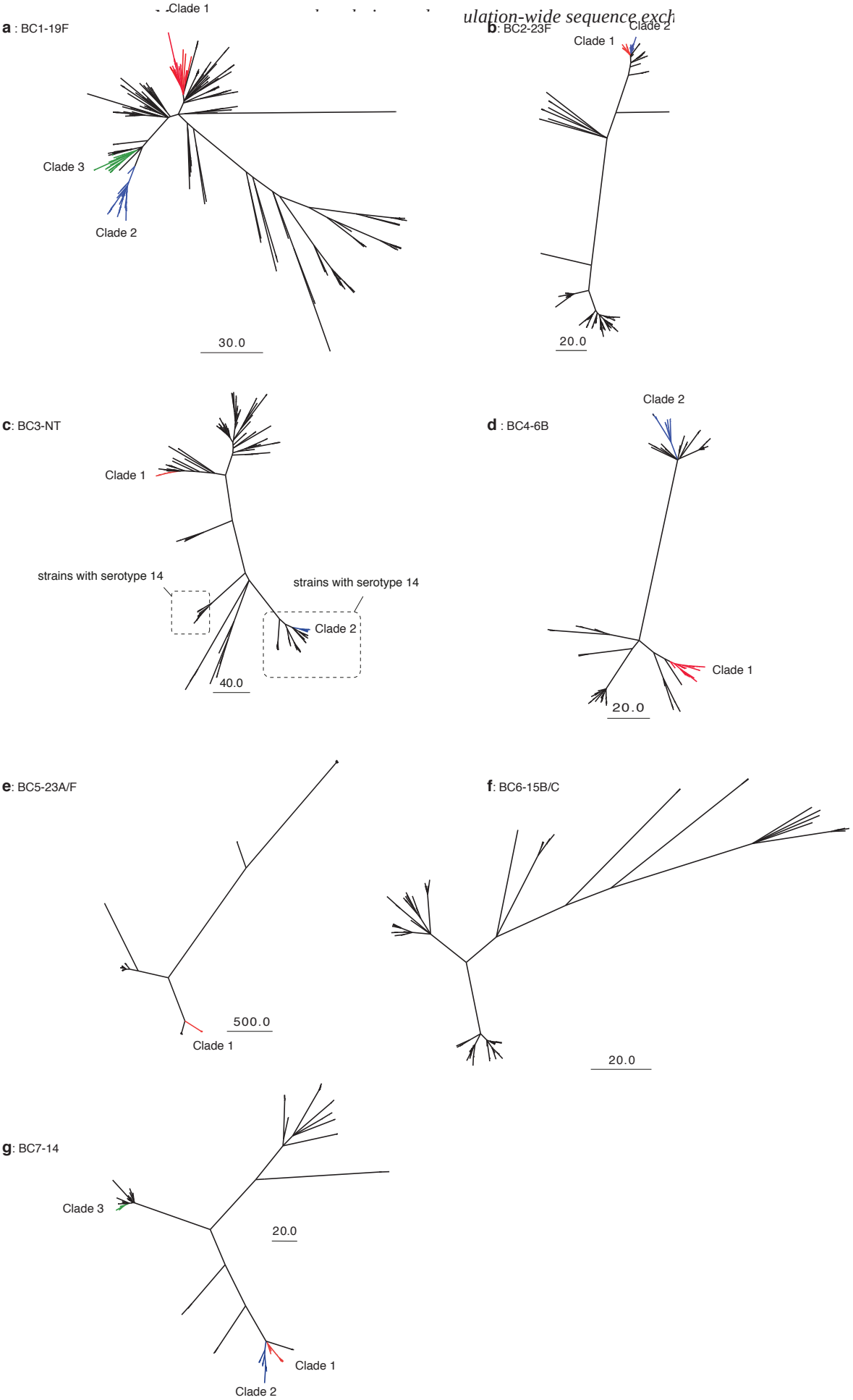
large number of putative recombination events in total. Removal of all recombination signatures from every single isolate removed nearly all the SNPs from the alignment after the first iteration. This left an inadequate number of remaining SNPs to construct a robust phylogenetic tree and consequently terminated all further iterations after the first round. In a single lineage scenario like PMEN-1, recombination from every single isolate accounted for 74% of the reference genome (Croucher, Harris *et al.* 2011), which left sufficient SNPs from nucleotide substitution to draw a vertical phylogeny. Therefore, the method was applied separately to each dominant lineage in Maela instead of the whole dataset.

The largest seven primary BAPS clusters, here denoted by BC1 to BC7 (**Figure 3.4 a**, **Appendix A**), were used as representatives for this analysis. Each cluster contains more than 100 samples and totalled 1,216 genomes accounting for 39.4% of the Maela pneumococcal population: BC1-19F (n=365 isolates), BC2-23F (n=213 isolates), BC3-NT (n=202 isolates), BC4-6B (n=126 isolates), BC5-23A/F (n=106 isolates), BC6-15B/C (n=102 isolates), and BC7-14 (n=102 isolates). It is possible that evolutionary rates from less prevalent lineages are significantly different from the dominant groups selected here and this should be investigated. However, they cannot be reliably assessed with their current sampling size nor can they be grouped together due to the complexity imposed on the computations. Therefore, minority groups were not considered here.

Analyses within focused clusters required genome alignment with higher resolution. To improve the resolution seen in previous chapter, sequence reads for each BAPS cluster were remapped against a closely related reference genome (see Methods) to allow greater sensitivity for detection of variants including single polymorphic changes and small insertions and deletions (indels). Applying the method described in (Croucher, Harris *et al.* 2011) on the improved alignment of each BAPS cluster separated recombination signals from nucleotide substitutions. This resulted in a final tree where only single nucleotide substitution accounts for its evolutionary history (**Figure 4.1**). Subsequently, predicted recombination events were allocated to each phylogenetic branch.

Figure 4.1 Nucleotide substitution based phylogeny and the clusters from which the nucleotide substitution rates were estimated.

Each panel represents a major Maela cluster: (a) BC1-19F, (b) BC2-23F, (c) BC3-NT, (d) BC4-6B, (e) BC5-23A/F, (f) BC6-15B/C and (g) BC-14. Subclades where substitution rates were estimated are highlighted in different colours (blue, green and red) and labelled accordingly. Please note that substitution rates cannot be confidently estimated from any clades in BC6-15B/C. The scale bar represents the number of SNPs. (**Figure is shown on the next page**)


a : BC1-19F
Clade 1
Clade 3
Clade 2
30.0
b: BC2-23F
Clade 1
Clade 2
20.0
c: BC3-NT
Clade 1
strains with serotype 14
strains with serotype 14
Clade 2
40.0
d : BC4-6B
Clade 2
Clade 1
20.0
e: BC5-23A/F
500.0
Clade 1
f: BC6-15B/C
20.0
g: BC7-14
Clade 3
20.0
Clade 1
Clade 2

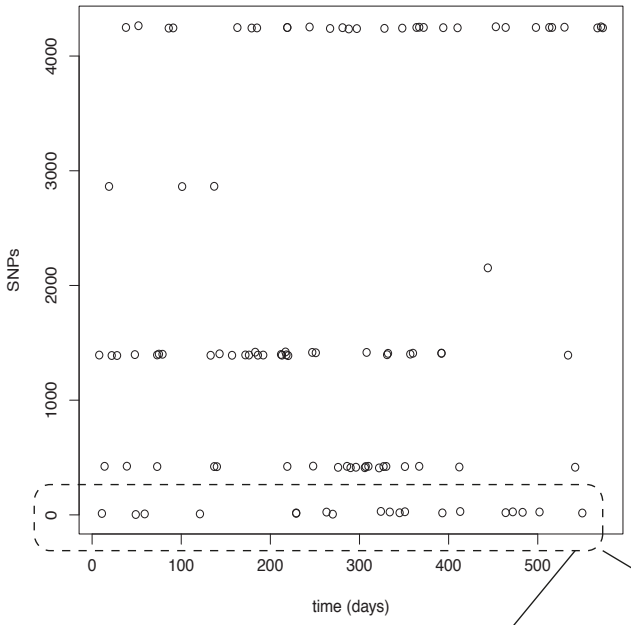
4.3.1.2 Rates of single nucleotide substitution

For each dominant lineage, the rate of single nucleotide substitution can be calculated from a rooted, time-measured phylogeny inferred from the nucleotide substitution tree discussed in previous section. However, there was difficulty in correlating the accumulation of nucleotide substitutions through time with the whole BAPS cluster. As each cluster consists of smaller subclades that co-evolve at the same time, the three-year time span of the isolate collection is not long enough to resolve the combined evolutionary signals from multiple subclades. To illustrate the difficulty in correlating nucleotide substitution and temporal signals from the whole cluster, **Figure 4.2** (top panel) illustrates the absence of nucleotide substitution-time signal when combined signals were considered using BC5-23A/F as an example. A linear regression cannot be determined from the plot as a whole. However, there are distinct clusters representing small subclades where there are good correlations between nucleotide substitution and time. When considering each subclade individually, the temporal signal and clock-likeness of each subclade phylogeny can be captured (**Figure 4.2** bottom panel). A correlation between substitution and time of each subclade was validated using Path-O-Gen (**Figure 4.3**). The subclades showing a positive correlation are highlighted in **Figure 4.1** and were used to estimate the substitution rates with Bayesian MCMC analysis (BEAST, see Methods).

Mean estimated substitution rates fell within the range of 1.45-4.81 x 10^{-6} substitutions per site per year with overlapping 95% credibility intervals (**Table 4.1**). These estimated rates are consistent with the previous report of the PMEN-1 lineage (1.57 x10^{-6} substitutions per site per year, 95% confidence interval 1.34 to 1.79 x 10^{-6}) (Croucher, Harris *et al.* 2011). Although the results appear to be consistent, the rates estimated here have broader confidence intervals due to the much shorter time span of the sampled collection.

Figure 4.2 Demonstration that clock-like signals can be detected from the subclades but not from the whole population.

Each dominant cluster is comprised of more than a single subclade that coevolve together. This plot used BC5-23A/F as an example. The clock signal cannot be detected using the whole cluster as there is confounding from the signals of the subclades.



When all members of the cluster was considered, there was no relationship between time and accumulation of nucleotide polymorphisms. However, distinct subclades can be observed from the plot.

A zoom into the subclade show that there is a positive correlation between time and SNPs, allowing mutation rate to be calculated.

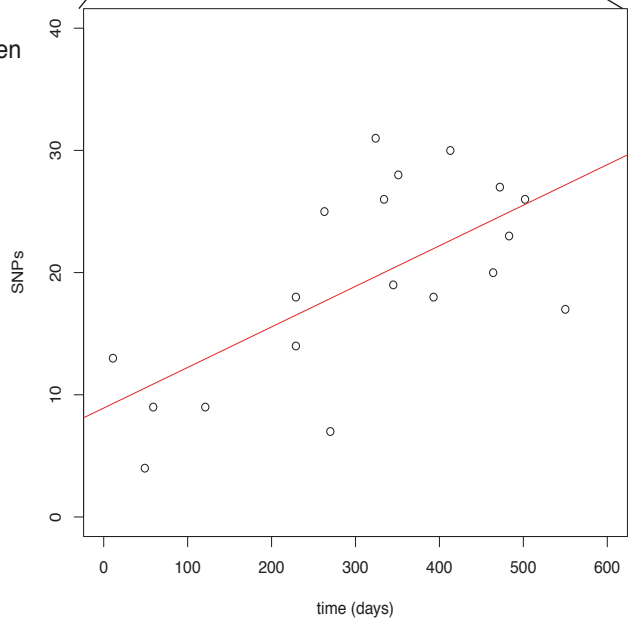


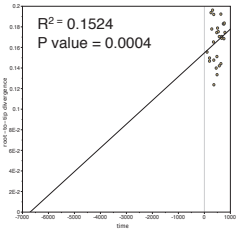
Figure 4.3 Clock-like signals from Path-O-Gen in the subclades where substitution rates were estimated

These subclades were highlighted in Figure 4.1. The y-axis reports root-to-tip divergence while the x-axis represents the time scale in days from the first date of collection, which was 12th November 2007. The first date (time = 0) is shown as a vertical dashed line.

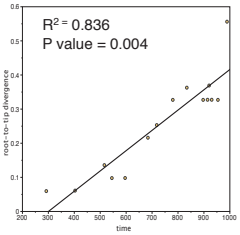
Please note that while a combination of high R^2 and low p-value is expected, regression with low R^2 and low p-value are observed in some cases. These do not necessarily conflict with each other as each represents different predictions. While R^2 measures how close the data are to the fitted regression line, the p-value tests the null hypothesis that the coefficient is equal to zero (no correlation).

(Figure is shown on the next page).

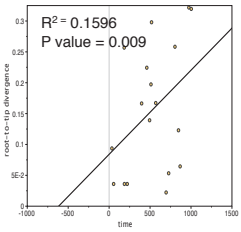
BC1-19F: clade 1



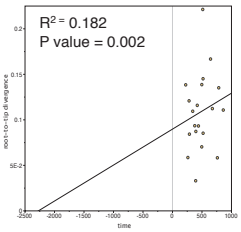
BC1-19F: clade 2



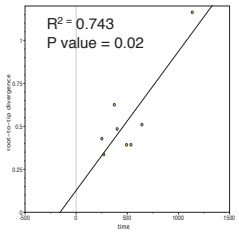
BC1-19F: clade 3



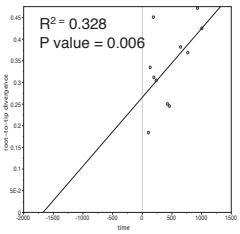
BC2-23F: clade 1



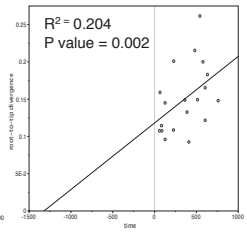
BC3-NT: clade 1



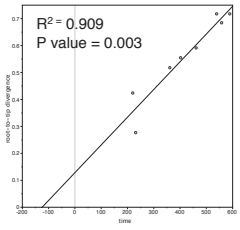
BC3-NT: clade 2



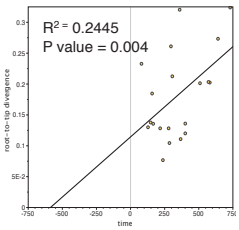
BC4-6B: clade 1



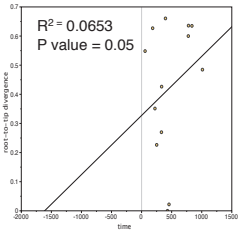
BC4-6B: clade 2



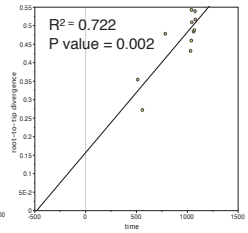
BC5-23A/F: clade 1



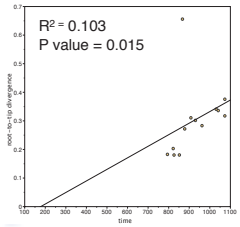
BC5-23A/F: clade 2



BC7-14: clade 1



BC7-14: clade 2



BC7-14: clade 3

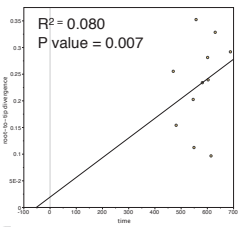


Table 4.1 Nucleotide substitution rates estimated by BEAST

Mean nucleotide substitution rate, the lower bound of the 95% highest posterior density (HPD) and the upper bound of the 95% HPD were tabulated respectively.

	Nucleotide substitution rate (substitutions per site per year)		
	mean	**lower bound**	**upper bound**
BC1-19F clade 1	2.60×10^{-6}	1.58×10^{-6}	3.71×10^{-6}
BC1-19F clade 2	2.86×10^{-6}	1.69×10^{-6}	4.07×10^{-6}
BC1-19F clade 3	2.35×10^{-6}	1.15×10^{-6}	3.68×10^{-6}
BC2-23F clade 1	1.83×10^{-6}	9.96×10^{-7}	2.70×10^{-6}
BC2-23F clade 2	1.45×10^{-6}	2.89×10^{-7}	2.66×10^{-6}
BC3-NT clade 1	4.49×10^{-6}	1.61×10^{-6}	9.78×10^{-6}
BC3-NT clade 2	2.39×10^{-6}	8.49×10^{-7}	4.38×10^{-6}
BC4-6B clade 1	3.08×10^{-6}	1.29×10^{-6}	5.13×10^{-6}
BC4-6B clade 2	3.07×10^{-6}	4.36×10^{-7}	9.83×10^{-6}
BC5-23A/F	3.26×10^{-6}	1.02×10^{-6}	6.60×10^{-6}
BC7-14 clade 1	2.79×10^{-6}	1.16×10^{-6}	4.67×10^{-6}
BC7-14 clade 2	4.81×10^{-6}	5.31×10^{-7}	9.77×10^{-6}
BC7-14 clade 3	4.39×10^{-6}	2.65×10^{-6}	8.14×10^{-6}

4.3.1.3 Rates of recombination

Recombination levels of the 7 dominant clusters were calculated, given numbers of recombination events and number of nucleotide substitutions computed for each branch of the phylogenetic tree as described earlier. Recombination signals can result from both site-specific recombination, potentially associated with mobile genetic elements, and homologous recombination. Signals localised in the regions of mobile genetic elements including prophages and integrative conjugative elements (ICEs) were removed so that only homologous recombination was considered.

Here, the level of recombination was estimated using the relative scale of **r**ecombination events per number of nucleotide substitutions or **m**utations (r/m) observed in each branch. For the numerator (r), the number of recombination events was used instead of numbers of polymorphic sites introduced by recombination as originally given in (Feil, Maiden *et al.* 1999). Depending on the genetic distance between the donors and recipients of a recombination event, numbers of polymorphic sites incorporated in a single recombination event can be variable. Closely related DNA donors may have lower sequence variation compared to distant donors, and thus potentially bring a bias when comparing level of recombination across different clusters. With the use of recombination events instead of polymorphic sites introduced by recombination, the r/m calculated here is expected to be lower than reported earlier (Croucher, Harris *et al.* 2011, Croucher, Finkelstein *et al.* 2013).

The ratio r/m was calculated and compared by two different approaches.

i) By modelling the relationship between recombination events and mutations as a linear regression: recombination events ~ nucleotide substitutions (**Figure 4.4**), using the ranked recombination events as the outcome, and ranked number of nucleotide substitutions as the predictor variable. The slope of each plot represents r/m. Where the assumptions of linearity were met, r/m was calculated and reported in **Table 4.2**.

ii) By using the arithmetic mean of r/m of a cluster. For each cluster, the r/m was calculated separately for each branch and then averaged. Mean and distribution of the r/m of each cluster are tabulated in **Table 4.2**.

With the exception of BC7-14, the r/m calculated from two different approaches shows a good overlap within each cluster, generating a result in which one can have more confidence. The ratio was found to be less than 1 in all studied clusters, indicating that recombination events occur less frequently than nucleotide substitutions.

Figure 4.4 Recombinations per mutation (r/m) of each cluster calculated by linear regression.

Due to the large sample size available in our studies, we alternatively calculated the ratio of recombination events (y axis) over single nucleotide substitutions (x axis) observed on each branch of the slope (r/m) of the linear regression. The number is tabulated in **Table 4.2**. For comparison of r/m by linear regression, all the data are ranked to accommodate the non-parametric ANCOVA analysis. (**Figure is shown on the next page**)

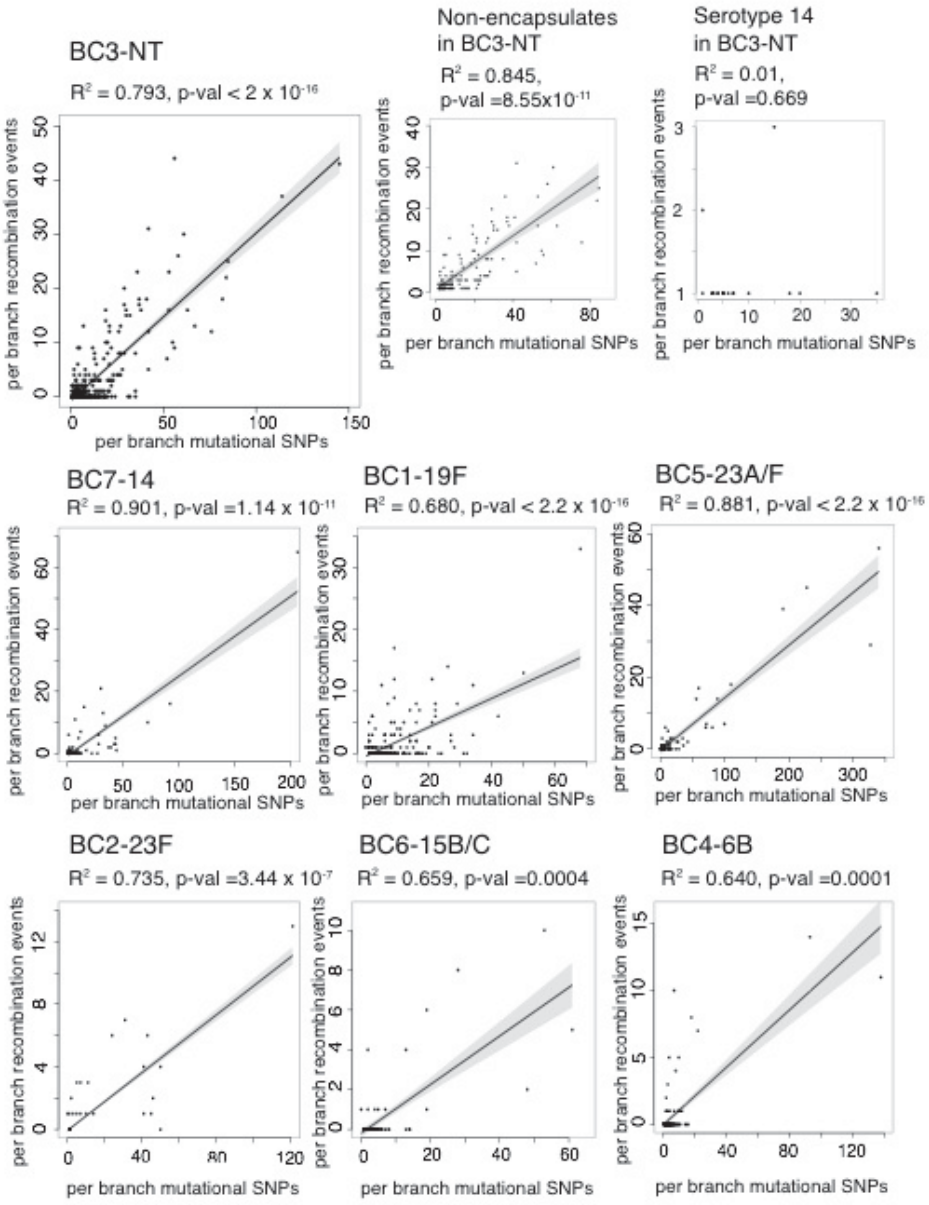

BC3-NT
R^2 = 0.793, p-val < 2 x 10^{-16}
Non-encapsulates in BC3-NT
R^2 = 0.845, p-val =8.55x10^{-11}
Serotype 14 in BC3-NT
R^2 = 0.01, p-val =0.669
BC7-14
R^2 = 0.901, p-val =1.14 x 10^{-11}
BC1-19F
R^2 = 0.680, p-val < 2.2 x 10^{-16}
BC5-23A/F
R^2 = 0.881, p-val < 2.2 x 10^{-16}
BC2-23F
R^2 = 0.735, p-val =3.44 x 10^{-7}
BC6-15B/C
R^2 = 0.659, p-val =0.0004
BC4-6B
R^2 = 0.640, p-val =0.0001
per branch recombination events
per branch mutational SNPs

Table 4.2 Recombination per mutation (r/m) calculated from linear regression and arithmetic mean.

Testing clusters	r/m		Hypothesis	Test and p-value	
	Estimated by linear regression (95% confident interval)	Estimated by arithmetic mean (95% confident interval)		ANCOVA (difference in slope calculated by linear regression)	Kruskal-Wallis (difference in arithmetic mean)
BC1-19F	0.233 (0.210-0.256)	0.299 (0.195-0.263)	r/m of BC3-NT > other clusters	$1.10x10^{-3}$	$1.76x10^{-5}$
BC2-23F	0.092 (0.068-0.117)	0.140 (0.068-0.212)			
BC3-NT	0.310 (0.284-0.336)	0.320 (0.289-0.351)			
BC4-6B	0.107 (0.005-0.209)	0.132 (0.037-0.227)			
BC5-23A/F	0.147 (0.132-0.162)	0.146 (0.100-0.192)			
BC6-15B/C	0.122 (0.070-0.174)	0.200 (0.115-0.285)			
BC7-14	0.257 (0.211-0.395)	0.148 (0.086-0.210)			
NT within BC3-NT	0.341 (0.288-0.395)	0.343 (0.307-0.379)	Within BC3-NT, r/m of NT > serotype 14	NA	$2.44x10^{-3}$
Serotype 14 within BC3-NT	Assumptions of the linear regression models were not met	0.203 (0.164-0.242)			

4.3.1.4 Comparison of evolutionary rates within the population

Next, evolutionary parameters estimated in dominant clusters were compared (**Figure 4.5** top panel). There was no significant difference in rates of nucleotide substitutions between major clusters (Kruskal-Wallis test p value = 0.98). However, the levels of recombination, estimated by the r/m ratio, were significantly different between clusters (Kruskal-Wallis test p value = 1.24×10^{-8}) (**Figure 4.5** bottom panel). The difference in levels of recombination is consistent with previous genome-based (Croucher, Finkelstein *et al.* 2013) and *vitro* studies (Ravin 1959, Yother, McDaniel *et al.* 1986, Hsieh, Wang *et al.* 2006). This might suggest a potential difference in speed of adaptation to changes in environment within the population.

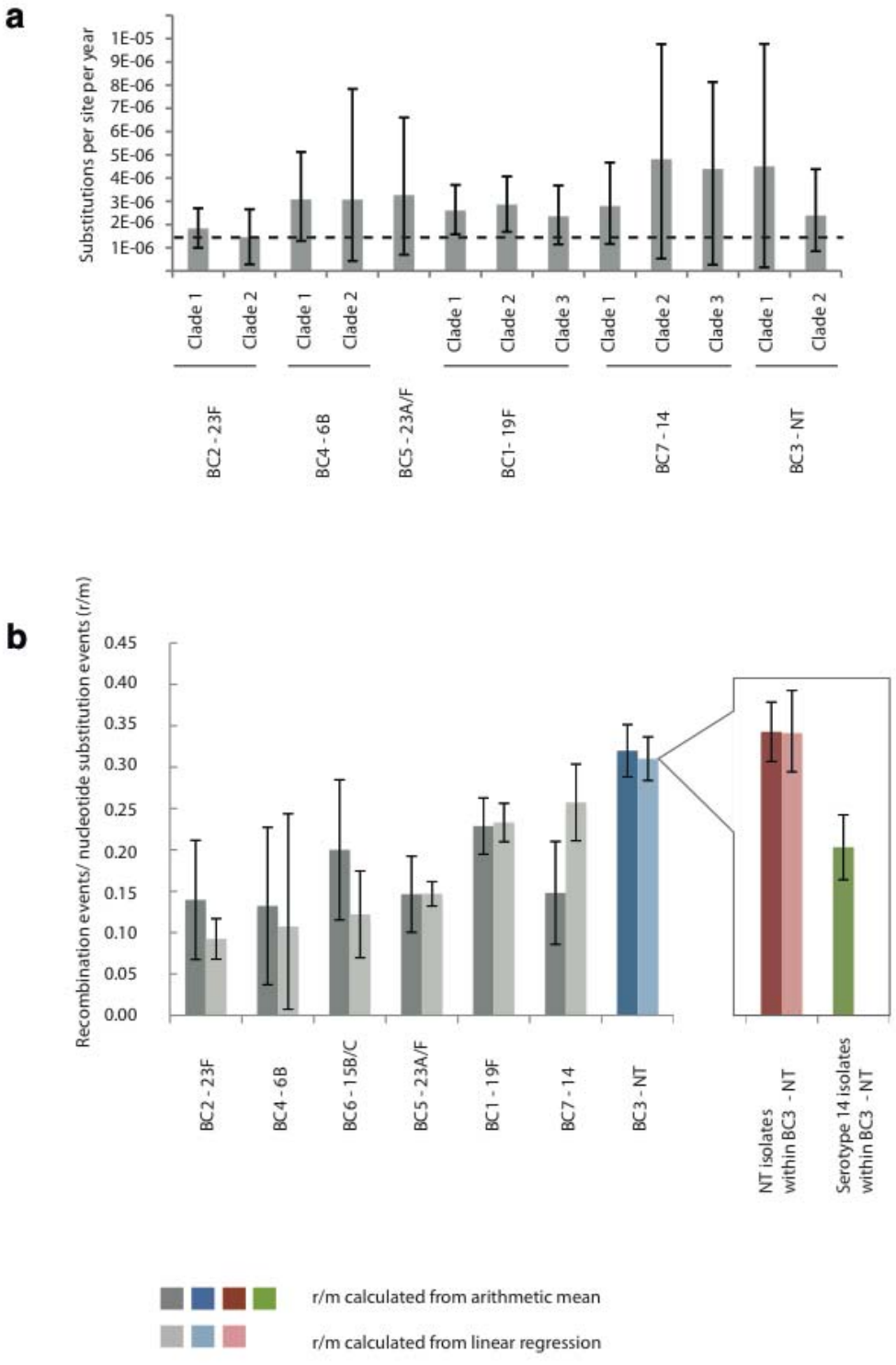
The highest recombination rate was observed in a group dominated by nontypable isolates (BC3-NT) (comparing NT against other groups: Mann-Whitney test p value = 1.76×10^{-5}, ANCOVA test p value = 0.0011, **Table 4.2**). This is consistent with the observation of increased transformation efficiency in capsule defective mutants (Pearce, Iannelli *et al.* 2002); a property widely exploited in laboratory experiments. This is consistent with the general concept that the capsule might act as a physical barrier to DNA uptake *in natura*. Cluster BC3-NT includes a sub-set of isolates able to express the serotype 14 capsule, thereby providing a route to test the idea of the capsule limiting DNA uptake within the same genetic background. When calculated separately, the r/m was significantly higher in the NT isolates compared to serotype 14 isolates (Man-Whitney test p value = 2.44×10^{-3}, **Table 4.2**), indicating that encapsulation reduces recombination efficiency.

Given similar substitution rates, the differences in rates of recombination are likely to be an important factor in the ability of clusters to acquire exogenous DNA with potential selective advantages. With higher level of DNA uptake observed in NTs, one may speculate that they might be fast adapters to environmental changes, and potentially promote the whole population in adjusting to new challenges by disseminating their selectively advantageous genes. However, such a role cannot be established without knowing how much NTs contribute their DNA to the population DNA pool. This question will be tackled in the next section.

Figure 4.5 Comparison of evolutionary parameters estimated in dominant clusters

(a) Comparison of single nucleotide substitution rates estimated using BEAST. Error bars correspond to 95% credibility intervals. The dashed line represents the mutation rate estimated in a previous pneumococcal study of 1.57×10^{-6} substitutions per site per year (95% confidence interval of $1.34\text{-}1.79 \times 10^{-6}$) (Croucher, Harris *et al.* 2011). (b) Comparison of recombination events per mutation (r/m) across the dominant clusters quantified by two separate methods: linear regression on each branch of the appropriate phylogeny and the arithmetic mean of r/m on each branch. Error bars represent 95% confidence intervals. BC3-NT (in blue) has the highest r/m ratio, with its subclusters NT and serotype 14 highlighted in red and green, respectively. Please note that the assumptions of the linear regression models were not met for serotype 14.

(**Figure is shown on the next page**)


a
Substitutions per site per year
1E-05
9E-06
8E-06
7E-06
6E-06
5E-06
4E-06
3E-06
2E-06
1E-06
Clade 1
Clade 2
BC2 - 23F
BC4 - 6B
BC5 - 23A/F
BC1- 19F
BC7 - 14
BC3 - NT
b
Recombination events/ nucleotide substitution events (r/m)
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
BC2 - 23F
BC4 - 6B
BC6 - 15B/C
BC5 - 23A/F
BC1 - 19F
BC7 - 14
BC3 - NT
NT isolates within BC3 - NT
Serotype 14 isolates within BC3 - NT
r/m calculated from arithmetic mean
r/m calculated from linear regression

4.3.2 Population-wide sequence exchange

With high sampling density, potential sources of recombination fragments (referred to as the "donor blocks") can be determined given the sequence's identity to the recombined fragments detected in the recipient strains (referred to as the "recipient block"). This allowed a unique opportunity to capture potential recombination donors, which are important players at the other end of the gene flow.

4.3.2.1 Searching for potential recombination donors and search criteria

The recipient blocks identified in BC1- BC7 were searched using BLAST against the rest of the pneumococcal genomes in the Maela data set for identical matches (donor blocks). As the Maela pneumococcal population has co-evolved in the same geographical area, the chance of different sampled pneumococcal lineages exchanging genetic content is high. Several criteria, which will be discussed in the next section, have been tested and applied in this analysis to reduce the false positives and maximise the search specificity. As a result, 443 out of 928 unique recipient blocks were found to have identical matches elsewhere in the data set (**Table 4.3**).

Table 4.3 Numbers of recombination events used for the search, and number of recipient blocks where potential donors were identified following strict search criteria

Total number of recipient blocks used for the search following different criteria			Number of recipient blocks where potential donor blocks were identified
A) All recombination events predicted by method described in Croucher *et al.* (generated in 4.1.1)	**B)** Recombination detected at the external nodes using method in Croucher *et al.*, which are likely to represent recent recombination (criteria discussed in 4.2.1.1)	**C)** Recombination blocks predicted by method described in Marttinen *et al.* that show overlap with regions in B) (criteria discussed in 4.2.1.2)	**D)** Recipient blocks in C) where potential donors can be found (applying criteria discussed in 4.2.1.3, 4.2.1.4, and 4.2.1.5)
2,209	620*	928*	443

* Please note that some recombination blocks predicted by the method of Croucher *et al.* contain multiple blocks predicted by that of Marttinen *et al.*

Criteria used for identifying potential recombination donors include: focusing the search on recent recombination events, using overlapping recombination blocks predicted by two independent algorithms, allowing only identical hits with no unknown mapping character “N”, and an overall check on the search specificity. Each point and its rationale are discussed as follows.

4.3.2.1.1 Only recent recombination events were considered

A focus on recent recombination occurring on the external branches alone reduces the chances of the donor detection being confounded by subsequent recombination events. Therefore only recent recipient blocks detected on the external branches (identified using the algorithm described in Croucher, Harris *et al.* 2011 were considered.

4.3.2.1.2 Using two independent algorithms for predicting recombination fragments in the recipients

Any methods, to a certain level, report false positives. Based on detecting the density of SNPs in a sliding window, the method described in Croucher, Harris *et al.* 2011 allows for high flexibility such that poor mapping regions flanked by high SNP densities can be counted as recombination fragments. This can be observed, for example, for surface protein encoding genes where high sequence diversity may affect mapping against the reference genome due to sequence mismatches. The algorithm is designed to merge such mismatched gaps flanked by SNPs dense region together as single recombination fragments. Results generated by this method thus represent all possible cases of recombination.

However, a search for donor blocks using recipient blocks as the sequence query required a good sequence quality as well as a confident prediction of recombination fragments. To reduce false positives generated in one method, as well as eliminate regions with poor mapping quality due to naturally high sequence diversity allowed in previous algorithm, another algorithm was used to co-detect recombination regions.

BratNextGen was developed to identify foreign DNA fragments that were introduced into the genome by recombination and this has been successfully applied to

pneumococcal genomes (Marttinen, Hanage *et al.* 2012). Unlike the previous algorithm, which takes poor quality regions with flanking recombination signals into account, BratNextGen handles missing data in a different manner and leads to better sequence quality in predicted recombination fragments.

Here, recombination fragments predicted by two independent algorithms were compared. A large majority of predicted regions show overlaps. This can be demonstrated in **Figure 4.6 a** using the phylogeny and overlapping recombined regions detected in BC7-14, one of the smallest dominant clusters, as an example. The length and quality of the predicted fragments from both algorithms were investigated. **Figure 4.6 b-c** and **Table 4.4** summarise comparisons of the length and percent of unknown characters "N" found in the fragments from both algorithms respectively. The results show that while the method described in Croucher, Harris *et al.* 2011 generally predicts larger recombination blocks, BratNextGen (Marttinen, Hanage *et al.* 2012) gave predicted sequences with higher quality. To optimise the output, only sequences co-predicted from both algorithms were used in the next part of the analysis.

Figure 4.6 Comparison of two recombination detection methods

(a) Genome view of recombination fragments predicted by both algorithms. Recombination regions are aligned with taxa on the phylogenetic tree (left). Genome coordinates are labelled on top. Recombination regions exclusively predicted by methods described in Croucher, Harris *et al.* 2011 and Marttinen, Hanage *et al.* 2012 are highlighted in red and blue, respectively. Overlapping regions predicted by both algorithms are highlighted in dark grey. (b) A histogram showing the length of recombination fragments (bp) predicted by two algorithms. (c) Sequence quality of recombination fragments predicted by two algorithms reported as percent "N". For (b) and (c), fragments predicted by tools described in Croucher, Harris *et al.* 2011 and Marttinen, Hanage *et al.* 2012 are shaded in red and blue respectively.

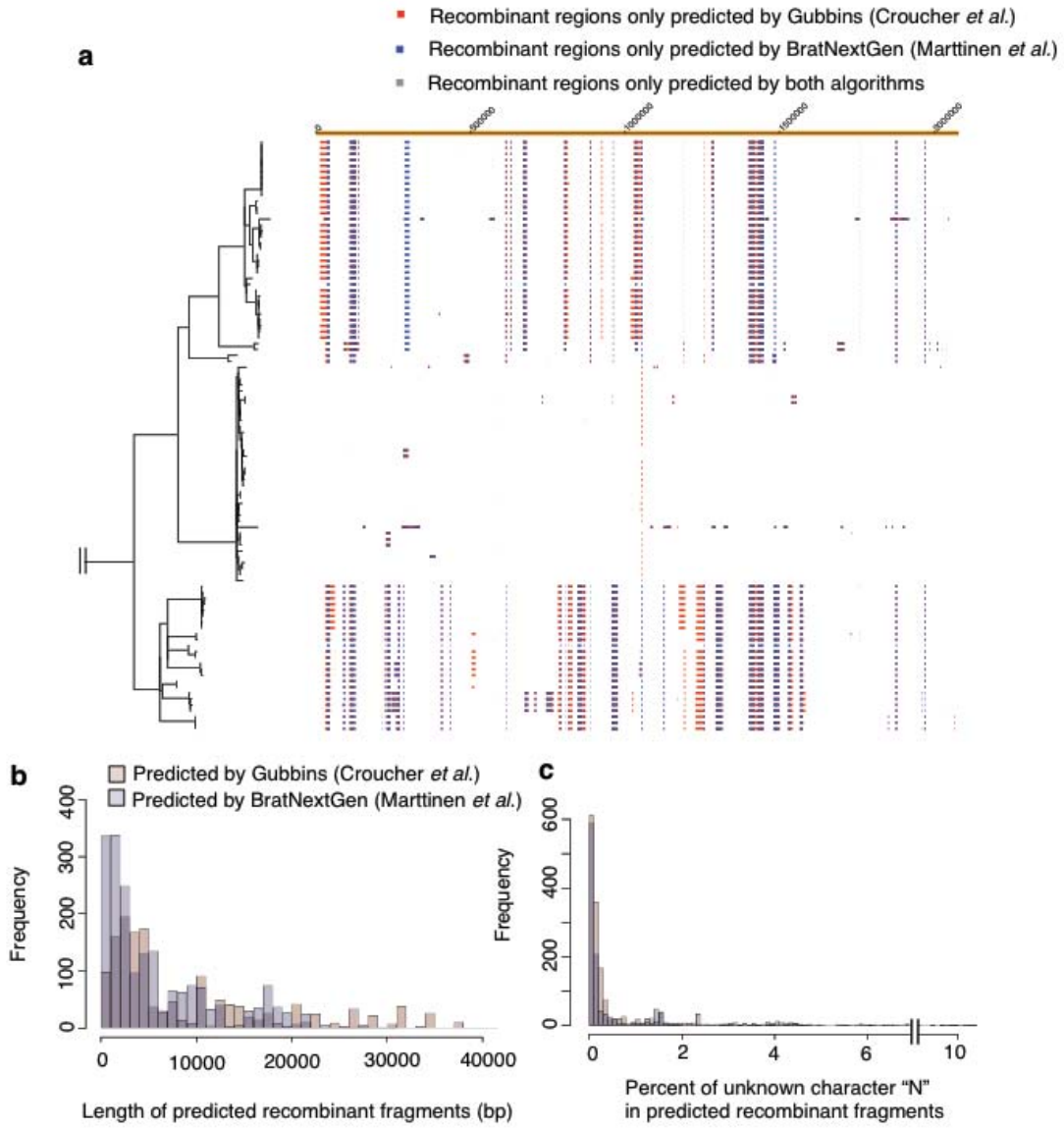


Table 4.4 Comparison of two recombination detection methods as given by Figure 4.6

	methods	1st Quantile	Median	Mean	3rd Quantile
Length of recipient blocks in **Figure 4.6 b**	Croucher *et al.*	2403	4558	10150	14350
	Marttinen *et al.*	1189	3427	5777	8500
	Co-predicted	1177	3057	5418	21263
Percent of "N" in recipient blocks in **Figure 4.6 c**	Croucher *et al.*	0.084	0.200	6.05	2.36
	Marttinen *et al.*	0.066	0.153	5.59	2.00

4.3.2.1.3 No unknown mapping character "N" was allowed

As "N" s, unknown nucleotides, generate non-specific matches in BLAST searches, recipient and donor blocks were checked for sequence quality such that no sequences with "N"s were used.

4.3.2.1.4 Hits must be identical matches

Recombination recipient blocks were used as query sequences for nucleotide blast searches. They were blasted against themselves as positive controls. Any hits that have an exact match to the score given by the positive control are likely to be potential donors. Although recombination may lead to insertions or deletions over recombining regions (Claverys, Lefevre *et al.* 1980, Lefevre, Mostachfi *et al.* 1989, Pasta and Sicard 1996), indels were not considered here as the searches were performed in closely related strains, which require high specificity for the recipient-donor relationship to be drawn.

4.3.2.1.5 A relationship between query length and search specificity

Length of recipient blocks (blast queries) and the number of potential donors detected (blast hits) were checked for search specificity. Queries with successful hits ranged

between 10 - 6,846 bp, with a mean length of 1,162 bp (**Table 4.5**). **Figure 4.7 a** shows similar distribution in length of recipient blocks shown by all queries and queries where hits were detected. This suggests that the blast search did not impose any preference over query length. A decrease in query length usually correlates with increasing broadness or generality of the hits (Phan 2006). The distribution of hits can be modelled with a negative exponential function (**Figure 4.7 b**) in this result. The decay constant of 2.56×10^{-4} bp^{-1} observed here means that on average, the numbers of hit clusters are reduced by half for every 3,903 bp extension of sequence query. These behaviours gave some confidence to the search results. Therefore, all sequence queries that passed criteria described earlier were used for blast searches.

Figure 4.7 Query length and search specificity

(a) A histogram showing distribution of length of recipient blocks from recombination events detected at the tip of the phylogenies. Shaded in grey are all recipient blocks where identical hits were detected from the rest of the population. (b) A plot showing association between the length of sequence queries (recipient blocks) and the diversity of detected hits (potential donor blocks classified by secondary BAPS clusters). The data was modelled as an exponential decay with the line of best fit (red line) and the 95% confidence interval (dashed red lines)

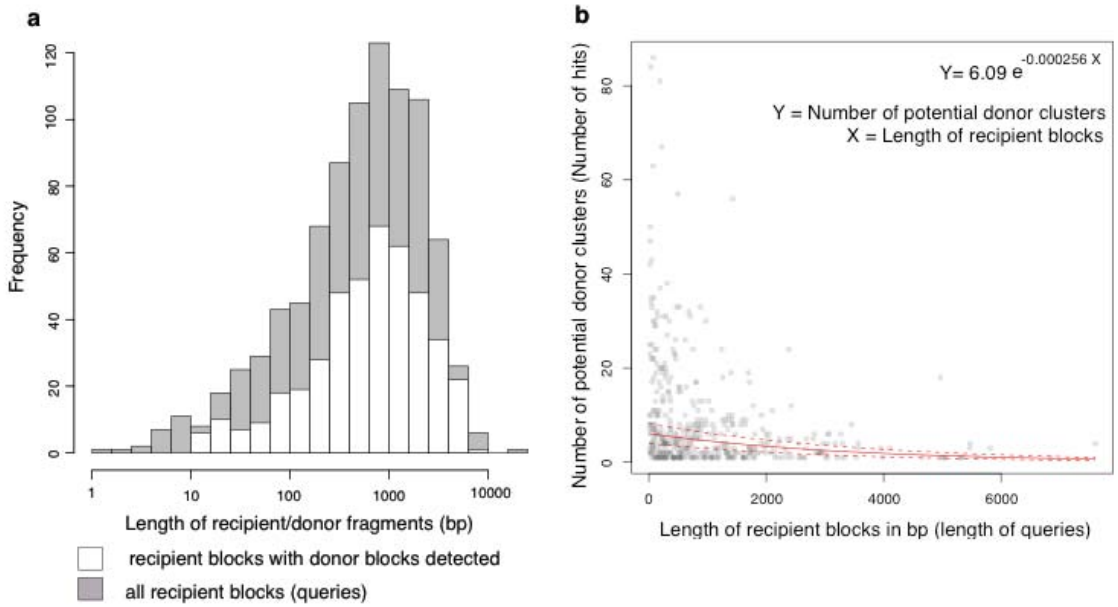


Table 4.5 Distribution of length of recipient blocks described in Figure 4.7 a

Length of fragments (bp)	**Min**	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Max**
All recipient blocks (queries)	3	197	604	1,072	1,412	22,970
Recipient blocks with donor blocks detected (identical hits)	10	275	741	1,162	1,513	6,846

4.3.2.2 Nature of sequence exchange

Following identification of potential recombination donors in 4.2.1, donor strains were classified based on their BAPS clusters (**Appendix A**). This highlighted some interesting biologically features of the genetic flow between the recipient and donor strains.

4.3.2.2.1 A single recipient strain can have multiple donors

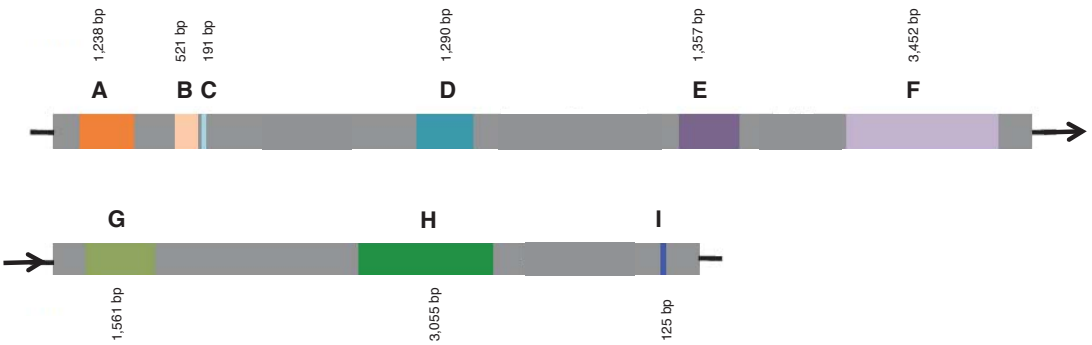
When considering recombination events found in an individual recipient isolate, donors for a single recipient strain could come from a single or multiple genetic backgrounds. For example, isolate SMRU1452 had nine recipient blocks with identical hits detected in eight different clusters, each with a different serotype (**Figure 4.8, Table 4.6**). Eight of the nine recipient blocks (block A, B, C, D, F, G, H, I) were repeatedly detected in one particular cluster, sBC145-sBC152 (serotype 6B), while the remaining block (block E) was only detected in sBC140-sBC143 (serotype NT). Taken together, these observations suggest that the recent ancestor of isolate SMRU1452 had recombined with members of sBC145-152 (serotype 6B) and sBC140-143 (serotype NT), resulting in the import of eight and one DNA region of diversity, respectively.

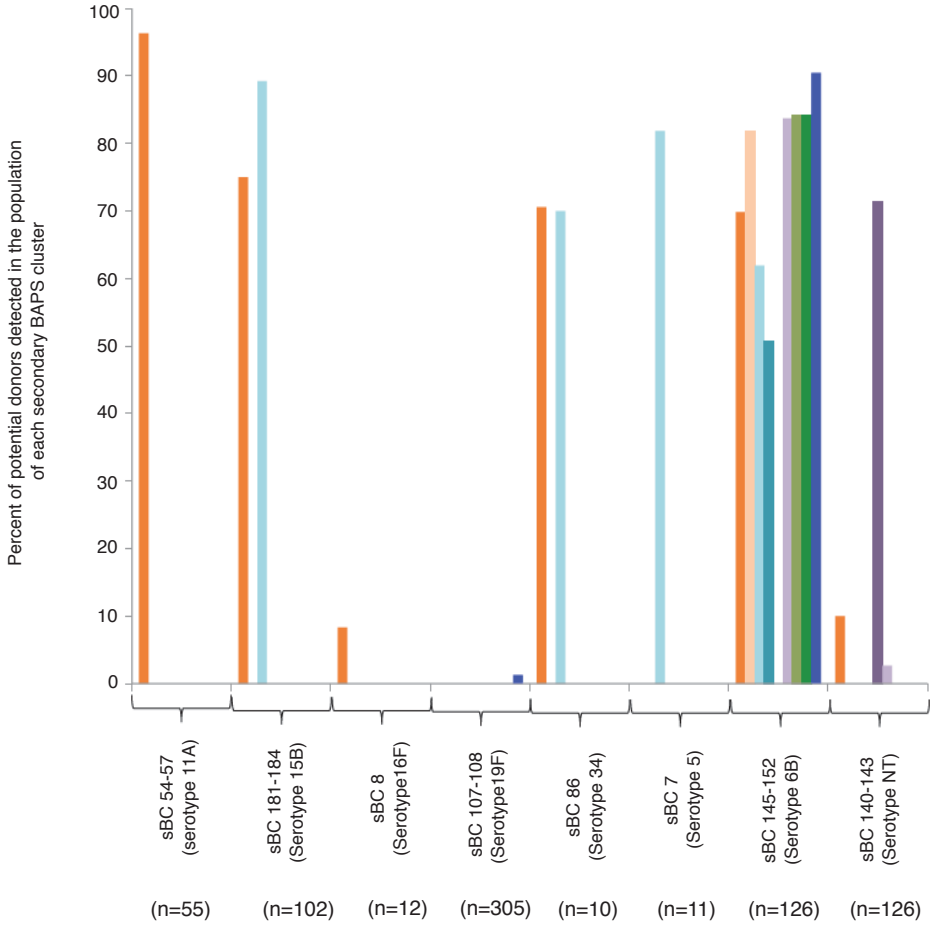
In addition, the shared ancestry of multiple blocks A, B, C, D, E, F, G, H and I possibly suggests that a single transformation event can result in the replacement of DNA at multiple non-adjacent loci of the recipient strain. This observation is consistent with (Hiller, Ahmed *et al.* 2010) where simultaneous replacement of multiple loci following a single transformation event was captured from a clinical strain. Together, these highlight the magnitude of genetic changes potentially introduced by a single transformation event.

Figure 4.8 Multiple potential donors for a single recipient

Top panel: Nine predicted recombination fragments in SMRU1452 (fragments A-I) are highlighted in different colours and are ordered according to their locations on the genome with their size labelled. Bottom panel: the bar chart presents the possible sources of each recombination fragment based on the above colour scheme. The y-axis gives the proportion of hits detected per population of particular lineage. For example, recombination fragment A of 1,236 bp in length was found to have identical matches in 96.29%, 75%, 70.59%, 69.84%, 10% and 8.33 % of the population of secondary BAPS clusters of serotype 11A, 15B, 34, NT, and 16F, respectively. The number of isolates in each secondary cluster was given in the parentheses. (**Figure is shown on the next page**)

Size and relative position of each recombinant blcoks





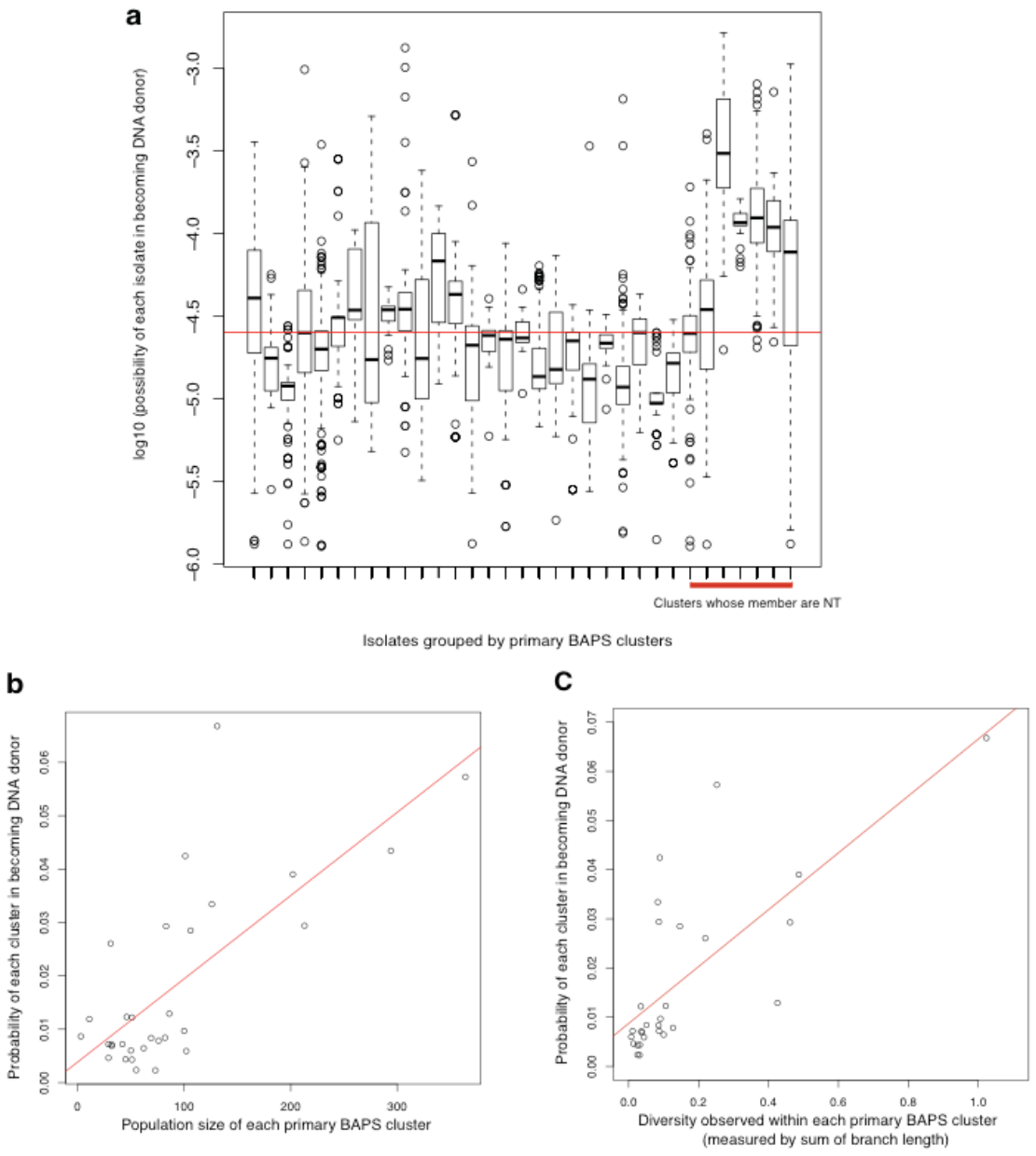
4.3.2.2.2 Probability of a single isolate acting as a donor

The probability of each isolate being a donor was calculated. For each recipient isolate where "n" potential donor isolates were identified, a probability of "1/n" of having been the donor was assigned to each potential donor isolate. Isolates showing no hit for a particular search were assigned a probability of 0. The total likelihood of each isolate for being a donor was represented by the sum of the above probabilities from all donation events. This gave a mean frequency of 2.53×10^{-5} donation events per isolate i.e. each isolate has a probability of 1/39,537 to donate its DNA in recombination events detected here.

The probabilities from individual isolates were then grouped into lineages based on their primary BAPS clusters. The boxplots (**Figure 4.9 a**) showed the distribution of probabilities of isolates within that cluster acting as a donor. Heterogeneity in the donation frequency between each cluster within the population was observed. There was a higher probability of NT isolates acting as the donor than the rest of the population (Mann-Whitney U test between NT isolates and other clusters, p value $< 2.2 \times 10^{-16}$). However, identical matches could come from shared recipient of recombinant fragments as well as true donors. As the two events cannot be distinguished based on the sequenced isolate alone, this finding should be interpreted with caution. As shown earlier, NT isolates are efficient recipients with higher acquisition of recombinant DNA. Therefore the results presented here may be confounded by NT isolates being co-recipients as well as recombination donors.

Figure 4.9 Trends in genetic exchange

(a) Boxplots represent distribution of donation probability of isolates within each cluster. A red bar represents a mean frequency of donation event of any isolates ($2.53x10^{-5}$), i.e. each isolate has a probability of 1/39,537 to donate DNA in a recombination event. (b) and (c) respectively show positive correlations between potential donor clusters (based on primary BAPS clusters) and outer population size, and separately cluster diversity.



4.3.2.2.3 Probability of a cluster acting as a donor and its relationship with cluster size and diversity

To reduce the number of detections from non-donor isolates, the probability of being the donor was considered as a cluster. Cluster filters were applied to remove random hits from non-related clusters, particularly the co-recipients that potentially confounded the result. Either i) clusters most commonly detected as sources for each recipient isolate or ii) clusters detected as the sole source of recombinant DNA in each recipient isolate, were allowed. To demonstrate, a case where multiple potential donor clusters were identified from nine recipient blocks of a single recipient isolate SMRU1452 (discussed in 4.2.2.1) was used here. Tabulated in **Table 4.6**, a hit to each cluster was scored as "1" while no hit was scored "0". The most common sources came from sBC145-152 and this was selected based on the first criterion. Also, sole-source donors were detected in sBC145-152 (sole sources for recipient blocks B, D, G, and H) and sBC140-143 (sole source for recipient block E). These clusters were chosen according to the second criterion. This effectively reduced the number of potential donor clusters from eight to two confident clusters.

For a search matching "n" potential donor clusters for each recipient block, each identified cluster was assigned a probability of "1/n" of being the donor. Clusters that did not contain any potential donors were given a probability of 0. The total probability of each cluster acting as donor is presented as the sum of above probabilities.

The probability of being recombination donors was associated with two other characters detected in each cluster: the cluster population size, and the genetic diversity within each cluster. The population size represented the number of isolates detected in a particular cluster. Plotting the probability of a cluster being a donor against the cluster size gave a positive correlation ($\rho = 0.592$, p-value = 2.69×10^{-4}, **Figure 4.9 b**). Another feature, the diversity within a cluster was calculated as the sum of total branch length, which is proportional to the number of polymorphic sites observed in a particular cluster. A positive correlation was also observed between the probability of a cluster being a donor and the cluster diversity ($\rho = 0.773$, p-value =

1.45 x 10^{-6}, **Figure 4.9 c**). This is consistent with the concept that higher diversity gives a greater chance to discover minority instances (Wang, Yao *et al.* 2009), some of which were captured as the donor for recombinant fragments.

This result suggests that the likelihood for each cluster of being a recombination donor increases with the cluster population size and the cluster diversity. Both features were observed in the clusters of NT isolates. Therefore, NTs could potentially be major donors, contributing their DNA more frequently to the population gene pool.

Table 4.6 Potential donors for each recombinant fragment detected in isolate SMRU1452.

		Recipient blocks								
		A	B	C	D	E	F	G	H	I
Donor clusters	sBC 54-57	1	0	0	0	0	0	0	0	0
	sBC 181-184	1	0	1	0	0	0	0	0	0
	sBC8	1	0	0	0	0	0	0	0	0
	sBC107-108	0	0	0	0	0	0	0	0	1
	sBC 86	1	0	1	0	0	0	0	0	0
	sBC 7	0	0	1	0	0	0	0	0	0
	sBC 145-152	1	1	1	1	0	1	1	1	1
	sBC 140-143	1	0	0	0	1	1	0	0	0

4.4 Conclusion

This chapter summarises the evolution and genetic exchange observed in a densely sampled pneumococcal carriage population. A high sampling density of 3,085 isolates collected over a 3-year period in a refugee camp allowed comparisons of evolutionary rates within the population as well as identification of the source and sink of sequence exchange. Heterogeneity in rates of recombination for both donation and receipt of DNA suggests a structure to genetic flux within the population.

The high rate of receipt of recombination in NT pneumococci is consistent with that observed in NT lineages from some other species (Connor, Corander *et al.* 2012). This is consistent with the general concept that capsule might act as a physical barrier for DNA uptake. The higher rates of both receipt and donation of recombinant fragments observed here in NTs suggest that these clusters might function as hubs of gene flow in the Maela pneumococcal population. Though an increased recombination rate could bring transient benefit, there are potential long-term disadvantages due to increasing genomic instability (Giraud, Matic *et al.* 2001). So it is notable that sporadic switches between the NT and encapsulated states (discussed in chapter 3) may serve as a mechanism to modulate the trade-off between benefit and cost of increased recombination rates.

This chapter introduced the key players of genetic exchange in the Maela pneumococcal population. The next chapter will explore the genes that have been exchanged in relation to selection pressure, particularly the high consumption of antibiotics observed in this refugee camp.

Chapter 5: Recombination allows rapid adaptation in response to local selective pressure

---

Declaration of work contributions:

Clinical records on antibiotic usage described in 5.2.1.1 were kindly provided by Dr Paul and Claudia Turner.

5. Recombination allows rapid adaptation in response to local selective pressure

5.1 Introduction and aims

Homologous recombination is a major driving force in the evolution of *Streptococcus pneumoniae* genomes and a key contributor to genetic diversity and population dynamics. Its importance in the failure of clinical intervention such as antibiotics has been increasingly recognised (Dowson, Hutchison *et al.* 1989, Laible, Spratt *et al.* 1991, Hanage, Fraser *et al.* 2009, Croucher, Harris *et al.* 2011). Following the identification of population-wide sequence exchange discussed in chapter 4, this chapter explores the contents of the exchanged genetic materials, many of which are likely to be a reflection of host immunity and clinical practices. The analysis shows that highly recombined regions in the Maela pneumococcal population are concentrated on genes encoding cell surface antigens and genes associated with resistance to antibiotics. The analyses will next focus on the directionality of recombination in response to changes in selection pressure: monitoring bacterial response to an increase as well as a decrease in consumption of two classes of antibiotics, beta-lactams and co-trimoxazole respectively. An increase in beta-lactam consumption coincided with recombination events that lead to the population becoming beta-lactam non-susceptible, providing genomic evidence of adaptation in response to antibiotic usage. On the other hand, a reduction in co-trimoxazole consumption has been found to coincide with a decrease in the number of co-trimoxazole resistant isolates compared to sensitive isolates in strains that have undergone recent recombination. Together, these observations further support the role of recombination in pneumococcal adaptations, particularly in the presence and absence of clinical interventions.

This chapter aimed at:

i) Defining the regions where recombination was observed at heightened frequencies and their biological relevance.

ii) Associating the recombination trends observed in the population with known changes in selection pressure.

5.2 Methods

5.2.1 Preparation of nucleotide sequences for penicillin-binding proteins (*pbps*), dihydrofolate reductase (*dhfR*) and dihydropterpate synthase (*folP*)

The DNA sequences of the above genes, whose allelic forms are known to confer resistant to β-lactam (*pbp1a, pbp2b, pbp2x*) and co-trimoxazole (*dhfR, folP*), were extracted from the draft genome assembly using Glimmer3 (Delcher, Bratke *et al.* 2007) for gene prediction. Predicted genes were searched for similarity against all *pbp1a, pbp2b, pbp2x, dhfR, folP* genes available in the public dataset using blastall v 2.2.15 (Altschul, Gish *et al.* 1990) with default settings, thereby allowing large diversity of these genes especially for *pbp1a*, *pbp2b* and *pbp2x* to be captured.

5.2.2 Phylogenetic analyses

Phylogenies of individual gene trees were estimated with RAxML v7.0.4 (Stamatakis 2006) using a GTR model with a gamma correction for site rate variation using 100 bootstraps.

5.2.3 Statistical tests

5.2.3.1 Associations between recombination and resistant phenotypes

The trend of recombination was estimated through the detected phenotypes observed in the presence and absence of recombination in the sub-population including 7 most prevalent clusters. Please note that 5 isolates with missing phenotypes (**Appendix A**) were not included in this analysis. Based on the prediction of recombination from 7 dominant clusters, strains undergoing recombination at *pbp1a, pbp2b, pbp2x, dhfR* or *folP* and their phenotypic resistance to β-lactam and co-trimoxazole were compared against the strains with no recombination events observed at these sites. The statistical difference between the recombining group and the non-recombining group was estimated with two-tailed Fisher's exact test.

Please note that beta-lactam resistance phenotype were considered as the binary outcomes with the breakpoint defining the categories of susceptible and non-susceptible to beta-lactams as ≤ 0.06, and > 0.06 μg/ml respectively. The breakpoint defining categories sensitive, intermediate and resistant for co-trimoxazole were given at a 1/20 ratio of trimethoprim/sulfamethoxazole as ≤ 0.5/9.5, 1/19-2/38, and ≥ 4/76 respectively.

5.2.3.2 Associations between past and recent recombination events and resistant phenotypes

Temporal trends were determined by comparing phenotype differences in isolates showing evidence of recent recombination (events predicted at external branch using algorithm described previously (Croucher, Harris *et al.* 2011)) to isolates whose ancestors had undergone recombination (events predicted at internal nodes). Two-tailed Fisher's exact test was used to test the significant difference in trend between the two temporal groups.

Note that alternative *murM* and *murN* genes associated with high beta-lactam resistance (Smith and Klugman 2001) were also considered. However, only two candidates with partial matches were observed and are thus less likely to explain trends in beta-lactam resistance.

5.3 Results

5.3.1 Biological relevance of sequences that have undergone recombination

To investigate the impact of recombination introducing selectively advantageous sequences, the frequency of recombination events across all genomic sites for the seven dominant clusters (BC1-BC7, see 4.1.1) was analysed and correlated to potential selection pressures applied to the community.

Here, recombination hotspots are defined as the genome location where recombination events were observed with higher frequency as a result of selection. Based on the recombination events identified in 4.1.1, the recombination frequency for each site was counted from a range of genome coordinates where homologous recombination events were predicted (**Figure 5.1**). Recombination occurring at sites of mobile genetic elements, although having the potential to have been generated through both site-specific recombination and homologous recombination, was not considered here. Hotspots were defined as a recombination frequency above the 95th percentile of site frequencies detected for the cluster as a whole, thus accounting for both recombination frequency and population size of each cluster.

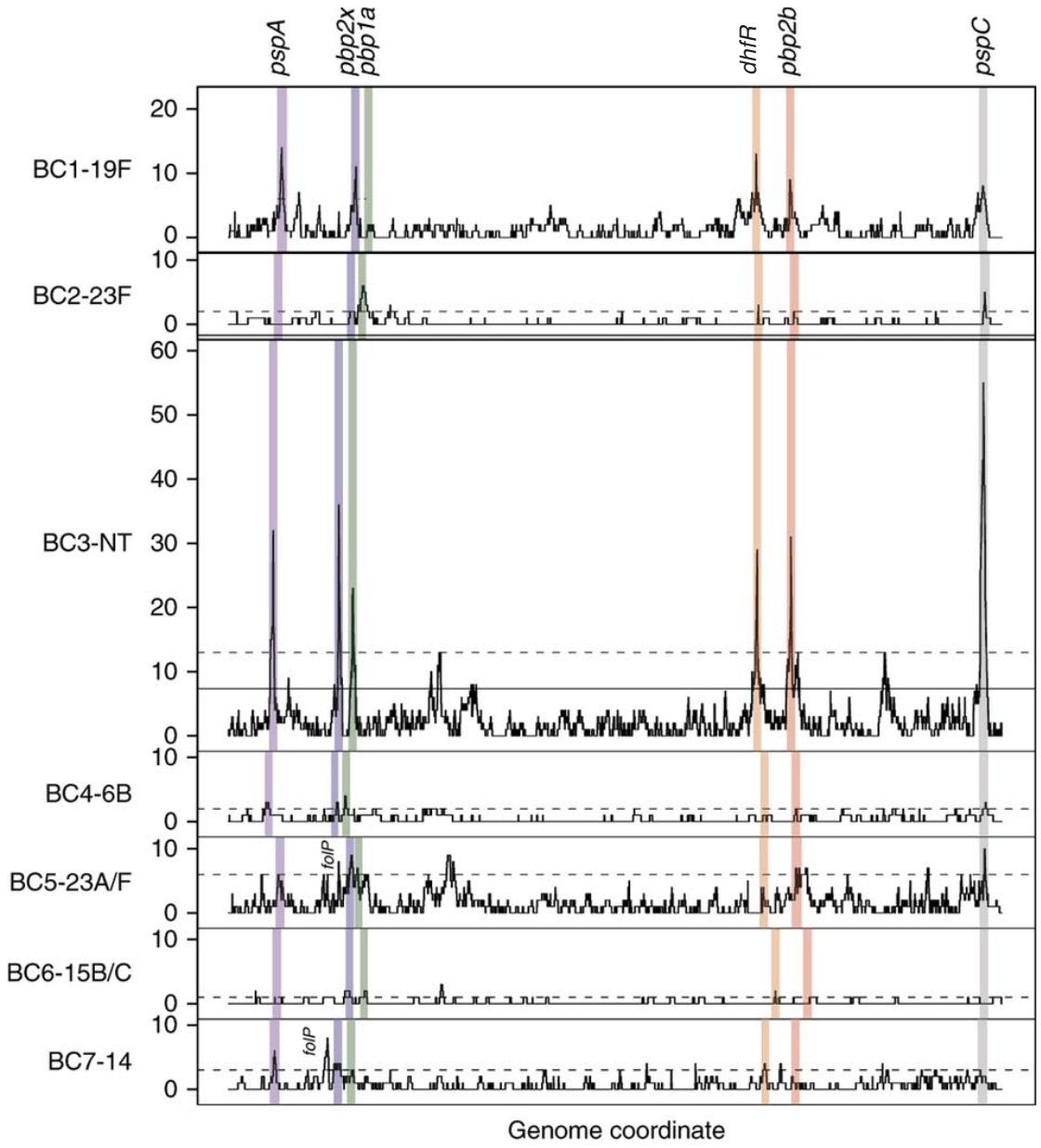
Consistent with the observed highly recombinogenic nature of the nontypable strains discussed in 4.1.3, the subpopulation cluster BC3-NT contained the highest frequency hotspots. This is despite its being third in term of population size (n=202 isolates), after two dominant encapsulated clusters, BC1-19F (n=365 isolates) and BC2-23F (n=213 isolates).

For each dominant cluster, a non-random distribution of detected recombination events across genomic sites was observed (**Figure 5.1**). Although recombination frequencies vary between clusters, the positions of recombination hotspots in most clusters were remarkably similar. This consistency within the Maela population potentially highlights two important points. Firstly, despite high levels of species-wide genotype diversity, major selective pressures appear to have equal impact across investigated clusters, resulting in similar patterns of selection. Secondly, the force of selection for these recombination sites must have acted rapidly enough for clusters to have similar hotspot patterns regardless of their time since transmission to the Maela community. Host immunity and clinical practices may account for these hotspot patterns. Indeed, the six leading hotspots in the Maela pneumococcal population are among genes encoding cell surface antigens (*pspA, pspC*) and genes associated with resistance to antibiotics (*pbp1a, pbp2b, pbp2x* and *dhfR*).

Figure 5.1 Recombination hotspots

Panels (top to bottom) are ordered by decreasing cluster population size. For each cluster, recombination hotspots were identified as sites with recombination frequency above the 95th percentile of the homologous recombination frequency detected in that cluster. The 95th percentile levels are represented as horizontal dashed lines. Recombination hotspots detected in at least four of the seven studied clusters are shaded in different colours. These common hotspots (in order by genomic coordinates) encode pneumococcal surface protein A (*pspA*, purple), penicillin-binding protein 2x (*pbp2x*, blue), penicillin-binding protein 1a (*pbp1a*, green), dihydrofolate reductase (*dhfR*, orange), penicillin-binding protein 2b (*pbp2b*, red) and pneumococcal surface protein C (*pspC*, grey). The figure includes 2,077 recombination events; the 132 events associated with mobile genetic elements are not displayed.

(**Figure is shown on the next page**)


pspA
pbp2x
pbp1a
dhfR
pbp2b
pspC
BC1-19F
BC2-23F
BC3-NT
BC4-6B
BC5-23A/F
BC6-15B/C
BC7-14
folP
folP
Genome coordinate

5.3.1.1 Hotspots associated with surface antigens

To allow asymptomatic carriage, it is essential for pneumococci to be transmitted without being recognised by the host immune system. Two evolutionary models have been proposed to explain the co-evolutionary processes taking place between bacteria and their hosts. The balanced polymorphism model, also called trench warfare model (Bergelson, Dwyer *et al.* 2001) predicts constant diversity in the population due to selective pressure to maintain polymorphisms especially to evade the host immune system. The model predicts that the fitness of the phenotypes increases as it becomes rarer. Another model, the evolutionary arms race or the red queen hypothesis suggests that there is co-evolution between the bacterial protein and the targeted host molecule (Dawkins and Krebs 1979). Therefore, bacterial surface proteins, which are directly involved in host-pathogen interactions, are likely to show greater diversity in sequence patterns than sequences that evolve in neutral manner according to these two models (Toft and Andersson 2010). This analysis suggests that recombination seems to be the tool to generate and spread the diversity in pneumococcal surface proteins.

Indeed, pneumococcal surface protein A (*pspA*) and pneumococcal surface protein C (*pspC*) have undergone recombination at a higher frequency, manifested as recombination hotspots in four and six of seven main Maela clusters respectively (**Figure 5.1**). This observation is consistent with hotspots reported in PMEN-1 (Croucher, Harris *et al.* 2011) and PMEN-14 (Croucher, Chewapreecha *et al.* 2014). PspA has been shown to be an important virulence factor for both invasive infections and nasopharyngeal colonisation. Pre-existing antibodies against PspA were shown to prevent carriage completely, suggesting that the protein is essential for colonisation in the human host (McCool, Cate *et al.* 2002). The protein acts to prevent the binding of host complement component C3 to the bacterial surface (Shaper, Hollingshead *et al.* 2004), the process by which the bacterium is marked for ingestion by host phagocyte known as opsonophagocytosis. This allows the bacterium to evade detection and destruction by the host immune system. Given its role in host-pathogen interaction, it

is not surprising to observe high recombination frequency at *pspA*, in consistent with its high sequence diversity reported previously (Hollingshead, Becker *et al.* 2000).

A paralogue of PspA, PspC also helps the bacterium to avoid host immune detection. It prevents the binding of host factor H to pneumococci, which again leads to the inhibition of complement-mediated opsonophagocytosis (Dave, Pangburn *et al.* 2004, Quin, Onwubiko *et al.* 2007). Moreover, a binding of a specific PspC subclass to a polymeric immunoglobulin receptor, a host protein highly expressed in the nasopharyngeal epithelium, results in an internalisation of receptor-bacterium complex into the host cell (Brock, McGraw *et al.* 2002). This process facilitates the invasion of pneumococci across mucosa epithelia.

Recombination hotspots were also observed in other surface proteins, but were limited to one or two clusters. These extra hotspots include Immunoglobulin A1 metalloprotease (*zmpA*) and pneumococcal serine-rich repeat protein (*psrP*). ZmpA is a protease that cleaves host immunoglobulin molecules attached to the surface of the bacterium and thereby prevents triggering of the host inflammatory response (Bek-Thomsen, Poulsen *et al.* 2012). Another integral membrane protein, PsrP is known to mediate pneumococcal aggregation in biofilms (Shivshankar, Sanchez *et al.* 2009). Although recombination frequency observed in *zmpA* and *psrP* did not reach the 95th percentile cut-off population-wide, their biological functions imply that diversifying selection, to a certain level, may well have acted on these genes to help the pneumococci thrive in their host.

5.3.1.2 Hotspots associated with antibiotic resistance determinants

Genes where allelic forms are known to confer resistance to beta-lactams and co-trimoxazole were also identified as recombination hotspots. These include penicillin binding proteins encoded by *pbp1a, pbp2b* and *pbp2x* which are known targets for a group of beta-lactam antibiotics (for an excellent review, see Hakenbeck, Bruckner *et al.* 2012). Beta-lactam functions as an inhibitor of penicillin-binding proteins (PBPs),

thereby preventing the complete formation of peptidoglycan in the cell wall during cell growth. The resistance to beta-lactams occurs with variant penicillin-binding proteins that have lower affinity for beta-lactams. This abolishes the inhibition and allows the formation of a complete cell wall. It has been long recognised that penicillin-binding proteins are highly altered in clinical isolates with mosaic gene structure; thereby indicating interspecies gene transfer from closely related commensal species mediated by recombination events (Dowson and Hutchison *et al.* 1989, Laible and Spratt *et al.* 1991).

Other antibiotic-resistance associated hotspots were detected in dihydrofolate reductase (*dhfR*) and, to a lesser extent dihydropteroate synthase (*folP*) (**Figure 5.1**). The allelic forms of *dhfR* and *folP* are known to confer resistance to trimethoprim and sulfamethoxazole, drugs that are commonly prescribed in combination as co-trimoxazole due to their synergistic effects (Adrian and Klugman 1997, Padayachee and Klugman 1999, Silver 2007). Both trimethoprim and sulfamethoxazole act by interfering with the synthesis of precursors in the thymidine synthesis pathway, thereby inhibiting bacterial DNA synthesis. Similar to beta-lactams, co-trimoxazole resistance develops with variant *dhfR* and *folP*, which have lower affinity to the drugs (Maskell, Sefton *et al.* 2001). This removes the inhibition and allows DNA synthesis to continue.

High recombination frequencies observed in *pbp1a*, *pbp2b*, *pbp2x*, *dhfR* and *folP* suggests that there have been extensive exchanges of different alleles at these gene loci. While replacements of sensitive by resistant alleles potentially lead to more isolates becoming non-susceptible, replacements of resistant by sensitive alleles may lead to a resistant being superseded by a sensitive population. The directionality of these exchanges cannot be determined by the data presented in **Figure 5.1** alone. This topic, as well as potential selection pressures driving the directionality will be discussed in 5.2.

5.3.1.3 Hotspot patterns vary over time

Not surprisingly, genetic reshuffling patterns resulting from recent recombination (events predicted at the external node, previously described in 4.2.1.1) appear to be more random compared to older recombination events (predicted at internal nodes) where more pronounced hotspot patterns can be observed (**Table 5.1**, Chi-square test comparing the ratio of hotspot patterns in recent and older recombination events = 13.979, two-tailed p value = 2.00×10^{-4}). This is expected given that the recombination process occurs at random. Under neutral evolution, recombination results in the reshuffling of alleles all over the genome. However, selection for advantageous alleles results in the pattern of hotspots where only a limited number of genes show a heightened recombination frequency relative to the rest of the genome. The selection process itself takes time. Therefore, the random pattern can still be observed in recent recombination events, in contrast to older recombination events where hotspot patterns are more pronounced, as there has been enough time for selection to act. This is consistent with the definition of recombination hotspots as a result of the combined action of recombination events in the past and selection pressure.

Table 5.1 Recombination signals have been refined through time

	Number of recombination events detected in seven dominant clusters (BC1-BC7) of method described in Croucher *et al.*		
	Total events	Events associated with recombination hotspots discussed in 5.1.1 and 5.1.2	Events occurring outside recombination hotspots
Recombination events predicted at internal nodes, which are likely to represent older recombination	1,589	356 (22.4%)	1,233 (77.6%)
Recombination events predicted at external nodes, which are likely to represent recent recombination	620	94 (15.2%)	526 (84.8%)

5.3.2 Changes in recombination trends reflect changes in selection pressure

Next, hotspots associated with antibiotic resistance described in 5.1.2 were used to demonstrate the associations between trends in recombination and temporal changes in selection pressures. The analysis focused on bacterial genetic and phenotypic responses to an increase in beta-lactam and a reduction in co-trimoxazole consumption.

5.3.2.1 Trend in antibiotic consumption in Maela

5.3.2.1.1 Measurable clinical prescriptions

With the availability of clinical records in Maela, the selection pressure applied to pneumococcal population can be estimated based on the trends in antibiotic prescriptions. **Table 5.2** summarises the trends in consumption of two antibiotics frequently used in the SMRU clinic, beta-lactams and co-trimoxazole. This local record (1994 – 2010) predated the pneumococcal collection used in the study (2007-2010), allowing genomic signals from evolutionary events years prior to the sampling time to be interpreted. Co-trimoxazole was recommended as a primary treatment for non-severe pneumonia in Maela from 1994 until 2002. However, due to increasing resistance across the region (Hoge, Gambel *et al.* 1998) and several side-effects (Medscape 2014), its use has been in decline. This is contrast to the rise in beta-lactam consumption. Based on this clinical information, it is likely that the pneumococci have been subjected to reducing pressure to become co-trimoxazole resistant. However, the population appeared to be under continuous pressure to be beta-lactam non-susceptible.

Table 5.2 Trend in antibiotic consumption based on the Burmese border guidelines (1994-2010)

Year	Co-trimoxazole consumption	Beta-lactam consumption
1994	Co-trimoxazole was the primary treatment for non-severe pneumonia, otitis media, urinary tract infection, and dysentery.	Ampicillin was used for severe pneumonia, meningitis and for infections in pregnant women where co-trimoxazole was contra-indicated.
1999	As in 1994.	As in 1994, but with amoxicillin being used as second line treatment for non-severe pneumonia (if no improvement with co-trimoxazole).
2002/2003	As in 1999 but ciprofloxacin replaced co-trimoxazole for dysentery.	As in 1999 with ceftriaxone appearing as an alternative drug for meningitis and typhoid.
2007	Amoxicillin now replaced co-trimoxazole for non-severe pneumonia.	Amoxicillin was recommended as primary treatment for non-severe pneumonia while ceftriaxone was first line for meningitis.

5.3.2.1.2 Non measurable self-prescriptions

While 5.2.1.1 describes predicted selection pressure from known clinical prescriptions, it is important to note that there might have been unknown pressure from self-prescriptions. Self-medication has been common in rural Thailand (Osaka and Nanakorn 1996). Based on a survey conducted in 1996 from Thai rural communities: only 42.2% of patients had proper clinical prescriptions from physicians; 37.5% of patients were self-prescribed with antibiotics purchased from local markets; and 9% were reported to “wait and see”. Although choices of drugs

used for self-medication have not been recorded, the drugs available in the markets largely matched those used in the clinic (personal communication from Dr Paul and Claudia Turner). Therefore it is highly likely that this unknown selective pressure is similar to that described from clinical prescriptions in 5.2.1.1.

5.3.2.2 Maela pneumococci have become more resistant to beta-lactams

Ampicillin (which has strong binding affinity to *pbp1a*, *pbp2x* and *pbp2b*), amoxycillin (strong binding affinity to *pbp2b*) and ceftriaxone (strong binding affinity to *pbp2x*) were first prescribed in Maela in 1994, 1999 and 2002, respectively (**Table 5.2)**. The usage of this group of beta-lactams has since become common, especially for amoxicillin. This has exerted a selective pressure with observable genetic signals from the phylogenies of *pbp* genes.

Figure 5.2 represents a phylogeny of each *pbp1a* (top), *pbp2b* (middle) and *pbp2x* (bottom) gene from all genomes in the study (n=3,085). Each phylogenetic tree reveals similar pattern, allowing two observations to be drawn. First, the sparseness of inner ring structure suggests that allelic groups are not exclusively linked to genomic population clusters. This indicates that the alleles have been distributed throughout the population, possibly through recombination. Notably, alleles found in the BC3-NT cluster (highlighted in red in **Figure 5.2**) are the most widely distributed on the tree. Second, the tree contains a mixture of short and long branches with beta-lactam non-susceptibility associated with longer branches. Long branches generally indicate recombination has occurred, thereby suggesting recombination as the main mechanism for acquisition of beta-lactam resistance.

To further test this hypothesis, the association between beta-lactam susceptibility and recombination was measured for the seven dominant clusters (1,126 genomes) and showed that strains that have undergone recombination at either *pbp1a*, *pbp2b* and *pbp2x* appear to be more phenotypically non-susceptible compared to strains without recombination (**Table 5.3**, Fisher's exact test p-value $< 2.20 \times 10^{-16}$).

Long-term consumption of beta-lactam antibiotics implies an on-going selection pressure for resistance. The data show that both recent recombination events (positioned on the external branches of the tree phylogeny) and older recombination events (on the internal nodes) are equally associated with resistance (no significant difference in recombination trend, Fisher's exact test, p-value = 0.6178), suggesting that selection pressure for recombination at these beta-lactam resistance loci has been continuous in the Maela community. Given the high beta-lactam consumption, these results are consistent with what was predicted earlier.

Figure 5.2 Association between recombining *pbp* genes and resistance phenotypes

(a) *pbp1a* gene tree. (b) *pbp2b* gene tree (c) *pbp2x* gene tree. The centre of each diagram shows a SNP-based phylogeny from 3,085 strains rooted on *Streptococcus mitis*. The inner ring is coloured according to dominant population clusters (BC1-7), with the rest of the population appearing in white. The outer ring is coloured according to resistance to penicillin with black and white showing non-susceptibility and susceptibility respectively.

(**Figure is shown on the next page**)

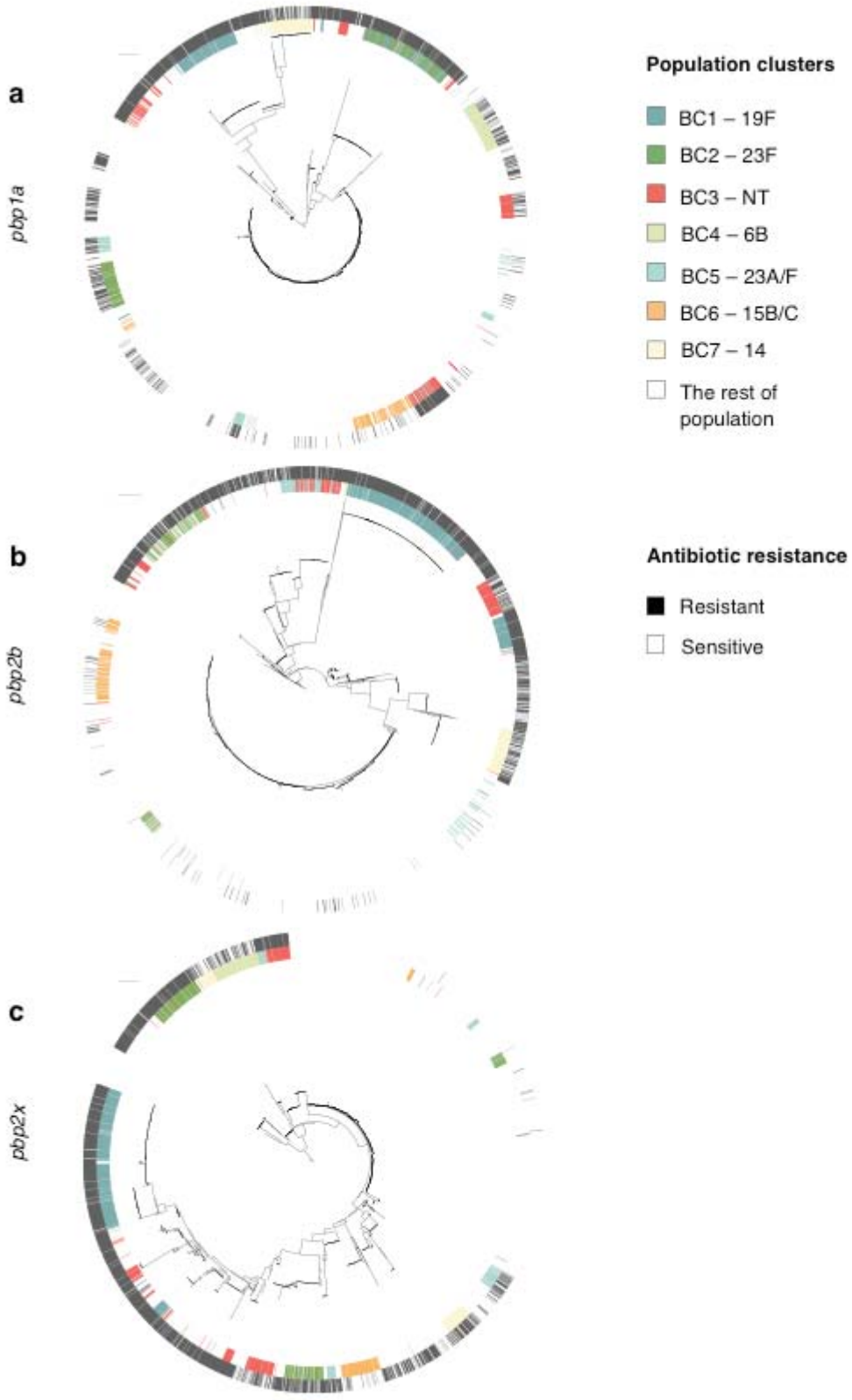

a
pbp1a
b
pbp2b
c
pbp2x
Population clusters
BC1 – 19F
BC2 – 23F
BC3 – NT
BC4 – 6B
BC5 – 23A/F
BC6 – 15B/C
BC7 – 14
The rest of population
Antibiotic resistance
Resistant
Sensitive

Table 5.3 Association between recombination, resistance phenotypes and temporal changes in recombination from seven dominant clusters

Observed phenotypes	No recombination at loci of interest	Recombination at loci of interest	Recent recombination (external node) at loci of interest	Older recombination (internal node) at loci of interest
Beta-lactam resistance: Resistant/sensitive (ratio)	120/146[a] (0.82)	795/150[a] (5.30)	25/6 (4.17)	770/144 (5.35)
Co-trimoxazole resistance: Resistant/ sensitive+intermediate (ratio)	210/28 (7.50)	873/100 (8.73)	10/9[b] (1.11)	863/91[b] (9.84)

[a] Significant difference between beta-lactam resistance phenotypes observed in strains with recombination at *pbp* genes and those without recombination (p-value $< 2.2 \times 10^{-6}$)

[b] Significant difference in cotrimoxazole resistance phenotypes between recent recombination and older recombination at *dhfR* and *folP* genes (p value = 3.49×10^{-5}). Note that the difference is still significant when ratios are grouped by resistant + intermediate/sensitive phenotype (p value = 0.00931)

5.3.2.3 Maela pneumococci have become less resistant to co-trimoxazole

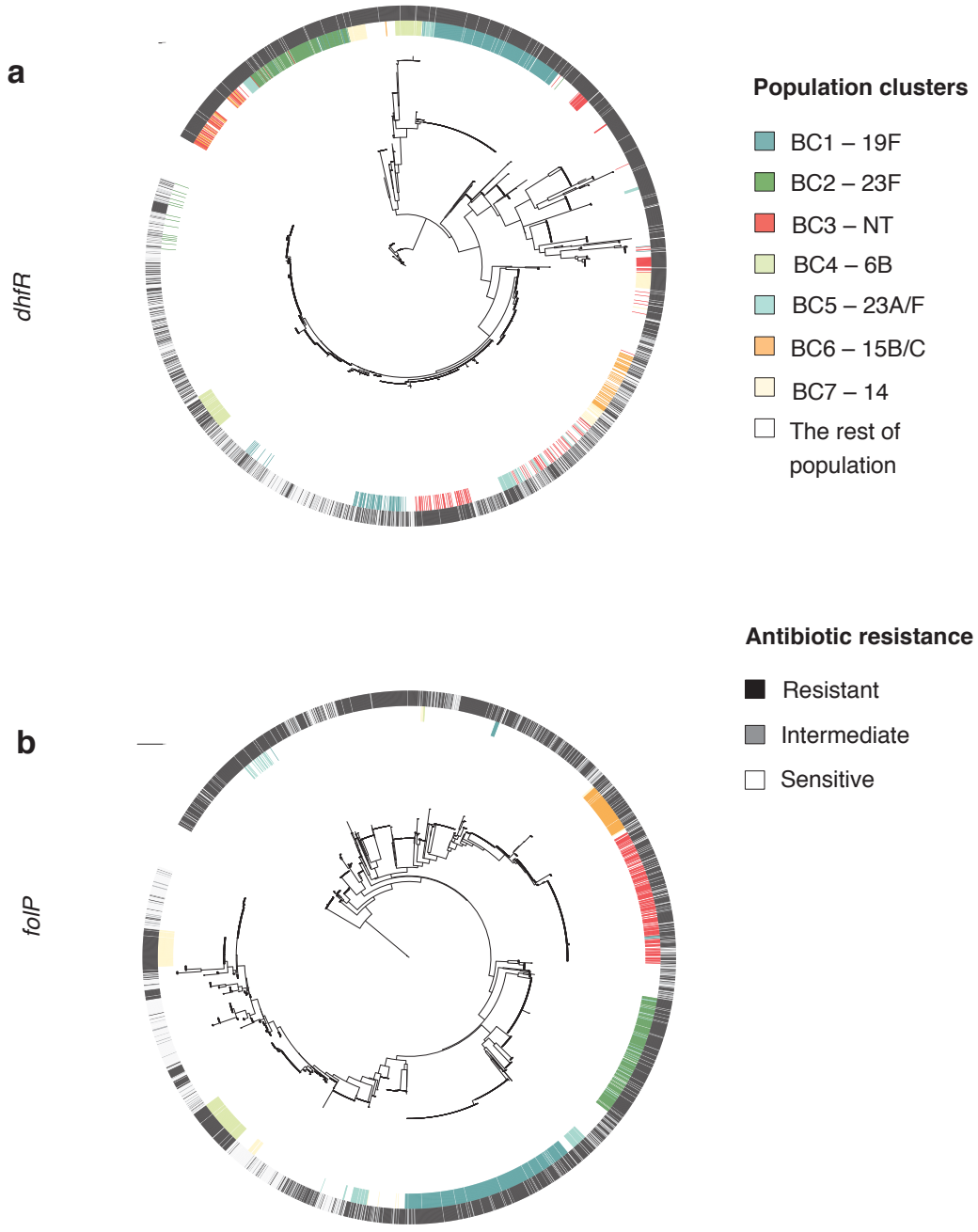
Based on clinical records described earlier, a reduced selection pressure for co-trimoxazole resistant was predicted. The results discussed next potentially provide genetic evidence in favour of this model.

Similar to the *pbp* genes, the phylogeny of *dhfR* (**Figure 5.3**) shows that alleles have been distributed through the population, suggesting a feasible role of recombination in

driving the evolution of *dhfR*. In contrast, the *folP* phylogeny revealed that the alleles were clustered according to their population clusters, suggesting a clonal expansion as the main mechanism for the spread of *folP*. Longer branches in both gene trees appeared to have a weaker association with phenotypic resistance than was seen for the *pbp* genes. Accordingly, there is no association between recombination at *dhfR* and *folP* genes and resistance (**Table 5.3**, Fisher exact test, p-value = 0.2229). It is possible that the lack of association might be due to the acquisition of resistance through nucleotide substitution followed by a clonal expansion. Another possible explanation could be that the signal is distorted due to changes in selective pressure over time. More specifically, the signal for positive selection for resistance before 2002 could be masked by a lack of strong pressure for resistance alleles thereafter (**Table 5.2**). A lack of selection pressure for resistance genes means that all alleles can be neutrally incorporated into the genome via recombination. This would be reflected by a significant difference in trend between recent and past recombination events, which can be seen (**Table 5.3,** Fisher exact test, p-value = 3.49×10^{-5}). While recent recombination events (on external branches) were associated with sensitive/intermediate susceptibility phenotypes, older recombination events (internal nodes) were associated with fully resistant phenotypes. This provides evidence that reduction in the use of an antibiotic may lead to a lower requirement for the isolates to be resistant. However, the magnitude by which the resistant phenotypes were observed to decrease after a decrease in co-trimoxazole usage is likely due to the cost of being resistant.

Figure 5.4 Association between recombining *fol* genes and resistant phenotypes.

(a) *dhfR* gene tree, (b) *folP* gene tree. The centre of each diagram represents a SNP-based phylogeny from 3,085 strains rooted on *Streptococcus mitis*. The inner ring is coloured according to the membership of the seven dominant population clusters (BC1-BC7). The outer ring is coloured on the basis of co-trimoxazole resistance phenotypes (white, sensitive; grey, intermediate; black, resistant). (**Figure is shown on the next page**)


a
dhfR
Population clusters
BC1 – 19F
BC2 – 23F
BC3 – NT
BC4 – 6B
BC5 – 23A/F
BC6 – 15B/C
BC7 – 14
The rest of population
b
folP
Antibiotic resistance
Resistant
Intermediate
Sensitive

5.4 Conclusion

The rise of resistant pneumococci in clinics has been observed since the late 1970s following the use of antibiotics (Kislak and Razavi *et al.* 1965, Hansman 1975). Moreover, high-level cephalosporin-resistant strains were observed in areas where the use of new-generation cephalosporins, which are a class of beta-lactams, was encouraged (McDougal and Rasheed *et al.* 1995, Smith and Botha *et al.* 2001). Together, these provided evidence that the use of antibiotics has driven the emergence of antibiotic resistance pneumococci (Hakenbeck and Bruckner *et al* 2012). Genetic culprits for the rise of resistance were soon after identified. Highly altered penicillin-binding protein genes, also known as mosaic structures, were observed among penicillin resistant pneumococci. The mosaic patterns were caused by recombination, thereby highlighting the role of recombination in spreading beta-lactam resistance (Dowson and Hutchison *et al.* 1989, Laible and Spratt *et al.* 1991). The genetic cause of co-trimoxazole resistance was also identified (Adrian and Klugman 1997, Padayachee and Klugman 1999, Silver 2007), but it was less clear which mechanism help disseminate the resistance alleles.

The depth of genomic sampling in this study allowed the identification of highly exchanged genes, which further led to the association of these genetic signals with antibiotic resistance patterns, and clinical records on antibiotic consumption. The results were consistent with the use of antibiotics and resistance patterns described previously. Genetic signals associated with antibiotic resistance could be linked to trends in antibiotic consumption. A reduction in co-trimoxazole consumption over time was reflected in the patterns of recombination, with more recent events having a weaker association with resistance phenotypes. On the other hand, a steady increase in the consumption of beta-lactams resulted in continuous acquisitions of resistance alleles via recombination.

The analyses also highlighted the remarkable similarity of sites under selection between dominant clusters, suggesting some uniformity in response to common selection pressure across the species. Aside signals related to antibiotic resistance,

recombination hotspots are also functionally linked to host immunity, reflecting the importance of this force in the evolution of *Streptococcus pneumoniae*.

Overall, this chapter highlights the role of recombination in mediating rapid adaptation of pneumococci. The nontypable clusters were described as a hub of gene flow in chapter 4. Therefore, it is plausible that the nontypable strains may be a principal driver of adaptation for a wider population, especially for antibiotic resistance. However, this role in mediating antibiotic resistance cannot be established without looking at the actual distributions of resistance alleles across the population. This will be explored in the next chapter where potential resistance markers for beta-lactams and their distributions in the wider population will be determined.

Chapter 6: Genome-wide association study identifies single nucleotide polymorphic changes associated with beta-lactam resistance

Introduction

Results

6.1 Identification of loci associated with beta-lactam non-susceptibility

- 6.1.1 Quality control and randomisation test
- 6.1.2 Population stratification
- 6.1.3 Case-control association analysis
- 6.1.4 Linkage disequilibrium
  - 6.1.4.1 Defining linkage blocks
  - 6.1.4.2 Larger linkage blocks observed in smaller dataset
- 6.1.5 Estimating the amount of the non-susceptible phenotype explained by co-detected loci in Maela and Massachusetts populations

6.2 Biological relevance of candidate loci

- 6.2.1 Candidate loci in genes participating in the peptidoglycan biosynthesis pathway
- 6.2.2 Candidate loci in genes outside the peptidoglycan biosynthesis pathway
- 6.2.3 Candidate loci in genes conferring resistance to other antibiotics

6.3 Beta-lactam specificity of resistance mutations

6.4 Distribution of candidate alleles in the Maela and Massachusetts populations

Conclusion

Declaration of work contributions:

Dr Nicholas Croucher kindly provided sequences for Massachusetts data.

6. Genome-wide association study identifies single nucleotide polymorphic changes associated with beta-lactam resistance

6.1 Introduction and aims

The previous chapter highlighted an increase in beta-lactam resistance and the mechanisms mediating the spread of resistance alleles in one local community. At the global level, a rise in beta-lactam resistant pneumococci over the recent decades has raised significant concerns (Potgieter, Carmichael *et al.* 1992, Carratalà, Alcaide *et al.* 1995, Donern amd Ferraro *et al.* 1996, and more recently WHO 2014). There have been tremendous efforts in identifying the genetic sources of resistance between 1980s-2000s. Comparative genomics between beta-lactam resistant and susceptible strains have identified highly variable regions in genes coding for penicillin-binding proteins (*pbp*) in resistant isolates; thereby highlighting *pbp* genes as key contributors to the resistance phenotype (Laible, Spratt *et al.* 1989, Dowson, Hutchison *et al.* 1989). Aside from determining nucleotide and amino acid alterations observed in beta-lactam resistant isolates, site-directed mutagenesis has been a useful tool in identifying mutations that give rise to resistant profiles (Laible and Hakenbeck 1991, Hakenbeck, Martin *et al.* 1994, Sifaoui, Kitzis *et al.* 1996, Smith and Klugman 1998 & 2003). Many of these studies were supported by further structural characterisations, revealing a change in the structure of penicillin binding proteins which lead to resistant phenotype (Gordon, Mouz *et al.* 2000, Dessen, Mouz *et al.* 2001, Job, Di Guilmi *et al.* 2003, Contreras-Martel, Job *et al.* 2006)

Identification of specific variants that might be associated with beta-lactam resistance can also be performed using a technique called genome-wide association study (GWAS). Although GWAS have been common in human genetics for many years (McCarthy, Abecasis *et al.* 2008), it was, at the time of this study, largely untried in bacteria because of the limited sample sizes and the confounding effects of bacterial clonal population structure. The clonal population structure may be less problematic in pneumococci as homologous recombination brings genetic admixture into pneumococcal populations in a manner akin to sexual reproduction in humans. Unlike human recombination, this does not occur every generation and only affects a small

part of the genome in each occurrence. On average, recombination in *S. pneumoniae* involves 2.3 kb of chromosomal DNA (Croucher, Harris *et al.* 2012), about twice the size of an average pneumococcal gene. This suggests that large numbers of recombination events must accumulate in order to disrupt the clonal structure, or break up haplotype blocks smaller than this size. Also, on average, one recombination event occurs every one to nine years, depending on the lineages. Therefore, to observe sufficient numbers of recombination events to disrupt clonal structure, a diverse and large population size would be required. The large size of the species-wide samples used in this study (3,071 pneumococcal genomes), and the highly recombinogenic nature of *S. pneumoniae* had the potential to allow sufficient resolution to be achieved and help refine the genetic determinants of resistance from large recombination fragments, previously described as mosaic genes, to discrete causative sites or smaller linkage blocks.

This gave an opportunity to investigate the molecular mechanisms of resistance down to single polymorphic changes. Ultimately, an increase in the resolution of detection and improved insights into the biology of resistance mechanisms for beta-lactams might contribute to the foundation for future application of genome sequencing in predicting antibiotic sensitivity in clinical settings and surveillance studies (Levine, O'Brien *et al.* 2012, Goldblatt, Ramakrishnan *et al.* 2013, GPS 2013).

This chapter aimed at:
i) Identifying genetic variants associated with beta-lactam non-susceptibility.
ii) Determining differential association of variants identified in i) to particular class of beta-lactams.
iii) Determining the distribution of resistant variants in different pneumococcal lineages including vaccine, and non-vaccine targeted lineages.

6.2 Methods

6.2.1 Subject populations

Two currently largest datasets for which whole genome sequences and beta-lactam susceptibility phenotype were available – Maela (3,085 isolates, also the subject of this thesis) and Massachusetts (616 isolates, Croucher, Finkelstein *et al.* 2013) - were employed in this analysis and the results were used to cross-validate each other.

6.2.2 Genotype callings and quality control

Bases were called from mapped sequences as discussed in 2.5, resulting in 392,524 and 198,248 SNP calls from the Maela and Massachusetts data respectively. This haploid bacterial information was handled as human mitochondrial sequence in PLINK v. 1.07 (Purcell, Neale *et al.* 2007). Since many thousands of genotypes are generated, a small genotyping error can lead to spurious GWAS results. Quality control thus is a critical step in performing GWAS. Here, minor allele frequency, which represents very low frequency alleles that likely reflect genotyping errors and proportion of missing genotypes per strain (genotype missingness rate) were estimated. Variants with minor allele frequency < 0.01, missingness by strain > 0.1 and missingness by variants > 0.1 were excluded from the analysis. For each site, the top two most common variants were parsed to the next analysis to reduce complexity in the test statistic.

6.2.3 Phenotype information

Beta-lactam susceptibilities were determined in both datasets by disk diffusion following the CLSI 2008 guidelines (CLSI 2008), generating 1,501 non-susceptible, 1,568 susceptible and 16 unknown phenotypes in the Maela data; 228 non-susceptible, 383 susceptible and 5 unknown phenotypes from the Massachusetts data. The minimum inhibitory concentrations (MIC) of non-susceptible isolates were confirmed by the E-test method as described in 2.1.1 and (Croucher, Finkelstein *et al.* 2013).

6.2.4 Determining the cut-off threshold

Randomisation tests were performed for both Maela and Massachusetts population to estimate the level of intrinsic noises generated by genetic variations alone as well as to determine a suitable cut-off for the analysis. 100 GWAS permutations were run with true genotypes but randomised binary phenotypes. With Bonferroni correction for multiple testing, there were no significant associations observed beyond the p-value 0.01. Therefore it was selected as a conservative threshold in the study reported here.

6.2.5 Case-control analysis

Variants associated with beta-lactam resistance were determined based on binary phenotypes: susceptible or non-susceptible conditioned on the pneumococcal population structure. BAPS clustering (generated in 2.6.2.2 and (Croucher, Finkelstein *et al.* 2013)) were used to represent population structure in the analysis. Based on known cluster information, the Cochran-Mantel-Haenszel (CMH) test for 2x2xK binary phenotype x variants | population cluster was employed with sites corrected for multiple testing using the Bonferroni correction at a p-value of 0.01.

6.2.6 Linkage analysis

Linkage disequilibrium was explicitly tested using Haploview (Barrett, Fry *et al.* 2005), which was devised for human genetics. However, bacterial recombination is not equivalent to human crossing over where linkage over long distance can be ignored. Therefore, the human genetics tool used (Barrett, Fry *et al.* 2005) would ignore any pairwise comparison over 500 kb. Here, the default setting was adjusted so that the tool considered all pairwise comparisons under 2,200 kb, which is equal to the size of the whole genome. The information was treated as male human X-chromosome to retain its haploidy, thereby incorporating all possible linkage predictions into our analysis. Using 95% confidence intervals as described in (Gabriel, Schaffner *et al.* 2002), a linkage block was identified as a region within a low recombination rate (here referred to as linkage loci). Physical linkage size detected in Maela and Massachusetts data were compared to illustrate the effect of population size on the power for separating causative SNPs from linked SNPs.

6.2.7 Estimation of percentage of resistance in the population explained by candidate loci

High-stringency SNPs co-detected in both Maela and Massachusetts datasets were used to cross-predict resistance in each population separately. The proportion of resistance in the population that could be explained by co-predicted SNPs (here grouped into linkage loci) were plotted for each of the test populations. The order of loci added was permutated to accommodate all possible combinations.

6.2.8 Specificity to different classes of beta-lactams

To test whether or not there were SNPs conferring more specific resistance to certain classes of beta-lactam antibiotics, GWAS was repeated on the SNPs co-detected in both populations. The binary phenotypes were replaced with continuous phenotypes, penicillin MIC values and ceftriaxone MIC values. P-values calculated from penicillin MIC and ceftriaxone MIC for each SNP were grouped by the linkage structure computed as discussed above.

6.2.9 Prevalence of candidate loci in the population

For each BAPS cluster in both the Maela and Massachusetts data, the mean prevalence of candidate loci was calculated by averaging the frequency of linked SNPs detected in each locus per cluster size.

6.3 Results

6.3.1 Identification of loci associated with beta-lactam non-susceptibility

Genome-wide association study (GWAS) is a broadly used approach in human genetics to identify SNPs associated with complex diseases, ranging from cancer to mental health (Sullivan, Daly *et al.* 2012, Goldstein, Allen *et al.* 2013, Pharoah, Tsai *et al.* 2013). The test compares SNPs across a large population including individuals with and without the disease. GWAS reports SNPs enriched in the disease population (case) but absent in healthy population (control) as potential risk factors indicating that the individual that carries the risk alleles is more likely to develop the disease.

Similar to these studies conducted in humans, SNPs across beta-lactam susceptible and non-susceptible pneumococcal populations were compared here. GWAS was independently performed with genetic variants called from whole genome alignments of 3,085 pneumococcal strains collected from a carriage cohort in Maela and 616 strains from a carriage cohort in Massachusetts (Croucher, Finkelstein *et al.* 2013). The binary phenotypes, which are based on susceptibility and non-susceptibility to beta-lactams were determined using the Clinical and Laboratory Standard Institute guidelines (CLSI, 2008). Strains with penicillin minimum inhibitory concentration (MIC) $\leq$ 0.06 μg/ml are classified as susceptible; applying these cut-offs across the Maela and Massachusetts data gave 1,729 non-susceptible (case) and 1,951 susceptible (control) samples for performing GWAS (with 21 unknown). This section discusses how GWAS was performed, including essential corrections needed to minimise false positive rates. The result reports both discrete and linked genetic variants (here called loci) associated with beta-lactam non-susceptibility in the Maela and Massachusetts pneumococcal populations.

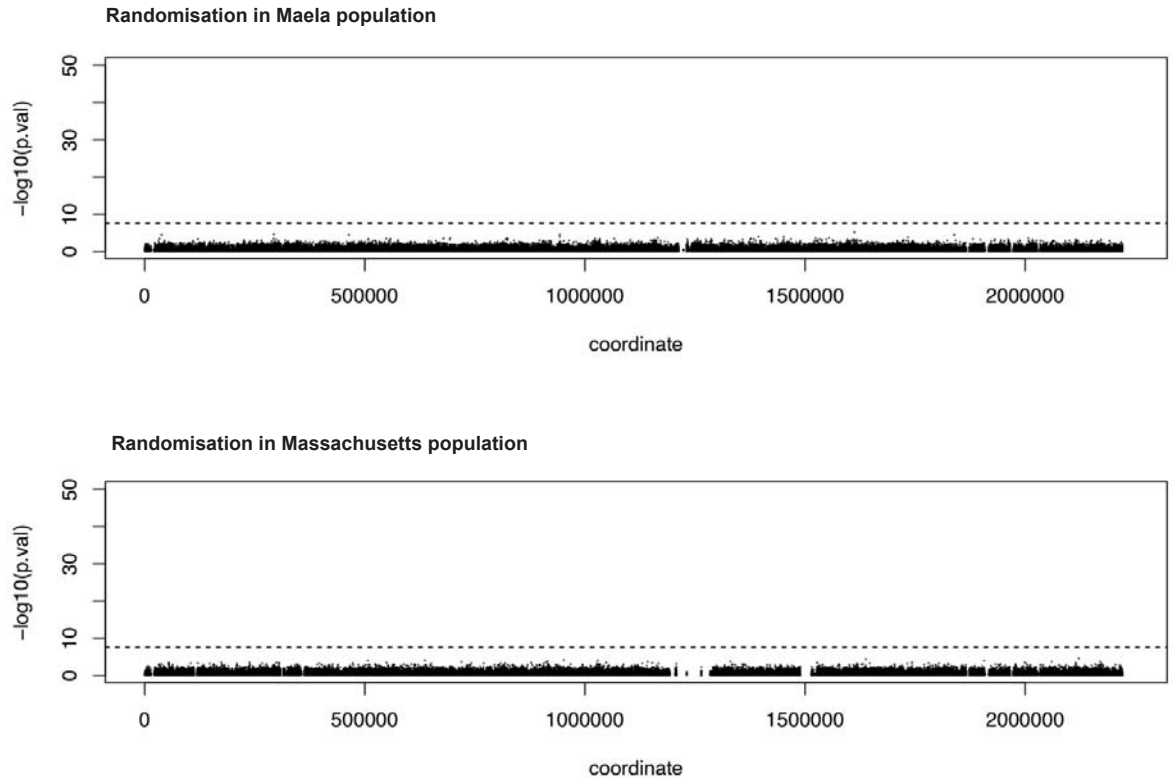
6.3.1.1 Randomisation test

Following quality control and data filtering, the basal intrinsic noise caused by genetic variation was estimated. As this noise can lead to false positives, it is essential to draw the result cut-off above the signals generated by noise. For each Maela and

Massachusetts data set, 100 GWAS permutations, with true genotypes but randomised binary phenotypes, were performed using PLINK (Purcell, Neale *et al.* 2007). Manhattan plots (**Figure 6.1**) summarise the randomised results from Maela (top) and Massachusetts data (bottom). None of the permutations of either datasets achieved significant association at p-value 0.01 with a Bonferroni correction for multiple testing, indicating a stringent threshold. Consequently, a significance p-value of 0.01 was applied throughout the study.

Figure 6.1 Randomised control for intrinsic noise based on genetic variation alone

Manhattan plots demonstrate no significant associations in either Maela or Massachusetts data using real genotype data and randomised resistance phenotype assignments. Horizontal dotted lines mark the cut-off with Bonferroni correction at p value = 0.01.



6.3.1.2 Population stratification

GWAS is sensitive to bias induced by population stratification. The test statistic is based on the assumption of independent observations. However, this is often violated as in humans; the cases may be overrepresented in a certain group of the population compared to the others (Wray, Yang *et al.* 2013), and in bacteria, the case population may be a part of the same clonal complex (Sheppard, Didelot *et al.* 2013). This leads to a true association locus being confounded by the underlying population structure, resulting in excessive false positive discoveries. As a consequence, population stratification is a necessary step in GWAS to make the studies consistent and replicable.

The Maela and Massachusetts populations consist of strains from a species-wide samplings; they respectively represent 277 and 154 known multilocus sequence types. Their population structures were predefined based on whole genome sequence similarity. The Bayesian based software BAPS was used to estimate the structure in both populations. This resulted in 33 and 16 initial clusters for the Maela and Massachusetts data, respectively. Due to the large sample size of the Maela data set, BAPS was additionally run in a hierarchical manner, generating secondary clusters within each primary cluster. These secondary clusters were used to represent the Maela population structure.

Based on this clustering information, the Cochran-Mantel-Haenszel (CMH) association statistic was employed to test for association between beta-lactam non-susceptibility and specific variants, conditioned on the population clusters. The Bonferroni correction for multiple testing at p-value of 0.01, as discussed in 6.1.1, was used as the cut-off. The reduction in false positive rates after correction for underlying population structure was estimated by a parameter called genomic inflation factor. The inflation factor is defined as the ratio of observed distribution of the test statistic to the expected mean, thereby allowing the extent of inflation and false positive rate to be quantified (Devlin and Roeder 1999). A high inflation factor typically indicates a high rate of false positives where associations are influenced by population structure. The application of CMH test with clustering condition reduced

the inflation factors from 80.16 (mean chi-squared statistic = 68.99) to 2.56 (mean chi-square statistic = 3.05) in the Maela data, and 13.18 (mean chi-square statistic = 14.17) to 3.76 (mean chi-square statistic = 4.73) in the Massachusetts data. The decreases in genomic inflation factors in both populations indicate lower false positive rates due to underlying population structure. However, these observed inflation factors are still relatively high compared to GWAS conducted on the human nuclear genome, suggesting that intrinsic clonal population structure remains an issue for bacterial association studies.

6.3.1.3 Case-control association analysis

The CMH test described in 6.1.2 was performed, giving 858 and 1,721 SNPs associated with beta-lactam non-susceptibility in the Maela and Massachusetts populations, respectively (**Figure 6.2**, SNPs tabulated in **Appendix C - D**). Among these, 301 SNPs were found to be associated with non-susceptibility in both populations (**Figure 6.3**, tabulated in **Appendix E**). Given that the two data sets have different population structures, which have evolved independently, these co-detected SNPs represent a set of candidates in which one can have more confidence. The 301 co-detected SNPs consist of three SNPs localised in intergenic regions, and 298 SNPs found in coding sequences. The latter can be further divided into 71 non-synonymous and 227 synonymous SNPs. The detection of non-synonymous SNPs implies a functional effect which might contribute to beta-lactam non-susceptibility. Synonymous SNPs, on the other hand, might not play a causative role but could be tightly linked to causative SNPs with insufficient recombination in the data set to separate the link (here called “hitchhiking” SNPs), and therefore form part of the same haplotype block. These linkage structures and the limitation they imposed on the predictions will be explored in the next section.

Figure 6.2 Summary of the genome-wide association study conducted in two separate datasets

Manhattan plots summarise the association of whole-genome SNP variant with beta-lactam susceptibility in the Maela and Massachusetts data as well as particular gene regions which show strong associations. Top panel represents the statistical significance of association (y-axis) for each variant arranged in order on the genome (x-axis) in the Maela (red) and Massachusetts (blue) data. Horizontal dotted lines in both top and bottom panels indicate a significance cut-off after Bonferroni correction of p value = 0.01. Genes with significant associations are annotated on top. Genes coding for penicillin binding proteins: *pbp2x, pbp1a,* and *pbp2b*, whose roles in beta-lactam resistance are well characterised, are highlighted in grey. Bottom panel expands the view of penicillin binding protein genes where most of the significant associations are detected: from left to right *pbp2x, pbp1a,* and *pbp2b*. Protein domains identified within these genes are shaded in pale grey and labelled. The vertical dotted lines represent the active sites of the transpeptidase domain. Plus signs denote synonymous SNPs and dots denote non-synonymous SNPs.

(**Figure is shown on the next page**)

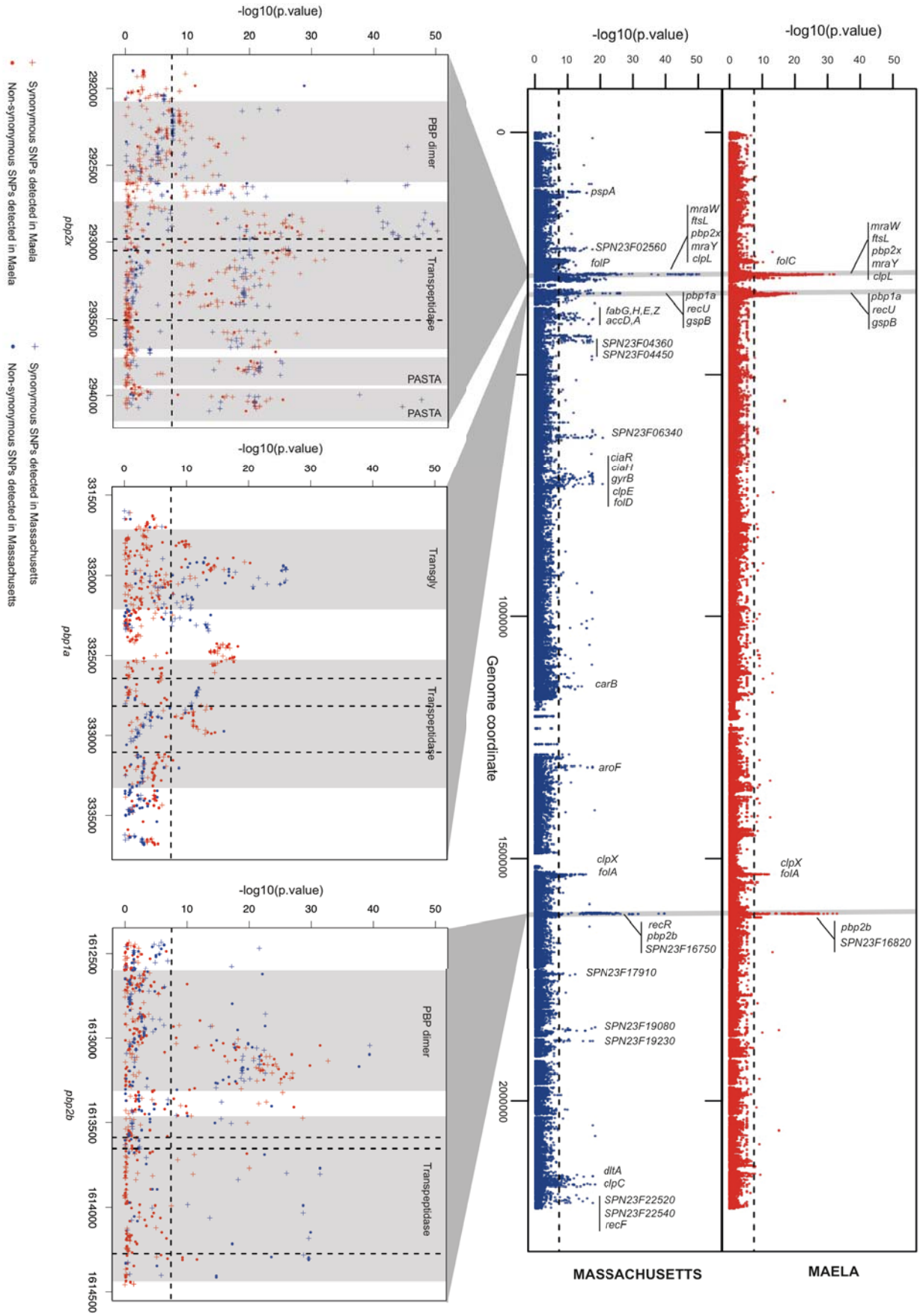
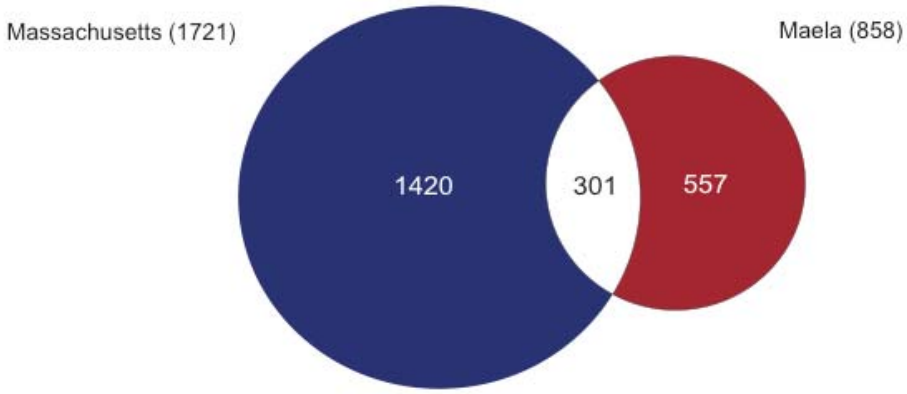

Synonymous SNPs detected in Maela
Non-synonymous SNPs detected in Maela
Synonymous SNPs detected in Massachusetts
Non-synonymous SNPs detected in Massachusetts
-log10(p.value)
pbp2x
PBP dimer
Transpeptidase
PASTA
PASTA
pbp1a
Transgly
Transpeptidase
pbp2b
PBP dimer
Transpeptidase
Genome coordinate
MASSACHUSETTS
MAELA
pspA
SPN23F02560
folP
mraW
ftsL
pbp2x
mraY
clpL
fabG,H,E,Z
accD,A
pbp1a
recU
gspB
SPN23F04360
SPN23F04450
SPN23F06340
ciaR
ciaH
gyrB
clpE
folD
carB
aroF
clpX
folA
recR
pbp2b
SPN23F16750
SPN23F17910
SPN23F19080
SPN23F19230
dltA
clpC
SPN23F22520
SPN23F22540
recF
folC
SPN23F16820

Figure 6.3 Summary of single nucleotide polymorphisms (SNPs) associated with beta-lactam non-susceptibility

A Venn diagram summarises the number of SNPs reaching significance in each of the Maela and Massachusetts datasets, and those that are co-detected in both.



6.3.1.4 Linkage disequilibrium

6.3.1.4.2 Defining linkage blocks

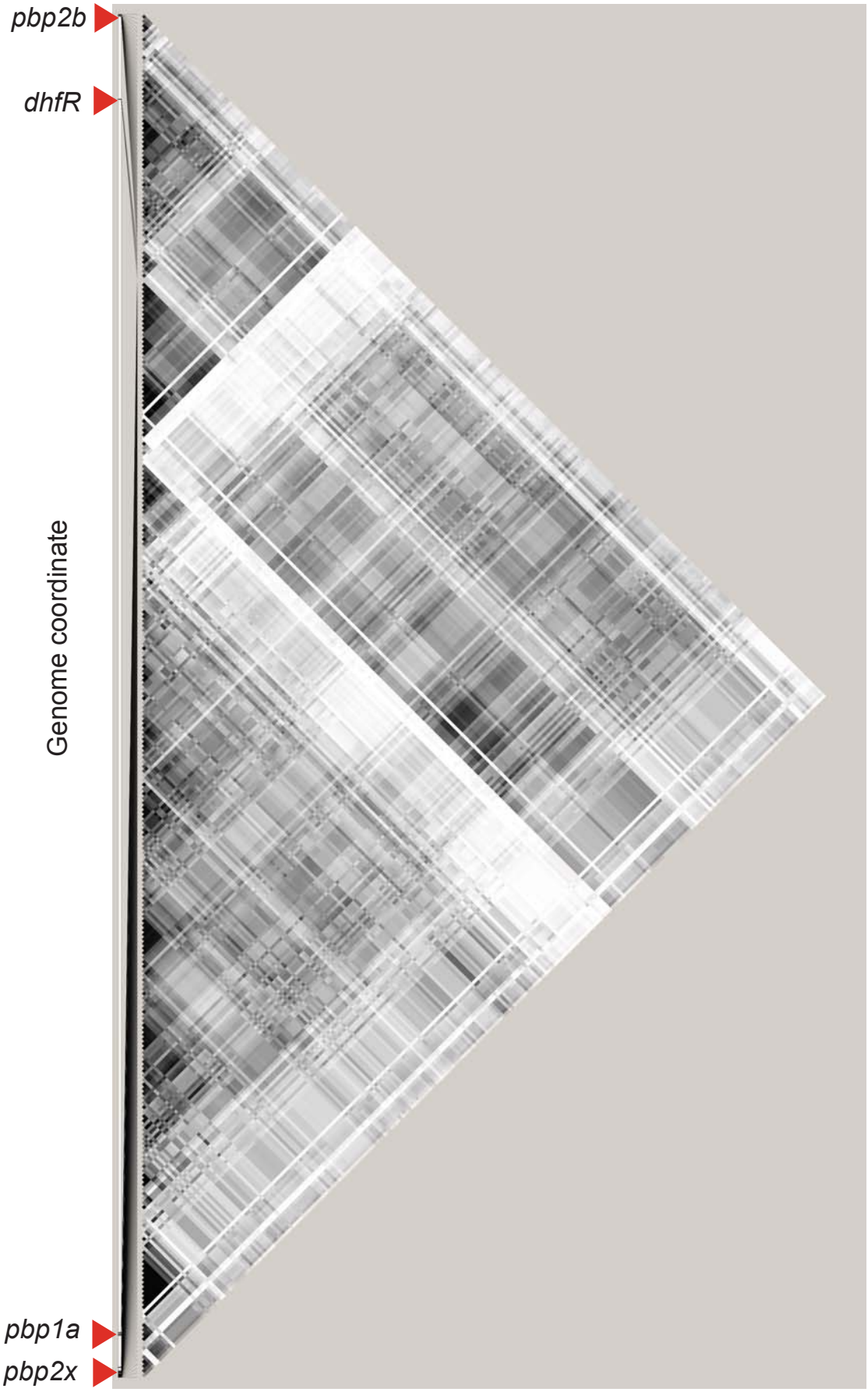
Linkage between candidate SNPs reported in 6.1.3 was explicitly tested using Haploview (Barrett, Fry *et al.* 2005). In humans, linkages between distant sites are disrupted by crossing-over, the recombination process where homologous chromosomes pair up and exchange different segments of their genetic material. However, bacterial recombination does not necessarily break long distance linkage. Therefore, Haploview was set to consider all possible pairwise comparisons over the entire size of *S. pneumoniae* genome (approximately 2,200 kb). This additional

setting allowed the application of Haploview, which was devised for human genetics, to be used in bacteria. Using 95% confidence bounds as described in (Gabriel, Schaffner *et al.* 2002), a haplotype blocks were identified as regions with a low recombination rate. Linkage information for SNPs detected in Maela and Massachusetts are listed in **Appendix C** and **D**, respectively. For SNPs co-detected in both Maela and Massachusetts populations, 51 linked loci were detected (**Figure 6.4)**. Among these, nine were single SNPs and 42 were in linkage blocks of between two and 19 SNPs, of which 12 contain only a single non-synonymous SNPs (**Appendix E**).

Figure 6.4 Linkage analysis for SNPs co-detected in two separate datasets

The Haploview plot illustrates linkage disequilibrium (r^2) between co-detected SNP candidates from the Maela and Massachusetts datasets. The bar of the left represents the genome position of the SNPs, connected by lines to the diamond plot on the right. A complete black diamond represents complete linkage disequilibrium between candidate SNPs (r^2=1), while a white diamond represents a perfect equilibrium (no linkage) (r^2 = 0)

(**Figure is shown on the next page**)


pbp2b
dhfR
Genome coordinate
pbp1a
pbp2x

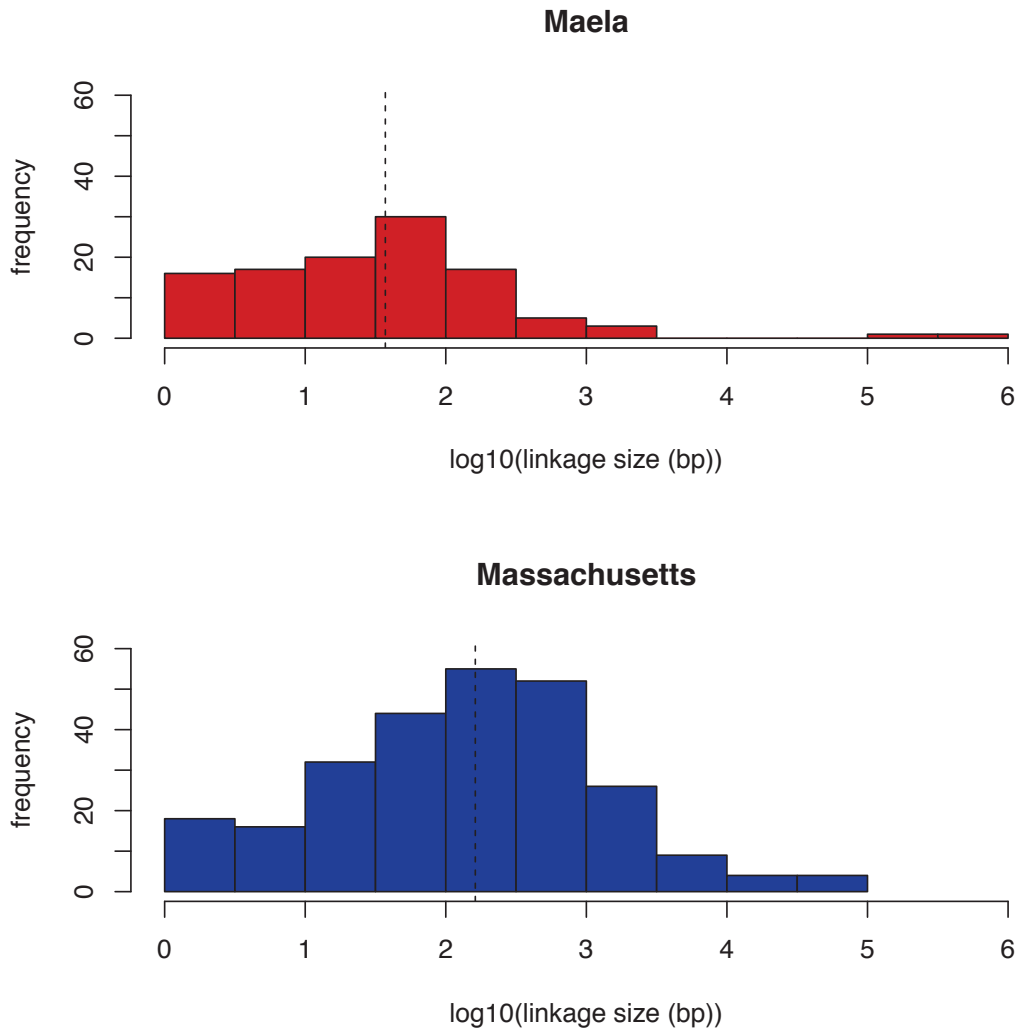
6.3.1.4.2 Larger linkage blocks observed in smaller dataset

As noted earlier, a total of 858 and 1,721 SNPs were detected in Maela and Massachusetts populations which comprise 3,085 and 616 strains respectively. The result was rather counter-intuitive, as one would expect higher number of candidate SNPs to be detected from the larger sample set of Maela than Massachusetts, which was not observed here. This could be due to the fact that population structures in the two settings were independently determined in previous works (discussed in 3.3.1.1 and Croucher, Finkelstein *et al.* 2013). Therefore, it is possible that the clustering information from the two data sets is not equivalent in their stringency. This subsequently leads to more strict control over population stratification in one population than the other.

Another potential explanation is that the Maela and Massachusetts data sets have different linkage structure. The sizes of the linkage blocks detected from the two populations in 6.1.4.1 were compared. Indeed, linkage block sizes detected in the Maela data were significantly smaller than the Massachusetts data (Mann Whitney test p-value 6.53×10^{-9}). Shown in **Figure 6.5**, the medians of Maela and Massachusetts linkage blocks were 37.5 bp and 165 bp respectively, presenting an inverse relationship between the size of the linkage structure and the sample size. This suggests that many of the candidate SNPs detected in the Massachusetts data are potentially hitchhikers, thereby resulting in greater false positives. Such observation possibly reflects the limitation of small data sets where there might not be enough recombination detection to sufficiently break the linkage structure. However, longer linkage blocks may also be expected from the Massachusetts population where carriage rates are lower, potentially reducing the frequency of opportunities for the strains to recombine.

Figure 6.5 Summary of physical linkage structure in two separate datasets

Size of linkage detected in the Maela (red) and Massachusetts (blue) association studies were plotted as histograms on log10 scale. Vertical dotted lines mark the median size of haplotype blocks that harbour candidate SNPs (37.5 bp in Maela data and 165 bp in Massachusetts data)



6.3.1.5 Estimating the amount of the non-susceptible phenotype explained by co-detected loci in the Maela and Massachusetts population

To estimate how much of the phenotypic resistance in the samples could be explained by the identified SNPs, cross-prediction tests were performed using only the SNPs co-

detected in both Maela and Massachusetts association studies. The co-predicted SNPs, grouped by their linkage structure, were tested back against each population. The results (**Figure 6.6**) show that close to 100% of the resistance in each population could be explained by all of the co-detected SNPs.

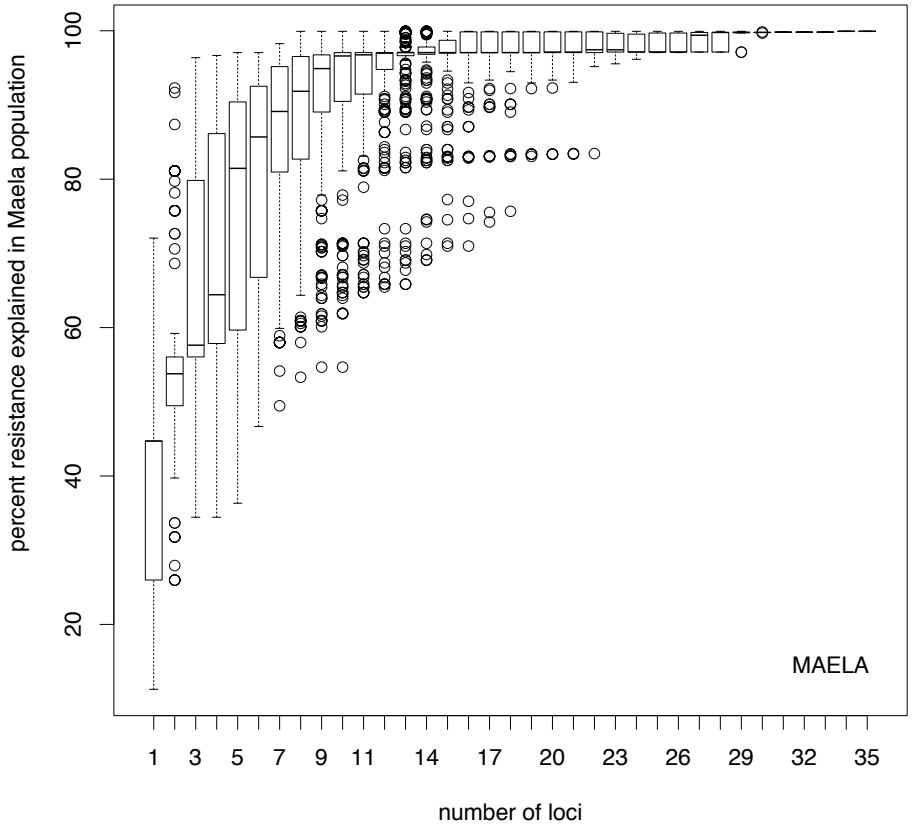
Unlike human polygenic traits where each locus contributes only a small effect on the phenotype, each of these bacterial loci appears to have much stronger effect. This is consistent with experimental characterisations where introductions of a single variant can sometimes lead to a change in pneumococcal beta-lactam susceptibility (Hakenbeck, Bruckner *et al.* 2012). This can be demonstrated using odds ratios, which indicate the size of the effect of each associated SNP. While human GWAS report a median odds ratio of 1.33 per SNP (Ku, Loy *et al.* 2010, Manolio 2010), this analysis gave a median odds ratio of 11.09 per SNP, indicating a stronger effect size.

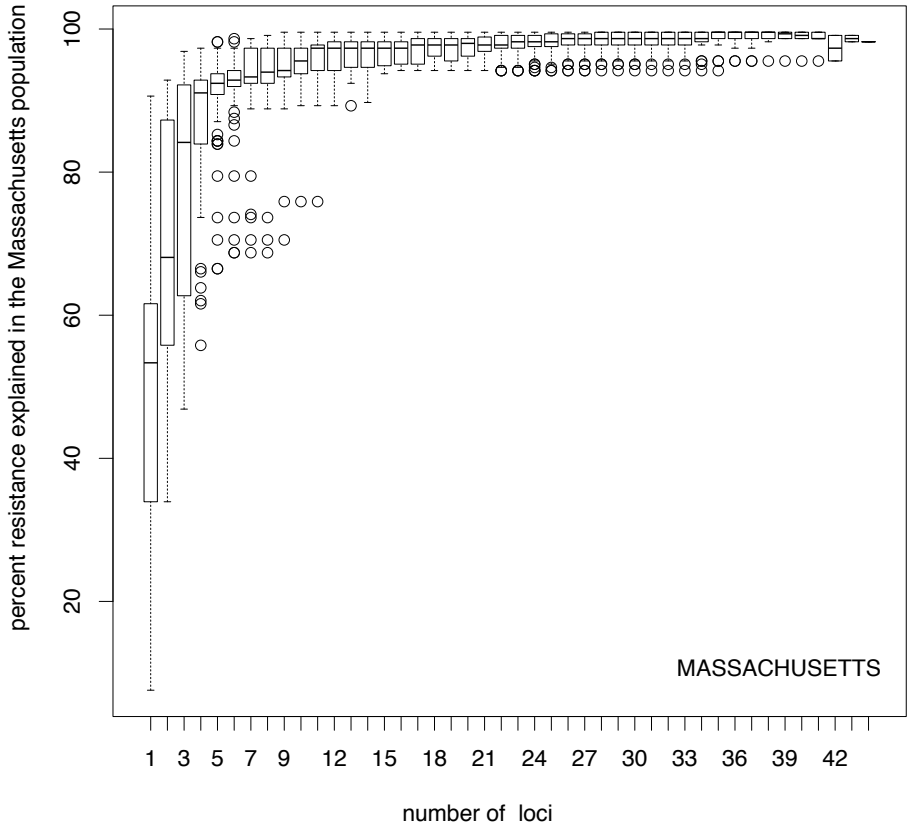
For both the Maela and Massachusetts populations, the percentage of resistance explained plateaued after the addition of approximately 10 loci in any order. This suggests that, at most, about 10 loci are required to make a susceptible strain non-susceptible and this can be done through multiple different combinations. However, in each resistant isolate, combinations of more than ten loci are commonly detected, perhaps indicating that not all loci are involved in conferring resistance. Some may play a compensatory role in reducing the fitness cost of resistance variants. In total, the co-detected variants are present in 100% and 98% of the Maela and Massachusetts resistant strains respectively, highlighting that a large proportion of possible resistance variants were captured in this analysis.

Figure 6.6 Percentage of the non-susceptible phenotype explained by co-detected loci in the Maela and Massachusetts populations

The plots represent proportions of resistance in the population (y-axis) explained by all combinations of increasing numbers of co-detected loci (x-axis), based on combinations of loci observed from both Maela and Massachusetts data.

(**Figure is shown on next page**)


percent resistance explained in Maela population
20
40
60
80
100
MAELA
1 3 5 7 9 11 14 17 20 23 26 29 32 35
number of loci


percent resistance explained in the Massachusetts population
20
40
60
80
100
MASSACHUSETTS
1 3 5 7 9 12 15 18 21 24 27 30 33 36 39 42
number of loci

6.3.2 Biological relevance of candidate loci

Following the identification of potential candidate loci associated with beta-lactam non-susceptibility, this section tries to explain the finding in 6.1 in biological context.

For both population settings, candidate loci show a higher enrichment in genes compared to intergenic regions than would be expected by chance (Fisher's Exact Test p-value > 0.0001). These loci are not randomly distributed across the whole genome, but clustered within certain genes (**Figure 6.2**). Loci co-detected in both populations are localised in genes participating in the peptidoglycan biosynthesis pathway, including penicillin binding proteins (*pbp2x*, *pbp1a*, *pbp2b*), two transferases required for cell wall biogenesis (*mraW*, *mraY*), the cell division pathway (*ftsL*, *gpsB*), heat shock protein and chaperones (*clpL*, *clpX*), the recombination pathway (*recU*) and a metabolic gene known to be involved in resistance to co-trimoxazole (*dhfR*). Some of these sites, particularly in the *pbp* genes, matched those previously reported to play an important role in beta-lactam resistance in the literature (**Appendix E**, providing independent validations to the methodology and some of the results. To my knowledge, out of 71 non-synonymous SNPs reported here, 43 SNPs are novel and potentially contribute to beta-lactam non-susceptibility in addition to those identified in previous studies.

6.3.2.1 Candidate loci in genes participating in the peptidoglycan biosynthesis pathway

Since beta-lactam antibiotics work by inhibiting cell wall biosynthesis, it is not surprising to observe significant associations between non-susceptible phenotypes and variants in genes participating in the peptidoglycan biosynthesis pathway, including *pbp2x*, *pbp1a*, *pbp2b*, *mraW* and *mraY*. Many single amino acid alterations in *pbp2x*, *pbp1a* and *pbp2b* have been previously demonstrated experimentally to increase pneumococcal resistance to beta-lactams (Hakenbeck, Bruckner *et al.* 2012). Mutations within or close to the active sites of the transpeptidase domain in penicillin binding proteins have been reported to be associated with penicillin resistance. By interfering with the formation of a covalent complex between the active site serine

and antibiotic molecules, these mutations help reduce the binding affinity of beta-lactam rings to the transpeptidase enzyme. This allows the pneumococci to form a functional cell wall, and thereby becoming non-susceptible. Many predicted loci co-localise with or surround the transpeptidase active sites. These are recognised as three conserved amino acid motifs, SXXK, SXN and KT(S)G and are highlighted as vertical dotted lines in the bottom panel of **Figure 6.2**. Amino acid alterations close to the active site often lead to conformational changes. An association at T338A in *pbp2x*, which is located next to the active site at position 337 was observed. The side chain of T338 is required for hydrogen-bonding, and the T338A substitution results in the distortion of the active site which lowers the binding affinity to beta-lactam (Hakenbeck, Bruckner *et al.* 2012). In *pbp1a*, an alteration from TSQF to NTGY at position 574-577 also matched previous reports. This alteration results in lower acylation efficiency *in vitro* (Job, Carapito *et al.* 2008). In addition to candidates known to confer structural changes, some candidates are consistent with compensatory mutations identified earlier. A substitution E285Q in *pbp1a* proposed to ameliorate the fitness cost caused by resistance in *pbp2b* (Albarracin Orio, Pinas *et al.* 2011) was also detected here. Other alterations that are consistent with previous literature were tabulated in **Appendix E**.

In addition to *pbp* genes, associations were observed in *mraY* and *mraW*, which encode transferases. Both function upstream of the *pbp* genes in the peptidoglycan biosynthesis pathway. They could potentially affect antibiotic susceptibility or represent compensatory mutations that interact epistatically with changes associated with resistance.

6.3.2.2 Candidate loci in genes outside the peptidoglycan biosynthesis pathway

The genome-wide screen provided an opportunity to identify associations outside the peptidoglycan biosynthesis pathway, which is the direct target of beta-lactams. The results obtained here highlighted nine loci found outside this pathway. The loci comprised 31 SNPs co-detected in both Maela and Massachusetts datasets; some of which were detected in the gene for heat shock protein *clpL*. Mutants lacking ClpL have been reported to be more susceptible to penicillin. The effect was attributed to

the ability of ClpL to interact with PBP2x and to stabilise *pbp2x* expression (Hakenbeck, Bruckner *et al.* 2012). In the Massachusetts data alone, amino acid alterations in a histidine kinase sensor, *ciaH* and its response regulator *ciaR*, also show an association with beta-lactam non-susceptibility. Previous reports showed that mutations in *ciaH* resulted in higher expression of *ciaR*. Hyperactivation of the regulon CiaR in turn leads to increased beta-lactam resistance (Muller, Marx *et al.* 2011, Hakenbeck, Bruckner *et al.* 2012). Associations between genes functioning in cell division, *ftsL* and *gpsB,* and beta-lactam resistance were observed in both Maela and Massachusetts datasets. Both proteins are required for complete cell wall formation. Depletion of GpsB leads to cell deformation (Land, Tsui *et al.* 2013). Based on known functions, these identified candidate loci potentially interact with *pbp* genes, either directly or indirectly through regulation or participating in cell wall formation. However, the hypotheses arising from these associations will require further experimental validation to explore the mechanisms of how these alterations might influence beta-lactam susceptibility.

6.3.2.3 Candidate loci in genes conferring resistance to other antibiotics

In both Maela and Massachusetts data, association signals were unexpectedly detected in dihydrofolate reductase (*dhfR*) and dihydropteroate synthase (*folP*), genes whose allelic variants are known to confer resistance to another group of antibiotics, co-trimoxazole. The drug interferes with folate synthesis, which is essential for nucleotide biosynthesis, thereby inhibiting the bacterial DNA synthesis pathway (Maskell, Sefton *et al.* 2001). Since beta-lactams and co-trimoxazole target different pathways, and no known protein-protein interactions between the two pathways have been reported, it is unlikely that variants detected in *dhfR* and *folP* would contribute to a rise of beta-lactam resistance mechanistically. The linkage analysis discussed in 6.1.4.1 has shown that loci at *dhfR* and *folP* genes were not genetically linked to *pbp* genes or other beta-lactam targets, suggesting that they are not hitchhikers but were selected through advantages that their allelic variants had conferred to on the strains.

Based on clinical records from Thailand, co-trimoxazole has been listed as the second most frequently used antibiotic for upper respiratory treatment after beta-lactam

(Thamlikitkul and Apisitwittaya 2004). This contemporaneous use of both beta-lactam and co-trimoxazole antibiotics in the studied populations may have driven co-selection for resistance to the two unrelated groups of antibiotics. In both Maela and Massachusetts datasets, strains that are phenotypically resistant to beta-lactams are more likely to be phenotypically resistant to co-trimoxazole (**Table 6.1**, Fisher's exact test p-value < 2.2×10^{-16}). This suggests that the frequent use of two antibiotics might have created co-selection pressures, resulting in the detection of linked yet unrelated association signals.

Table 6.1 Co-occurrence of co-trimoxazole and beta-lactam resistance phenotypes

			Beta-lactam		**Fisher's exact test p-value and (odds ratio)**
			resistant	sensitive	
Maela	**Co-trimoxazole**	resistant	1,356	771	< 2.2×10^{-16} (10.36)
		intermediate	77	280	
		sensitive	68	517	
Massachusetts		resistant	102	38	< 2.2×10^{-16} (7.29)
		sensitive	125	341	

6.3.3 Beta-lactam specificity of resistance mutations

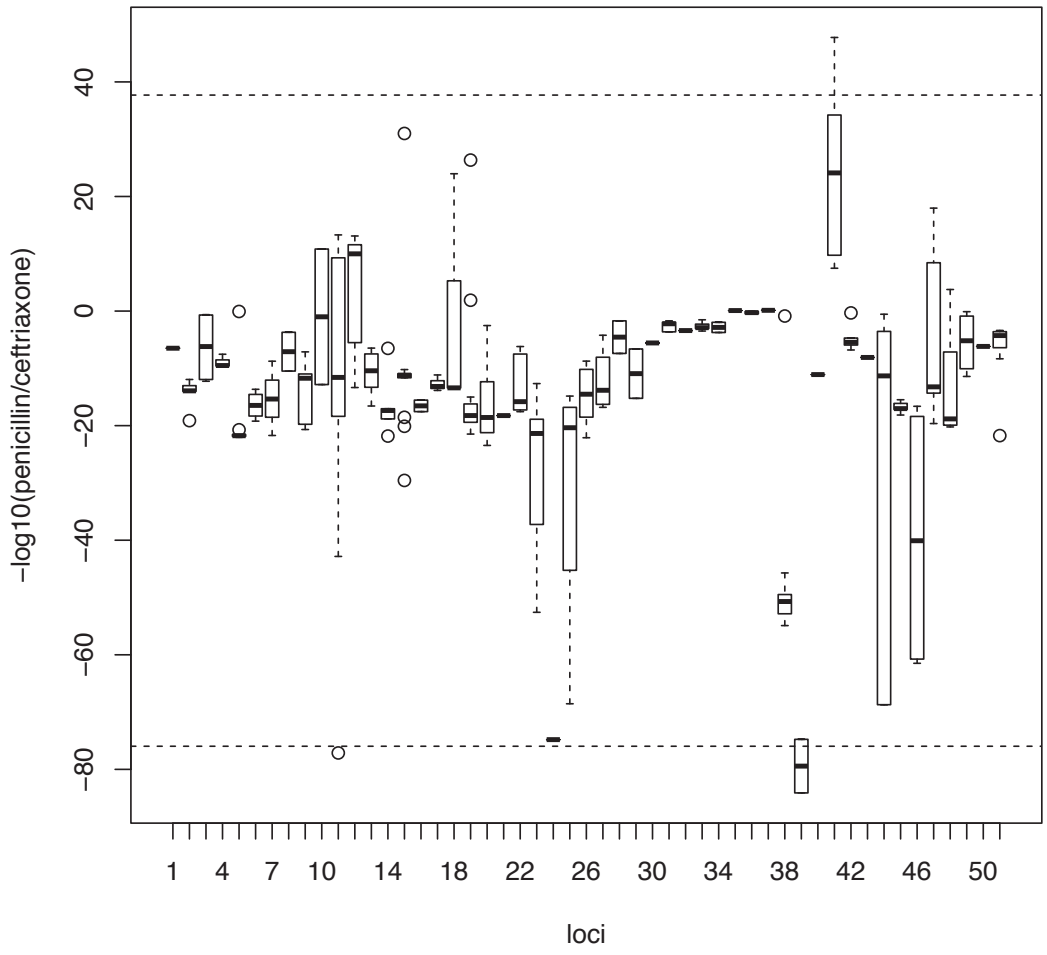
Although all beta-lactam antibiotics target peptidoglycan biosynthesis pathway, the group encompasses a range of drug molecules with different structures and chemical properties. Therefore, it is possible that each alteration detected earlier might confer greater resistance to certain classes of beta-lactam antibiotic than others. To test this hypothesis, the analyses were performed on the candidate SNPs identified in 6.1.3 using the continuous phenotypes recorded as the MIC value of two subclasses of beta-lactam antibiotics, penicillins and cephalosporins (here represented by ceftriaxone). Penicillins and cephalosporins display different structures. While the beta-lactam ring

is fused to a 5-membered thiaxolidine ring in penicillins, it is fused to a 6-membered dihydrothiazine ring in cephalosporins. Also, the side chains of the two drugs differ, leading to different kinetic properties (DePestel, Benninger *et al.* 2008).

The differential association of each locus with resistance to either penicillins or cephalosporins was summarised in **Figure 6.7**. Loci with stronger association with penicillin are distributed along the positive y-axis, while those showing a stronger association with cephalosporins are distributed along the negative y-axis. The result shows that some loci do not contribute equally to resistance to the different classes of beta-lactam antibiotic (Kruskal-Wallis rank sum test, p-value $< 2.2 \times 10^{-16}$). As can be seen in **Figure 6.7**, some loci show a strong preference towards penicillins or cephalosporins, suggesting specificity in resistance patterns.

Figure 6.7 Specificity of association signals for co-detected candidate loci with different classes of beta-lactam antibiotics

Bonferroni-adjusted p-values from associations with continuous phenotypes with each co-detected SNP were grouped into their linkage loci. Positive values on the y-axis show stronger association with penicillin resistance while negative values show stronger association with cephalosporin resistance. Horizontal dotted lines represent the 99th percentile.



6.3.4 Distribution of candidate alleles in the Maela and Massachusetts populations

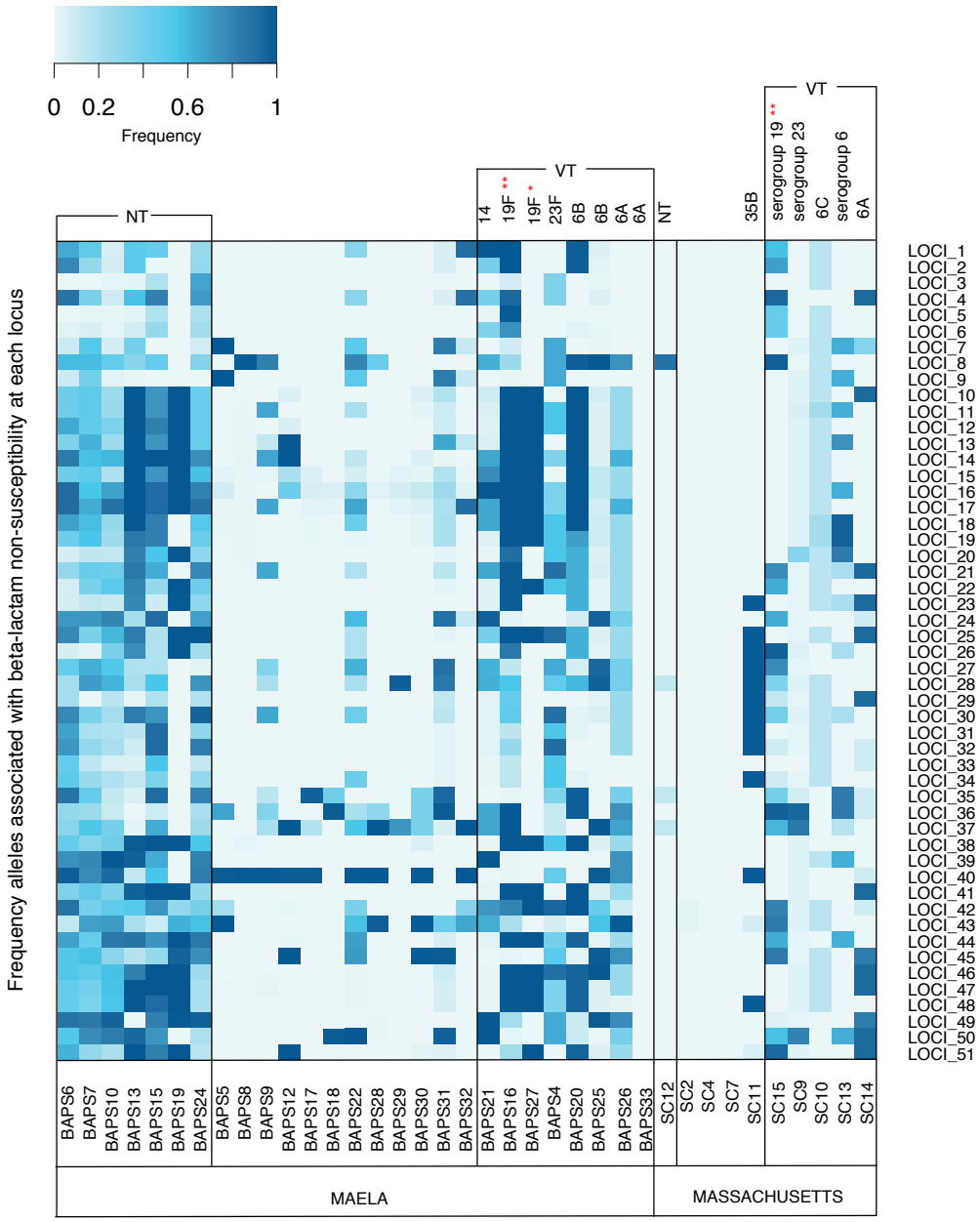
Next, the question posed in the previous chapter on the role of the nontypable strains in mediating the spread of beta-lactam resistance was investigated. Given that the pneumococcal population structure in both data sets is known (see 3.3.1.1 and Croucher, Finkelstein *et al* 2013), the prevalence of predicted beta-lactam resistance alleles in subpopulations was explored in each setting, particularly among nontypable clusters in the Maela population.

The result shows that candidate loci for beta-lactam resistance are heterogeneously distributed within subpopulations of both Maela and Massachusetts pneumococcal populations (**Figure 6.8**). The globally dispersed multidrug resistant lineages PMEN-14 and PMEN-1, along with other vaccine target lineages appear to carry candidate resistance alleles at a higher frequency. This reflects the vaccine's design to target serotypes associated with antibiotic resistance (Dagan 2009, Croucher, Finkelstein *et al.* 2013). Given this high frequency of resistance alleles, vaccine administration might be expected to help reduce beta-lactam resistance within the population as a whole. However, this has not been observed following the administration of pneumococcal conjugate vaccine (PCV7) (Huang, Platt *et al.* 2005, Huang, Hinrichsen *et al.* 2009), and this observation might provide a mechanism at the allelic level explaining why this may not occur. Nontypable lineages in Maela and serotype 35B in Massachusetts are non-vaccine lineages. Shown in **Figure 6.8**, each of these non-vaccine lineage harbours a high frequency of resistance alleles, allowing them to act as a reservoir for beta-lactam resistance following vaccination programmes.

For nontypable lineages, this observation provides an independent validation to their role as the source and sink of resistance alleles previously proposed in chapter 4 and chapter 5. Such a role might help generate more combinations of beta-lactam resistance alleles that are then seeded into the wider population. Such a trend observed in the NT lineages may thus have clinical consequences with respect to the spread of beta-lactam resistance in a community dominated by nontypable lineages like Maela.

Figure 6.8 Frequency of putative resistance alleles from candidate loci in the Maela and Massachusetts data

The resistance allele frequency at each locus (rows) in each population cluster (columns) is represented by different shades on a blue-grey scale. Vaccine and non-vaccine serotype status of each cluster are shown at the top of each column. Asterisks highlight two drug resistant global clones: * represents the PMEN1 lineage, ** represents the PMEN14 lineage. Note that mixed population clusters have been removed from the diagram.



6.4 Conclusion

Bacterial GWAS has allowed researchers to link individual elements of the genotype, including core genes, mobile genetic elements and SNPs, to specific phenotypes (Falush and Bowden 2006). This approach has been successfully performed in bacteria: *Campylobacter* (Sheppard, Didelot *et al.* 2013), *Staphylococcus aureus* (Laabei, Recker *et al.* 2014) and in this work, *Streptococcus pneumoniae*. Together, this holds a great promise for microbial functional genomics where one could design an experiment to systematically test the phenotypes followed by large-scale sequencing to draw the phenotype-genotype associations. Proposed by (Dutilh, Backus *et al.* 2013), a GWAS based approach may also be useful to link metagenomic entities including functions or taxa (operational taxonomic units) observed across metagenomic samples to clinical or environmental metadata. The term "metagenome-wide" association linking environmental parameters to metagenomic entities has been suggested.

Although these bacterial genome-wide and metagenome-wide associations are theoretically possible, their resolution will inherently be limited by the bacterial clonal population structure. This analysis used an *S. pneumoniae* dataset of large sampling size than previously, and the power to detect associated variants is therefore enhanced. Moreover, *pbp* genes, which are known targets of beta-lactams have been previously shown to be recombination hotspots, thereby significantly reducing the effect of long haplotype blocks. Together, this enabled a refinement of genetic determinants for beta-lactam non-susceptibility from "mosaic genes" to a single SNP or smaller linkage group. An increased resolution in the assignment of genetic variants will be useful for the prediction of antibiotic resistance/sensitivity from whole genome sequencing in surveillance and clinical studies.

A genome-wide screen allowed a search for loci associated with resistance beyond the known targets for beta-lactams in the peptidoglycan biosynthesis pathway, and reported associations in many previously characterised as well as many novel loci. The latter will require experimental validation to verify their contributions to resistance mechanisms. The results also show that loci can either contribute

universally to all beta-lactam resistance, or exert a stronger effect against certain classes of antibiotics. Moreover, these identified loci have a highly non-uniform distribution in the populations. They are enriched not only in vaccine-targeted but also non vaccine-targeted lineages, including the nontypable lineages detected in Maela. This provides further supportive evidence for the role of nontypables as the hub of genetic exchanges and a potential reservoir for antibiotic resistance genes in the Maela pneumococcal population, a concept that has been discussed throughout this thesis.

Chapter 7: Conclusion and future directions

7. Conclusions and future directions

In summary, this thesis has described diversity in pneumococcal populations, characterised evolutionary rates and genetic exchanges, and identified genetic determinants contributing to antibiotic resistance in isolates from healthy carriage in the human nasopharynx. Carriage is a prerequisite for the development of pneumococcal invasive diseases (Bogaert, De Groot *et al.* 2004). It is also the phase where evolution shapes the wider population structure and, thereby, the prevalence of susceptibility to clinical interventions such as antibiotics and vaccines (O'Brien and Santosham 2004, Dagan and Klugman 2008). Due to the high level of genomic plasticity, the pneumococci have rapidly developed antibiotic resistance. This area has been under intense focus since the late 1960s, when the first resistant pneumococcal isolates were identified (Klugman 1990, Appelbaum 1992, Crook and Spratt 1998, Hanage, Fraser *et al.* 2009, Donkor, Bishop *et al.* 2011). Moreover, the genomic plasticity also led to capsular switching, creating new variants that are not targeted by vaccines. The switches were observed throughout the history of pneumococci (Moore, Gertz *et al.* 2008, Donati, Hiller *et al.* 2010, Wyres, Lambertsen *et al.* 2013, Croucher, Finkelstein *et al.* 2013) and have been under recent study to evaluate vaccine efficacy. These studies have provided greater understanding on pneumococcal evolution and development of antibiotic resistance and vaccine escape serotypes emerging from carriage. I hope this thesis helps add a few pieces to the jigsaw of our current picture of pneumococcal populations.

This thesis presents an unprecedented density of sampling of over 3,000 samples collected over a 3-year period, from a densely populated area of 2.4 km^2 allowed an opportunity to capture the exchange of genetic materials. Moreover, the relatively large sample size (Wyres, Conway *et al.* 2014) allowed robust statistics for measuring lineage-specific evolutionary patterns and performing genome-wide association studies which had previously been difficult in bacteria. This chapter summarises the findings from all chapters, their applications elsewhere and directions for future work.

7.1 Biological summary

7.1.1 Views from Maela data

7.1.1.1 Recombination allows rapid adaptation in response to environmental changes

The role of recombination in mediating sequence exchange and allowing the pneumococci to adapt in response to clinical interventions has long been recognised (Coffey, Dowson *et al.* 1995, Lipsitch 2001, Hanage, Fraser *et al.* 2009, Croucher, Harris *et al.* 2011). This role has been re-emphasised here with genomic evidence given in Chapter 5. Consistent with previous studies, the chapter demonstrated that the most frequently exchanged genes were those associated with antibiotic resistance and immune selection, with the former being sensitive to the levels of antibiotic consumption. This thesis additionally captured bacterial response to temporal changes in the consumptions of two types of antibiotics; beta-lactams and co-trimoxazole. A reduction in consumption of co-trimoxazole over time was manifest in the patterns of recombination, with more recent events having a weaker association with resistance. Vice-versa, a continuing high beta-lactam consumption was consistent with a higher resistance observed in recombining strains in both recent and older recombination events. This demonstrates the role of recombination in enabling the bacteria to adjust to temporal fluctuation in addition to its role in spatial differentiation, as reported previously (Shapiro, Friedman *et al.* 2012).

Importantly, this temporal change in antibiotic consumptions (2002) began several years prior to the Maela pneumococcal carriage study (2007-2010), yet the data was used able to identify evolutionary events from years prior to the sampling time. This highlights the ability to identify temporal changes in selection at loci linked to fluctuating selective pressures, and might allow a prediction of behaviour of this pathogen in response to changes in future. An encouraging prediction based on evidences described here is that a reduction in the use of antibiotic may lead to a decrease of a drug resistant population (in this case for co-trimoxazole), if there is no influx of resistant populations from outside. Evidence given here helps support the models of the relationship between antibiotic consumption and the frequency of

resistance in communities first put forward in 1999 (Austin, Kristinsson and Anderson 1999).

7.1.1.2 Behaviour of nontypable pneumococci

The observation of lineage specific variation in rates of recombination, both for donation and receipt of DNA, implies differential rates of response to environmental selection pressures between lineages. An elevated rate of acquisition and donation of recombinant DNA was observed in nontypable cluster BC3-NT (Chapter 4), suggesting a higher capacity for adaptation in nontyable isolates. The observation is consistent with the general idea that capsule could act as a barrier for DNA uptake. Though increased recombination could bring transient benefits, there are potential long-term disadvantages due to increasing genomic instability (Giraud, Matic *et al.* 2001). As noted in Chapter 3, sporadic switches between the NT and encapsulated states were observed. This may serve as a mechanism to modulate the trade-off between benefit and cost of increased recombination rates. Though recombination efficiency is promoted by the non-encapsulated status, other factors should also be considered. Molecular mechanisms limiting the acquisition of foreign DNA, like such as restriction modification systems, haves been described (Johnston, Martin *et al.* 2013). Also the genomic context, including sequence similarity and repeat elements which potentially promote strand exchanges, have been investigated in (Hiller, Ahmed *et al.* 2010) and (Croucher, Harris *et al.* 2012) (see introduction 1.2.1). These factors may also influence the recombinogenic behaviour of each pneumococcal lineage but were not fully investigated here.

7.1.1.3 A potential role of nontypables as a genetic reservoir

The elevated level of receipt and donation of recombinant DNA observed in Maela NT lineages suggests their role as a hub for genetic exchange in the Maela population. Moreover, a heightened level of beta-lactam resistant determinants harboured by BC3-NT, and other NT clusters (Chapter 6) further supports their role as a reservoir of antibiotic resistant genes, which may allows the pneumococcal population the potential to adapt to antibiotics more rapidly.

As many of the vaccine targeted serotypes are associated with antibiotic resistance, it was hoped that the removal of, or reduction in, such serotypes would reduce the pool of resistance alleles and hence decrease pneumococcal resistance to antibiotics (Dagan and Klugman 2008). Although resistance-encoding alleles might be passed from lineages that escape vaccines as a consequence of serotype switching, there is also a possibility that resistant alleles might come from highly recombinogenic lineages, including those that are not currently targeted by available vaccines such as the NT.

As NT lineages are more difficult to detect through conventional serotyping schemes and are less likely to cause invasive pneumococcal diseases, this group of pneumococci has received less attention (Hathaway, Stutzmann Meier *et al.* 2004). However, NTs are common in carriage, making up the majority of isolates collected in Maela camp (Chapter 3). Using MLST typing, (Hanage, Kaijalainen *et al.* 2006) showed that the NT lineages have been transmitted inter-continentally and have been included in recent Pneumococcal Molecular Epidemiology Network analyses as PMEN42 (ST344) and PMEN43 (ST448). This intercontinental spread may reflect an advantage in transmission and these lineages have been associated with outbreaks of conjunctivitis (Hanage, Kaijalainen *et al.* 2006).

This thesis has highlighted several characteristics of NT pneumococci detected in Maela. Given their high recombination tendency in Chapter 4, their predominance in carriage in Chapter 3, the prevalence of genetic determinants for antibiotic resistance detected in these lineages in Chapter 6, the evidence for successful inter-continental transmission and the fact that they are not currently targeted by vaccines, this thesis proposes that the NT could potentially be a genetic reservoir for all pneumococci, allowing harmful traits such as antibiotic resistance to circulate in the population.

7.1.2 Applications of views from Maela to other global collections

The capacity and speed of adaption mediated by homologous recombination seen in Maela may vary between different localities due to distinct population structures and the prevalence of NT pneumococci observed at each location.

7.1.2.1 Distinct population structures may have a distinct adaptive capacity

Chapter 3 presents the comparison of population structures between different locations: UK, US, Kenya, Gambia and Thailand, through MLST profiles collected between 2006-2010. Except for a high similarity between UK and US population structures, the results revealed only a small overlap in genotypes for each population despite the close timeframes of sampling. This small overlap is made up of recognised pneumococcal lineages that have been particularly successful in spreading worldwide. However, the effects of the globally spread clones on local population structure are likely to be buffered by locally distinct lineages, which are present at a higher proportion in each area.

A more in-depth comparison of population structure was performed between the Maela and Massachusetts data (Croucher, Finkelstein *et al.* 2013) where whole genome sequences were available. Consistent with the comparison made by MLST, the Maela and Massachusetts population phylogenies revealed a little overlap in the prevalent genotypes for each population. Each population comprised a large diversity of distinct genotypes manifested as star-like phylogenies, comprised of large clusters of closely related strains interspersed with smaller, looser clusters of divergent isolates.

The difference in population structures observed here might have an impact on differential response to clinical interventions implemented in different locations. What determines this impact remains obscure but detailed comparison of genomic datasets across a wide range of locations might provide clearer insights into the parameters that govern the process.

7.1.2.2 Differences in prevalence of nontypables in different population settings

Based on our previous hypothesis that NT lineages could act as the hub of genetic exchanges, the different proportion of NT pneumococci observed at each location might imply that strong adaptive responses detected in Maela (Chapter 5) may not be the same in other locations. While NT pneumococci appeared to be highly prevalent in Maela, they were observed at a lower frequency in Massachusetts, US; Southampton, UK; The Gambia; and Kilifi, Kenya (Chapter 3). This discrepancy may be due to different laboratory techniques, which might ignore NT due to its atypical morphology (Rolo, A *et al.* 2013) and subsequently lead to different numbers of NT being reported. However, laboratory protocols used in the two cohort studies: Maela, Thailand and Massachusetts, USA (Croucher, Finkelstein *et al.* 2013) were compared and neither showed a bias in NT detection; thereby confirming an actual lower prevalence of NT detected in Massachusetts compared to Maela (Personal communication with Dr Nicholas Croucher).

Although the NT population was confirmed to be low in Massachusetts, a resource-rich state of US, it has a high prevalence in poorer US communities including the Native American communities from Navajo and White Mountain Apache (Millar, O'Brien *et al.* 2009). In these Native American communities, the NT was the third largest serotype group, after 6A and 6B, detected in nasopharyngeal carriage of children < 6 years prior to vaccine introduction (Millar, O'Brien *et al.* 2009), and remained among the top 10 serotypes in all age strata post vaccine (Scott, Millar *et al.* 2012). Five percent (95% confidence interval (CI), 4.2% - 5.7%) of pneumococcal carriage isolates from Navajo and White Mountain Apache children collected between 2006-2008 were NT (Scott, Millar *et al.* 2012). In contrast, only 1.88% of carriage isolates from children aged < 7 years were NT in the study conducted in the Massachusetts during the same sampling time frame (Hanage, Huang *et al.* 2007). Based on multilocus sequence typing, the majority of STs detected in these Native American communities included ST344, ST448, ST1054, ST1186 and ST2011 (Scott, Hinds *et al.* 2012). The first two were members of PMEN clones: PMEN42 ($Norway^{NT}$-42) and PMEN43 (USA^{NT}-43) which have been reported globally, though with different prevalence.

It is unclear why NT pneumococci have a high prevalence in places like Maela refugee camp, Thailand and Native American communities from Navajo and White

Mountain Apache, US, but not in Massachusetts US. As the first two locations reported higher carriage rates than Massachusetts (Scott, Millar *et al.* 2012, Turner, Turner *et al.* 2012, Croucher, Finkelstein *et al.* 2013), it is possible that the non-encapsulated status allows some adaptive advantages in a densely colonised niche. A lack of capsule promotes greater bindings of pneumococcal surface attachment proteins to epithelia (Weiser, Austrian *et al.* 1994); thereby supporting their colonisation of the nasopharynx. Beyond this, little is known about factors governing the prevalence of NT pneumococci in different locations. Without densely sampling done in similar fashion to Maela, it will be difficult to predict the contributions of NT to genetic exchanges in different pneumococcal populations.

7.2 Methodological summary

In addition to pneumococcal biology, this thesis has documented some unprecedented technical challenges including the ability to handle large-scale data, and the application of genome-wide association approaches, which are common in human genetics, onto bacterial genomes.

7.2.1 Divide and conquer approach

With help from Professor Jukka Corander and his team, we showed that a rough population structure could be constructed from a large genomic data set by both the Bayesian clustering method (BAPS) and maximum likelihood phylogeny. However, the preference for large-scale analysis lies with BAPS as it operates much faster, while giving similar results when compared to phylogenetic tree generation (Chapter 3). By partitioning the diverse population into closely related clusters of smaller size, the data can be handled more effectively.

Population stratification is essential for genome-wide association studies (GWAS) (discussed in the next section) to determine whether the detected signals are due to true genetic associations or arise from shared common ancestry. BAPS allowed rapid estimation of the population structure in each the separate Maela and Massachusetts datasets. The tool has been further developed so it is now possible to perform BAPS

on the combined Maela and Massachusetts data and more, to determine their population structures at the same time. This will help reduce a potential bias introduced by running BAPS on the separate datasets, which might lead to more strict control over population stratification in one population than the other (discussed in Chapter 6).

7.2.2 Genome-wide association study

GWAS has been commonly used in human genetics to identify genetic loci associated with a particular trait. However, it has been difficult in bacteria due to the intrinsic clonal population structure, which hinders the differentiation of truly associated genetic variations from hitchhikers. Chapter 6 showed that it was possible to apply the technique on highly recombinogenic bacteria, where recombination has occurred frequently enough to disrupt the clonal structure. Moreover, the large sampling size used in this study: 3,085 isolates from Maela, and 616 isolates from Massachusetts cohorts, has allowed more robust statistics. This led to the identification of genetic determinants of antibiotic resistance down to single polymorphic changes or small loci. Apart from an application in *S. pneumoniae* as documented in this thesis, the method has been recently been applied to *Campylobacter* (Sheppard, Didelot *et al.* 2013) and *Staphylococcus aureus* (Laabei, Recker *et al.* 2014), suggesting GWAS as a promising avenue for identifying bacterial genes or genetic loci associated with traits such as antibiotic resistance, transmission and virulence.

7.3 Future directions

There are two potential areas that can be followed up from the works described here: first, transmission analysis; and second, bacterial components that would elicit host immune responses. The former may help control the spread of pneumococci in carriage and the latter might help identify suitable gene candidates for pneumococcal vaccines.

7.3.1 Pneumococcal transmission

Though it was not documented in this thesis, preliminary investigation showed that donors of recombinant fragments (Chapter 3) were largely observed to be strains colonising individuals living in the same household rather than those colonising individuals living in separate houses (Fisher's exact test, p value 2.2 x 10^{-16}). The observation implies high rates of household transmission, which is consistent with previous reports conducted elsewhere (Shimada, Yamanaka *et al.* 2002, Mosser, Grant *et al.* 2014). Whole genome sequencing is increasingly being used to investigate transmission and disease outbreaks caused by various bacterial pathogens (Bryant, Grogono *et al.* 2013, Peacock 2014, Price, Golubchik *et al.* 2014). It should be possible to adapt the techniques developed previously to track transmission in other bacteria for *S. pneumoniae,* although the high levels of recombination detected in this species will likely pose some challenges.

7.3.2 Bacterial-host interactions

Chapter 5 has captured the signals of selection pressure from host immunity that results in diversifying antigenic proteins, manifested here as recombination hotspots. The interactions between host immune systems and pneumococcal protein antigens are complex, and involve many players. However, given that information on i) whole genome sequences of pneumococcal strains colonising individuals and ii) measured antibody responses to pneumococcal proteins, are both available; one could apply the GWAS approach to identify both generic and lineage specific candidate loci that might elicit host responses. This would allows all signatures of diversifying selection by specific immune responses to be captured in the genomic data, beyond the classical dN/dS measures. Also, the allelic variation of antigens that have signals of diversifying selection could be demonstrated in *vitro* to confirm their selective advantage in the presence of antibodies targeting particular alleles.

7.4 Publications resulting from this thesis

Works described in Chapter 3, Chapter 4 and Chapter 5 form the publication

C. Chewapreecha, S. R. Harris, N. J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D. M. Aanensen, A. E. Mather, A. J. Page, S. J. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner and S. D. Bentley (2014). "Dense genomic sampling identifies highways of pneumococcal recombination." Nat Genet 46(3): 305-309.

Works described in Chapter 6 were published in

C. Chewapreecha, P. Marttinen, N. J. Croucher, S. J. Salter, S. R. Harris, A. E. Mather, W. P. Hanage, D. Goldblatt, F. H. Nosten, C. Turner, P. Turner, S. D. Bentley, and J. Parkhill (2014). “Comprehensive Identification of Single Nucleotide Polymophisms Associated with Beta-Iactam Resistance within Pneumococcal Mosaic Genes”. PLOS Genetics. DOI:10.1371/journal.pgen.1004547

Thank you so much for reading till the end of this thesis. I hope you enjoyed it.

8. References

Adetifa, I. M., Antonio M., CA Okoromah C. A, Ebruke, C., Inem, V., Nsekpong, D., Bojang, A., and Adegbola R. A. (2012). “Pre-vaccination nasopharyngeal pneumococcal carriage in a Nigerian population: epidemiology and population in biology.” *PLoS One* 7(1): e 30548.

Adrian, P. V. and Klugman K. P. (1997). "Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of *Streptococcus pneumoniae*." *Antimicrob. Agents Chemother.* **41**(11): 2406-2413.

Albarracin Orio, A. G., Pinas ,G. E., Cortes, P. R , Cian, M. B, Echenique J. (2011). "Compensatory evolution of pbp mutations restores the fitness cost imposed by beta-lactam resistance in Streptococcus pneumoniae." *PLoS Pathog* **7**(2): e1002000.

Appelbaum, P. C. (1992). “Antimicrobial resistance in *Streptococcus pneumoniae*: an overview” *Clin Infect Dis* 15(1):77-83

Austin, D. J., Kristinsson, K. G., Anderson, R. M. (1999) “The relationshop between the volume of antimicrobial consumption in human communities and the frequency of resistance” *Proc Natl Acad Sci USA* 96(3):1152-1156.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). "Haploview: analysis and visualization of LD and haplotype maps." *Bioinformatics* **21**(2): 263-265.

Bek-Thomsen, M., Poulsen K., Kilian M. (2012). "Occurrence and evolution of the paralogous zinc metalloproteases IgA1 protease, *ZmpB, ZmpC, and ZmpD* in *Streptococcus pneumoniae* and related commensal species." MBio **3**(5). Pii e00303-00312

Bellos, A., Mulholland, K., O'Brien, K. L., Qazi, S. A., Gayer, M., Checchi, F. (2010) “The burden of acute respiratory infections in crisis-affected populations: a systematic review.” *Confi Health* 4:3.

Bergelson, J., Dwyer, G., Emerson, J. J. (2001). "Models and data on plant-enemy coevolution." *Annu Rev Genet* **35**: 469-499.

Bogaert, D,, De Groot, R., Hermans, P. W. (2004). “*Streptococcus pneumoniae* colonisation: the key to pneumococcal disease.” *Lancet Infect Dis* 4(3),144–154(2004)

Brock, S. C., McGraw, P. A., Wright, P. F., Crowe, J. E. (2002). "The human polymeric immunoglobulin receptor facilitates invasion of epithelial cells by *Streptococcus pneumoniae* in a strain-specific and cell type-specific manner." *Infect Immun* **70**(9): 5091-5095.

Brueggemann, A. B., Griffiths, D. T., Meats, E., Peto, T., Crook, D. W., Spratt B. G. (2003) "Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential" *J Infect Dis* 187(9):1424-1432.

Brueggemann, A. B., Muroki, B. M., Kulohoma, B. W., Karani, A., Wanjiru, E., Morpeth, S., Kamau, T., Sharif, S., Scott J. A. (2013). "Population genetic structure of *Streptococcus pneumoniae* in Kilifi, Kenya, prior to the introduction of pneumococcal conjugate vaccine." *PLoS One* 8(11): e81539.

Bryant, J. M., Grogono, D. M., Greaves, D., Foweraker, J., Roddick, I., Inns, T., Reacher, M., Haworth, C. S., Curran, M. D., Harris S. R. *et al.* (2013). "Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study." *Lancet* 381(9877): 1551-1560.

Carratalà J., Alcaide, F., Fernàndez-Sevilla, A., Corbella, X., Liñares, J., Gudiol F. (1995) "Bacteremia due to viridans streptococci that are highly resistant to penicillin: increase among neutropenic patients with cancer." *Clin. Infect. Dis.* 20,1169–1173

Claverys, J. P., Lefevre, J. C., Sicard, A. M. (1980). "Transformation of *Streptococcus pneumoniae* with *S. pneumoniae*-lambda phage hybrid DNA: induction of deletions." *Proc Natl Acad Sci U S A* **77**(6): 3534-3538.

Coffey, T. J., Dowson, C. G., Daniels, M., Spratt, B. G. (1995) " Genetic and molecular biology of beta-lactam-resistant pneumococci" *Microb Drug Resist* 1(1):29-34

Connor, T. R., Corander J., Hanage W. P. (2012). "Population subdivision and the detection of recombination in non-typable *Haemophilus influenzae*." *Microbiology* 158(Pt 12): 2958-2964.

Corander, J., Marttinen, P., Siren, J., Tang, J. (2008). "Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations." *BMC bioinformatics* **9**: 539.

Crook, D. W., Spratt, B. G. (1998) "Multiple antibiotic resistance in *Streptococcus pneumoniae*." *Br Med Bull* 54(3):595-610.

Croucher, N. J., Chewapreecha, C, Hanage, W. P., Harris, S. R. , McGee, L., van der Linden, M., Song, J. H., Ko, K. S. , de Lencastre, H. , Turner, C. *et al.* (2014). "Evidence for soft selective sweeps in the evolution of pneumococcal multidrug-resistance and vaccine escape." *Genome Biol Evol* 6(7):1589-1602.

Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage W. P. and Lipsitch M. (2013). "Population genomics of post-vaccine changes in pneumococcal epidemiology." *Nat Genet* **45**(6): 656-663.

Croucher, N. J., Harris, S. R., Barquist, L., Parkhill, J., Bentley, S. D. (2012). "A high-resolution view of genome-wide pneumococcal transformation." *PLoS pathog* 8(6): e1002745.

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J. H., Ko, K. S. *et al.* (2011). "Rapid pneumococcal evolution in response to clinical interventions." *Science* 331(6016): 430-434.

Croucher, N. J., Walker, D., Romero, P., Lennard, N., Paterson, G. K., Bason, N. C., Mitchell, A. M., Quail, M. A., Andrew, P. W., Parkhill J. *et al.* (2009). "Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81." *J Bacteriol* **191**(5): 1480-1489.

Dagan, R. (2009). "Impact of pneumococcal conjugate vaccine on infections caused by antibiotic-resistant Streptococcus pneumoniae." *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 15 Suppl 3: 16-20.

Dagan, R. and Klugman, K. P. (2008) " Impact of conjugate pneumococcal vaccines on antibiotic resistance" *Lancet Infect Dis* 8(2):785-795.

Dave, S., Pangburn, M. K., Pruitt, C., McDaniel, L. S. (2004). "Interaction of human factor H with *PspC* of *Streptococcus pneumoniae*." *Indian J Med Res* 119 Suppl: 66-73.

Dawkins, R., Krebs J. R. (1979). "Arms races between and within species." *Proc R Soc Lond B Biol Sci* 205(1161): 489-511.

DePestel, D. D., Benninger, M. S., Danziger, L., LaPlante, K. L., May, C., Luskin, A., Pichichero, M., Hadley, J. A. (2008). "Cephalosporin use in treatment of patients with penicillin allergies." *J Am Pharm Assoc* 48(4): 530-540.

Dessen A., Mouz, N., Gordon, E., Hopkins, J., Dideberg, O. (2001). “Crystal structure of PBP2x from a highly penicillin-resistant *Streptococcus pneumoniae* clinical isolate: a mosaic framework containing 83 mutations.” *J Biol Chem* 276,45105–45112

Devlin, B., Roeder, K. (1999). "Genomic control for association studies." *Biometrics* 55(4): 997-1004.

Doern, G. V., Ferraro, M. J., Brueggemann, A. B., Ruoff, K. L. (1996). “Emergence of high rates of antimicrobial resistance among viridans group streptococci in the United States. *Antimicrob Agents Chemother* 40, 891-894.

Donkor, E. S., Bishop, C. J., Gould, K., Hinds, J., Antonio, M., Wren, B., Hanage, W. P. (2011) "High levels of recombination among *Streptococcus pneumoniae* isolates from the Gambia." *mBio* **2**(3): e00040-00011.

Dowson, C. G., Hutchison, A., Brannigan J. A., George, R. C., Hansman, D., Liñares, J., Tomasz, A., Smith, J. M., Spratt, B. G. (1989) “Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*.” *Proc. Natl Acad. Sci. USA* 86,8842-8846.

Dowson, C. G., Hutchison, A., Spratt, B. G. (1989). “ Extensive re-modelling of the transpeptidase domain of penicillin-binding protein 2b of a penicillin-resistant South African isolate of *Streptococcus pneumoniae*.” *Mol Microbiol* 3, 95-102

Dutilh, B. E., Backus, L., Edwards, R. A., Wels, M., Bayjanov, J. R., van Hijum, S. A. (2013). "Explaining microbial phenotypes on a genomic scale: GWAS for microbes." *Brief Funct Genomics* 12(4): 366-380.

Erchibengoa M., Arostegi, N., Marimón, J. M., Alonso, M., Pérez-Trallero, E. (2012) “Dynamics of pneumococcal nasopharyngeal carriage in healthy children attending a day care centre in northern Spain. Influence of detection techniques on the results” *BMC Infect Dis* 12:69

Everett, D. B., Cornick, J., Denis, B., Chewapreecha, C., Croucher, N., Harris, S., Parkhill, J., Gordon, S., Carrol, E. D., French, N. *et al.* (2012). "Genetic characterisation of Malawian pneumococci prior to the roll-out of the PCV13 vaccine using a high-throughput whole genome sequencing approach." *PLoS One* **7**(9): e44250.

Falush, D., Bowden, R. (2006). "Genome-wide association mapping in bacteria?" *Trends Microbiol* 14(8): 353-355.

Feil, E. J., Maiden, M. C., Achtman, M., Spratt, B. G. (1999) "The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*." *Mol Biol Evol* **16**(11): 1496-1502.

Francisco, A. P., Vaz, C., Monteiro, P. T., Melo-Cristino, J., Ramirez, M., Carrico, J. A. (2012) "PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods." *BMC Bioinformatics* 13: 87.

Contreras-Martel C., Job V., Di Guilmi A. M., Vernet T., Dideberg O., Dessen A. (2006). "Crystal structure of penicillin-binding protein 1a (PBP1a) reveals a mutational hotspot implicated in beta-lactam resistance in *Streptococcus pneumoniae*." *J Mol Biol* 355,648-696.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) "The structure of haplotype blocks in the human genome." *Science* 296(5576): 2225-2229.

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., Taddei, F. (2001) "Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut." *Science* 291(5513): 2606-2608.

Goldblatt, D., Ramakrishnan, M., O'Brien, K. L. (2013) "Using the impact of pneumococcal vaccines on nasopharyngeal carriage to aid licensing and vaccine implementation; a PneumoCarr meeting report March 27-28, 2012, Geneva." *Vaccine* **32**(1): 146-152.

Goldstein, D. B., Allen, A., Keebler, J., Margulies, E. H., Petrou, S., Petrovski, S., Sunyaev, S. (2013) "Sequencing studies in human genetics: design and interpretation." *Nat Rev Genet* 14(7): 460-470.

Gordon, E., Mouz, N., Duee, E., Dideberg, O. (2000) "The crystal structure of the penicillin-binding protein 2x from *Streptococcus pneumoniae* and its acyl-enzyme form: implication in drug resistance." *J Mol Biol* 299,477-485

GPS. (2013). "20,000 Global pneumococcal project" from http://news.emory.edu/stories/2013/03/video_pnemonia_genome/

Grundman, H., Hori, S., Tanner, G. (2001) " Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms." *J Clin Microbiol* 39:4190-4192.

Hakenbeck, R., Bruckner, R., Denapaite, D., Maurer, P. (2012) "Molecular mechanisms of beta-lactam resistance in *Streptococcus pneumoniae*." *Future Microbiol* 7(3): 395-410.

Hakenbeck R., Martin, C., Dowson, C., Grebe, T. (1994) “Penicillin-binding protein 2b of *Streptococcus pneumoniae* in piperacillin-resistant laboratory mutants.” *J Bacteriol* 176, 5574-5577

Hanage, W. P., Bishop, C. J., Lee, G. M., Lipsitch, M., Stevenson, A., Rifas-Shiman, S. L., Pelton, S. I., Huang, S. S., Finkelstein, J. A. (2011) "Clonal replacement among 19A *Streptococcus pneumoniae* in Massachusetts, prior to 13 valent conjugate vaccination." *Vaccine* 29(48): 8877-8881.

Hanage, W. P., Fraser, C., Tang, J., Connor, T. R., Corander, J. (2009) "Hyper-recombination, diversity, and antibiotic resistance in pneumococcus." *Science* 324(5933): 1454-1457.

Hanage, W. P., Huang, S. S., Lipsitch, M., Bishop, C. J., Godoy, D., Pelton, S. I., Goldstein, R., Huot, H., Finkelstein, J. A. (2007). "Diversity and antibiotic resistance among nonvaccine serotypes of *Streptococcus pneumoniae* carriage isolates in the post-heptavalent conjugate vaccine era." *J Infect Dis* 195(3): 347-352.

Hanage, W. P., Kaijalainen, T., Saukkoriipi, A., Rickcord, J. L., Spratt, B. G. (2006) "A successful, diverse disease-associated lineage of nontypeable pneumococci that has lost the capsular biosynthesis locus." *J Clin Microbiol* 44(3): 743-749.

Hanage, W. P., Kaijalainen, T., Syrjanen, R. K., Auranen, K., Leinonen, M., Makela, P. H., Spratt, B. G. (2005) "Invasiveness of serotypes and clones of *Streptococcus pneumoniae* among children in Finland." *Infect Immun* **73**(1): 431-435.

Hanieh S., Hamaluba, M., Kelly, D. F., Metz, J. A., Wyres, K. L., Fisher, R., Pradhan, R., Shakya, D., Shrestha, L., Shrestha, A. *et al.* (2014) “ *Streptococcus pneumoniae* carriage prevalence in Nepal: Evaluation of a method for delayed transport of samples from remote regions and implications for vaccine implementation” *PLoS One* 9(6):e98739.

Hansman D. (1975) “Antibiotic sensitivity pattern of pneumococci relatively insensitive to penicillin and cephalosporin antibiotics.” *Med J Aust* 2, 740-742

Hathaway, L. J., Stutzmann Meier, P., Battig, P., Aebi, S., Muhlemann, K. (2004) "A homologue of *aliB* is found in the capsule region of nonencapsulated *Streptococcus pneumoniae*." *J Bacteriol* 186(12): 3721-3729.

Hiller, N. L., Ahmed, A., Powell, E., Martin, D. P., Eutsey, R., Earl, J., Janto, B., Boissy, R. J., Hogg, J., Barbadora, K. *et al.* (2010). "Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection." *PLoS pathog* 6(9): e1001108.

Hoge, C. W., Gambel, J. M., Srijan, A., Pitarangsi, C., Echeverria, P. (1998) "Trends in antibiotic resistance among diarrheal pathogens isolated in Thailand over 15 years." *Clin Infect Dis* 26(2): 341-345.

Hollingshead, S. K., Becker, R., Briles, D. E. (2000) "Diversity of *PspA*: mosaic genes and evidence for past recombination in *Streptococcus pneumoniae*." *Infect Immun* 68(10): 5889-5900.

Hsieh, Y. C., Wang, J. T., Lee, W. S., Hsueh, P. R., Shao, P. L., Chang, L. Y., Lu, C. Y., Lee, C. Y., Huang, F. Y., Huang, L. M. (2006) "Serotype competence and penicillin resistance in *Streptococcus pneumoniae*." *Emerging Infect Dis* 12(11): 1709-1714.

Huang, S. S., Hinrichsen, V. L., Stevenson, A. E., Rifas-Shiman, S. L., Kleinman, K., Pelton, S. I., Lipsitch, M., Hanage, W. P., Lee, G. M., Finkelstein, J. A. (2009) "Continued impact of pneumococcal conjugate vaccine on carriage in young children." *Pediatrics* 124(1): e1-11.

Huang, S. S., Platt, R., Rifas-Shiman, S. L., Pelton, S. I., Goldmann, D., Finkelstein, J. A. (2005) "Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004." *Pediatrics* 116(3): e408-413.

Hunter, P. R., Gaston, M. A. (1988) "Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity." *J Clin Microbiol* 26(11): 2465-2466.

Ip, M., Lyon, D. J., Yung, R. W., Tsang, L., Cheng, A. F. (2002) "Introduction of new clones of penicillin-nonsusceptible *Streptococcus pneumoniae* in Hong Kong." *J Clin Microbiol* 40(4): 1522-1525.

Jauneikaite, E., Jefferies, J. M., Hibberd, M. L., Clarke, S. C. (2012) "Prevalence of *Streptococcus pneumoniae* serotypes causing invasive and non-invasive disease in South East Asia: a review." *Vaccine* 30(24): 3503-3514.

Job, V., Carapito, R., Vernet, T., Dessen, A., Zapun, A. (2008) "Common alterations in PBP1a from resistant *Streptococcus pneumoniae* decrease its reactivity toward beta-lactams: structural insights." *J Biol Chem* 283(8): 4886-4894.

Job V., Di Guilmi A. M., Martin L., Vernet T., Dideberg O, Dessen A (2003). "Structural studies of the transpeptidase domain of PBP1a from *Streptococcus pneumoniae." Acta Crystallogr D Biol Crystallogr* 59,1067-1069.

Johnston, C., Martin, B., Granadel, C., Polard, P., Claverys, J. P. (2013) "Programmed protection of foreign DNA from restriction allows pathogenicity island exchange during pneumococcal transformation." *PLoS pathog* 9(2): e1003178.

Keenan J. D., Klugman, K., McGee, P. L., Vidal, J. E., Chochua, S., Hawkins, P., Cevallos, V., Gebre, T., Tadesse, Z., Emerson, P. M. *et al.* (2014). "Evidence for clonal expansion after antibiotic selection pressure: pneumococcal multilocus sequence types before and after mass azithromycin treatments." *J Infect Dis* pii: jiu552

Kislak, J. W., Razavi, L. M., Daly, A. K., Finland, M. (1968), "Susceptibility of pneumococci to nine antibiotics." *AM J Med Sci* 250, 261-268.

Klugman, K. P. (1990) "Pneumococcal resistance to antibiotics" *Clin Microbiol Rev* 3(2), 171-196

Ku, C. S., Loy, E. Y., Pawitan, Y., Chia, K. S. (2010) "The pursuit of genome-wide association studies: where are we now?" *J Hum Genet* 55(4): 195-206.

Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., Williams, P., Endres, J. L., Bayles, K. W., Fey, P. D. *et al.* (2014). "Predicting the virulence of MRSA from its genome sequence." *Genome Res* 24(5):839-849.

Laible, G., Hakenbeck, R. (1991) "Five independent combinations of mutations can result in low-affinity penicillin-binding protein 2x of *Streptococcus pneumoniae*." *J Bacteriol* 173, 6986-6990

Laible, G., Spratt, B. G., Hakenbeck, R. (1991) "Inter-species recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*." *Mol Microbiol* 5. 1993-2001.

Land, A. D., Tsui, H. C., Kocaoglu, O., Vella, S. A., Shaw, S. L., Keen, S. K., Sham, L. T., Carlson, E. E., Winkler, M. E. (2013). "Requirement of essential *Pbp2x* and *GpsB* for septal ring closure in *Streptococcus pneumoniae* D39." *Mol Microbiol* 90(5): 939-955.

Lefevre, J. C., Mostachfi, P., Gasc, A. M., Guillot, E., Pasta, F., Sicard, M. (1989) "Conversion of deletions during recombination in pneumococcal transformation." *Genetics* 123(3): 455-464.

Levine, O. S., O'Brien, K. L., Deloria-Knoll, M., Murdoch, D. R., Feikin, D. R., DeLuca, A. N., Driscoll, A. J., Baggett, H. C., Brooks, W. A., Howie, S. R. *et al.* (2012) "The Pneumonia Etiology Research for Child Health Project: a 21st century childhood pneumonia etiology study." *Clin Infect Dis* 54 Suppl 2: S93-101.

Lipsitch M. (2001). “Measuring and interpreting associations between antibiotic use and penicillin resistance in *Streptococcus pneumoniae*” *Clin Infect Dis* 32(7):1044-1054.

Manolio, T. A. (2010) "Genomewide association studies and assessment of the risk of disease." *N Engl J Med* 363(2): 166-176.

Marttinen, P., Hanage, W. P., Croucher, N. J., Connor, T. R., Harris, S. R., Bentley, S. D., Corander, J. (2012) "Detection of recombination events in bacterial genomes from large population samples." *Nucleic Acids Res* 40(1): e6.

Maskell, J. P., Sefton, A. M., Hall, L. M. (2001). "Multiple mutations modulate the function of dihydrofolate reductase in trimethoprim-resistant *Streptococcus pneumoniae*." *Antimicrob Agents Chemother* 45(4): 1104-1108.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., Hirschhorn, J. N. (2008) "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nat Rev Genet* 9(5): 356-369.

McCool, T. L., Cate, T. R., Moy, G., Weiser, J. N. (2002) "The immune response to pneumococcal proteins during experimental human carriage." *J Exp Med* **195**(3): 359-365.

McDougal, L. K., Rasheed, J. K., Biddle, J. W., Tenover, F. C. (1995) “Identification of multiple clones of extended-spectrum cephalosporin-resistant *Streptococcus pneumoniae* isolates in the United States.” *Antimicrob Agents Chemother* 39, 2282-2288.

McGee, L., Klugman, K. P., Wasas, A., Capper, T., Brink, A. (2001) "Serotype 19F multiresistant pneumococcal clone harboring two erythromycin resistance determinants (*erm(B)* and *mef(A)*) in South Africa." *Antimicrob Agents Chemother* 45(5): 1595-1598.

McGee, L., McDougal, L., Zhou, J., Spratt, B. G., Tenover, F. C., George, R., Hakenbeck, R., Hryniewicz, W., Lefevre, J. C., Tomasz, A., Klugman, K. P. (2001) "Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network." *J Clin Microbiol* 39(7): 2565-2571.

Medscape. (2014). "Bactrim, Bactrim DS (trimethoprim/sulpfamethoxazole) dosing, indications, interactions, adverse effects, and more." from http://reference.medscape.com/drug/bactrim-trimethoprim-sulfamethoxazole-342543#showall.

Millar, E. V., O'Brien, K. L., Zell, M., Bronsdon, A., Reid, R., Santosham, M. (2009) "Nasopharyngeal carriage of *Streptococcus pneumoniae* in Navajo and White Mountain Apache children before the introduction of pneumococcal conjugate vaccine." *Pediatr Infect Dis J* 28(8): 711-716.

Moore, M. R., Gertz, R. E., Woodbury, R. L., Barkocy-Gallagher, G. A., Schaffner, W., Lexau, C., Gershman, K., Reingold, A., Farley, M., Harrison, L. H. *et al.* (2008) "Population snapshot of emergent *Streptococcus pneumoniae* serotype 19A in the United States, 2005." *J Infect Dis*. 197(7):1016-1027

Mosser, J. F., Grant, L. R., Millar, E. V., Weatherholtz, R. C., Jackson, D. M., Beall, B., Craig, M. J., Reid, R., Santosham, M., O'Brien, K. L. (2014) "Nasopharyngeal carriage and transmission of *Streptococcus pneumoniae* in American Indian households after a decade of pneumococcal conjugate vaccine use." *PLoS One* 9(1): e79578.

Muller, M., Marx, P., Hakenbeck, R., Bruckner, R. (2011) "Effect of new alleles of the histidine kinase gene *ciaH* on the activity of the response regulator *CiaR* in *Streptococcus pneumoniae* R6." *Microbiology* 157(Pt 11): 3104-3112.

O'Brien, K. L., Santosham, M. (2004) "Potential impact of conjugate pneumococcal vaccines on pediatric pneumococcal diseases" *Am J Epidemiol* 159(7):634-644

Osaka, R., Nanakorn, S. (1996) "Health care of villagers in northeast Thailand--a health diary study." *Kurume Med J* 43(1): 49-54.

Padayachee, T., Klugman, K. P. (1999) "Novel expansions of the gene encoding dihydropteroate synthase in trimethoprim-sulfamethoxazole-resistant *Streptococcus pneumoniae*." *Antimicrob Agents Chemother* 43(9): 2225-2230.

Pasta, F., Sicard, M. A. (1996) "Exclusion of long heterologous insertions and deletions from the pairing synapsis in pneumococcal transformation." *Microbiology* 142 (Pt 3): 695-705.

Peacock, S. (2014) "Health care: Bring microbial sequencing to hospitals." *Nature* 509(7502): 557-559.

Pearce, B. J., Iannelli, F., Pozzi, G. (2002) "Construction of new unencapsulated (rough) strains of *Streptococcus pneumoniae*." *Res Microbiol* 153(4): 243-247.

Phan, N. B., Wilkinson, R. (2006) "Understanding the relationshop of information need specificity to search query length." SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval: 709-710.

Pharoah, P. D., Tsai, Y. Y., Ramus, S. J., Phelan, C. M., Goode, E. L., Lawrenson, K., Buckley, M., Fridley, B. L., Tyrer, J. P., Shen, H. *et al.* (2013) "GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer." *Nat Genet* 45(4): 362-370, 370e361-362.

Posada, D., Crandall, K. A. (2002) "The effect of recombination on the accuracy of phylogeny estimation." *J Mol Evol* 54(3): 396-402.

Potgieter, E., Carmichael, M., Koornhof, H. J., Chalkley, L. J. (1992) "*In vitro* antimicrobial susceptibility of viridans streptococci isolated from blood cultures." *Eur J Clin Microbiol Infect Dis* 11, 543-546.

Price, J. R., Golubchik, T., Cole, K., Wilson, D. J., Crook, D. W., Thwaites, G. E., Bowden, R., Walker, A. S., Peto, T. E., Paul, J. *et al.* (2014) "Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit." *Clin Infect Dis* 58(5): 609-618.

Price, M. N., Dehal, P. S., Arkin, A. P. (2009) "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix." *Mol Biol Evol* 26(7): 1641-1650.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. *et al.* (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am J Hum Gen* 81(3): 559-575.

Quin, L. R., Onwubiko, C., Moore, Q. C., Mills, M. F., McDaniel, L. S., Carmicle, S. (2007) "Factor H binding to *PspC* of *Streptococcus pneumoniae* increases adherence to human cell lines *in vitro* and enhances invasion of mouse lungs *in vivo*." *Infect Immun* 75(8): 4082-4087.

Ravin, A. W. (1959) "Reciprocal capsular transformations of pneumococci." *J Bacteriol* 77(3): 296-309.

Robinson, D. A., Edwards, K. M., Waites, K. B., Briles, D. E., Crain, M. J., Hollingshead, S. K. (2001) "Clones of *Streptococcus pneumoniae* isolated from nasopharyngeal carriage and invasive disease in young children in central Tennessee." *J Infect Dis* 183(10): 1501-1507.

Rolo, D., Domenech, A., Fenoll, A., Linares, J., de Lencastre, H., Ardanuy, C., Sa-Leao, R. (2013) "Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain." *PLoS One* 8(2): e57047.

Ruths, D., Nakhleh, L. (2005) "Recombination and phylogeny: effects and detection." *Int J Bioinform Res Appl* 1(2): 202-212.

Salter, S. J., Hinds, J., Gould, K. A., Lambertsen, L., Hanage, W. P., Antonio, M., Turner, P., Hermans, P. W., Bootsma, H. J., O'Brien, K. L., Bentley, S. D. (2012) "Variation at the capsule locus, cps, of mistyped and non-typable *Streptococcus pneumoniae* isolates." *Microbiology* 158(Pt 6): 1560-1569.

Scott, J. R., Hinds, J., Gould, K. A., Millar, E. V., Reid, R., Santosham, M., O'Brien, K. L., Hanage, W. P. (2012) "Nontypeable pneumococcal isolates among navajo and white mountain apache communities: are these really a cause of invasive disease?" *J Infect Dis* 206(1): 73-80.

Scott, J. R., Millar, E. V., Lipsitch, M., Moulton, L. H., Weatherholtz, R., Perilla, M. J., Jackson, D. M., Beall, B., Craig, M. J., Reid, R. *et al.* (2012). "Impact of more than a decade of pneumococcal conjugate vaccine use on carriage and invasive potential in Native American communities." *J Infect Dis* **205**(2): 280-288.

Shaper, M., Hollingshead, S. K., Benjamin, W. H., Briles, D. E. (2004) "*PspA* protects *Streptococcus pneumoniae* from killing by apolactoferrin, and antibody to PspA enhances killing of pneumococci by apolactoferrin [corrected]." *Infect Immun* 72(9): 5031-5040.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabo, G., Polz, M. F., Alm, E. J. (2012) "Population genomics of early events in the ecological differentiation of bacteria." *Science* 336(6077): 48-51.

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., Bentley, S. D., Maiden, M. C., Parkhill, J., Falush, D. (2013) "Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter." *Proc Natl Acad Sci USA* **110**(29): 11923-11927.

Shi, Z. Y., Enright, M. C., Wilkinson, P., Griffiths, D., Spratt, B. G. (1998) "Identification of three major clones of multiply antibiotic-resistant *Streptococcus pneumoniae* in Taiwanese hospitals by multilocus sequence typing." *J Clin Microbiol* 36(12): 3514-3519.

Shimada, J., Yamanaka, N., Hotomi, M., Suzumoto, M., Sakai, A., Ubukata, K., Mitsuda, T., Yokota, S., Faden, H. (2002) "Household transmission of *Streptococcus pneumoniae* among siblings with acute otitis media." *J Clin Microbiol* 40(5): 1851-1853.

Shivshankar, P., Sanchez, C., Rose, L. F., Orihuela, C. J. (2009) "The *Streptococcus pneumoniae* adhesin *PsrP* binds to Keratin 10 on lung cells." *Mol Microbiol* 73(4): 663-679.

Sifaoui F., Kitzis, M-D., Gutmann, L. (1996) "*In vitro* selection of one-step mutants of*Streptococcus pneumoniae* resistant to different oral β-lactam antibiotics is associated with alterations of PBP2x" *Antimicrob Agents Chemother* 40,152–156

Silver, L. L. (2007) "Multi-targeting by monotherapeutic antibacterials." *Nat Rev Drug Discov* 6(1): 41-55.

Simões A. S., Pereira, L., Nunes, S., Brito-Avô, A., de Lencastre, H., sá-Leão, R. (2011) "The widespread use of PCV7 led to an expansion of two PMEN clone ST63 (serotype 15A, and 19A) and Denmark ST230." *J Clin Microbiol* 49(8): 2810-2817.

Simpson, E. (1949) "Measurement of diversity" *Nature* 163:688

Smith, A. M., Botha, R. F., Koornhof, H. J., Klugman, K. P. (2001) "Emergence of a pneumococcal clone with cephalosporin resistance and penicillin susceptibility." *Antimicrob Agents Chemother* 45, 2648-2650.

Smith, A. M., Klugman, K. P. (1998) "Alterations in PBP1a essential for high-level penicillin resistance in *Streptococcus pneumoniae*." *Antimicrob Agents Chemother* 42,1329–1333

Smith, A. M., Klugman, K. P. (2003). "Site-specific mutagenesis analysis of PBP 1a from a penicillin-cephalosporin-resistant pneumococcal isolate." *Antimicrob Agents Chemother* 42,1329-1333

Sullivan, P. F., Daly, M. J., O'Donovan, M. (2012) "Genetic architectures of psychiatric disorders: the emerging picture and its implications." *Nat Rev Genet* 13(8): 537-551.

Tang, J., Hanage, W. P., Fraser, C., Corander, J. (2009) "Identifying currents in the gene pool for bacterial populations using an integrative approach." *PLoS Comput Biol* 5(8): e1000455.

Thamlikitkul, V., Apisitwittaya, W. (2004) "Implementation of clinical practice guidelines for upper respiratory infection in Thailand." *Int J Infect Dis* 8(1): 47-51.

The Border Consortium. (2012) "Camp and latest population." from http://www.tbbc.org/camps/mst.htm#ml.

Tocheva, A. S., Jefferies, J. M., Christodoulides, M., Faust, S. N., Clarke, S. C. (2013) "Distribution of carried pneumococcal clones in UK children following the

introduction of the 7-valent pneumococcal conjugate vaccine: a 3-year cross-sectional population based analysis." *Vaccine* 31(31): 3187-3190.

Toft, C., Andersson, S. G. (2010) "Evolutionary microbial genomics: insights into bacterial host adaptation." *Nat Rev Genet* 11(7): 465-475.

Turner, P., Turner, C., Jankhot, A., Helen, N., Lee, S. J., Day, N. P., White, N. J., Nosten, F., Goldblatt, D. (2012) "A longitudinal study of *Streptococcus pneumoniae* carriage in a cohort of infants and their mothers on the Thailand-Myanmar border." *PloS one* **7**(5): e38271.

Vestrheim D. F., Høiby, E. A., Aeberge, I. S., Caugant, D. A. (2010) "Impact of a Pneumococcal Conjugate Vaccination Program on Carriage among Children in Norway." *Clin Vaccine Immunol* 17(3):325-334.

Weiser, J. N., Austrian, R., Sreenivasan, P. K., Masure, H. R. (1994) "Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonization." *Infect Immun* 62(6): 2582-2589.

WHO (2014) World Heatlth Organisation Antimicrobial Resistance: Global Report on Surveillance 2014.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., Visscher, P. M. (2013) "Pitfalls of predicting complex traits from SNPs." *Nat Rev Genet* 14(7): 507-515.

Wyres, K. L., Conway, T. C., Garg, S., Queiroz, C., Reumann, M., Holt, K., Rusu, L. I. (2014). "WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare?" *Pathogens* 3(2): 437-458.

Wyres, K. L., Lambertsen, L. M., Croucher, N. J., McGee, L., von Gottberg, A., Liñares, J., Jacobs, M. R., Kristinsson, K. G., Beall, B. W., Klugman, K. P. *et al.* (2013) "Pneumococcal capsular switching: a historical perspective." *J Infect Dis* 207(3):439-449.

Wang, S., Yao, X. (2009) "Diversity analysis on imbalanced data sets by using ensemble models." *IEEE Symposium on Computational Intelligence and Data Mining.*

Yother, J., McDaniel, L. S., Briles, D. E. (1986) "Transformation of encapsulated *Streptococcus pneumoniae*." *J Bacteriol* 168(3): 1463-1465.

9. Appendices

The data listed in each appendix can be found in an enclosed CD at the back of this thesis. Alternatively, they can be accessed through the URLs provided below. Should more information is needed, please do not hesitate to contact me at cc12@sanger.ac.uk

Appendix A: Epidemiological data of Maela collection

The table contains strains, accession codes and associated epidemiological data for the Maela data set. From left to right, the columns represent strain name, accession codes associated with the data deposited in the ENA, dominant population clusters (BC1 to BC 7) used for focused analyses in chapter 4 and 5, primary BAPS cluster, secondary BAPS cluster, date of collection, serotype, sequence type, beta-lactam susceptibility and co-trimoxazole resistance.

doi:10.1038/ng.2895 (Supplementary Table 1)

http://www.nature.com/ng/journal/v46/n3/full/ng.2895.html#supplementary-information

Appendix B: Epidemiological data of PMEN14 collection

The table contains strains, accession codes and associated epidemiological data for the PMEN14 dataset. From left to right, the columns represent accession codes of each strain as deposited in the ENA, strain name, serotype, sequence type, region and country of origin (used for analyses in chapter 3), year of collection, and sources of diseases.

doi: 10.1093/gbe/evu120 (Supplementary Table S1)
http://gbe.oxfordjournals.org/content/suppl/2014/06/04/evu120.DC1

Appendix C: Associations to beta-lactam non-susceptibility co-detected in separate Maela and Massachusetts data sets

The table summarises all association statistics, linkage disequilibrium (LD) analysis, biological relevance and literature references for co-detected SNPs and associated loci in the Maela and Massachusetts data. From left to right, the columns represent coordinates from the reference genome (*S. pneumoniae* ATCC 700669), coding regions in which associations were detected, putative resistance nucleotide alleles, putative sensitivity nucleotide alleles, minor allele frequency (MAF), odds ratios (OR), synonymous and non-synonymous changes, positions on protein sequences with observed amino acid alterations, amino acid residues, alternative amino acid residues, literature reports, PubMed identifier (PMID), and linkage information.

doi:10.1371/journal.pgen.1004547.s005 (Table S1)
http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1004547#s5

Appendix D: Associations to beta-lactam detected in the Maela data.

The table summarises all association statistics and linkage disequilibrium (LD) analyses for variations that show significant associations in the Maela data. From left to right, the columns represent coordinate from the reference genome (*S. pneumoniae* ATCC 700669), majority detected SNPs, minor allele frequency, minor detected SNPs, odd ratios, linkage information and gene information.

doi:10.1371/journal.pgen.1004547.s006 (Table S2)
http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1004547#s6

Appendix E: Associations detected in the Massachusetts data

Summary of all association statistics and LD analysis for variations that show significant association in Massachusetts data. Columns were summarized using the same scheme described in Appendix C.

doi:10.1371/journal.pgen.1004547.s007 (Table S3)

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1004547#s7