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 (10th 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).

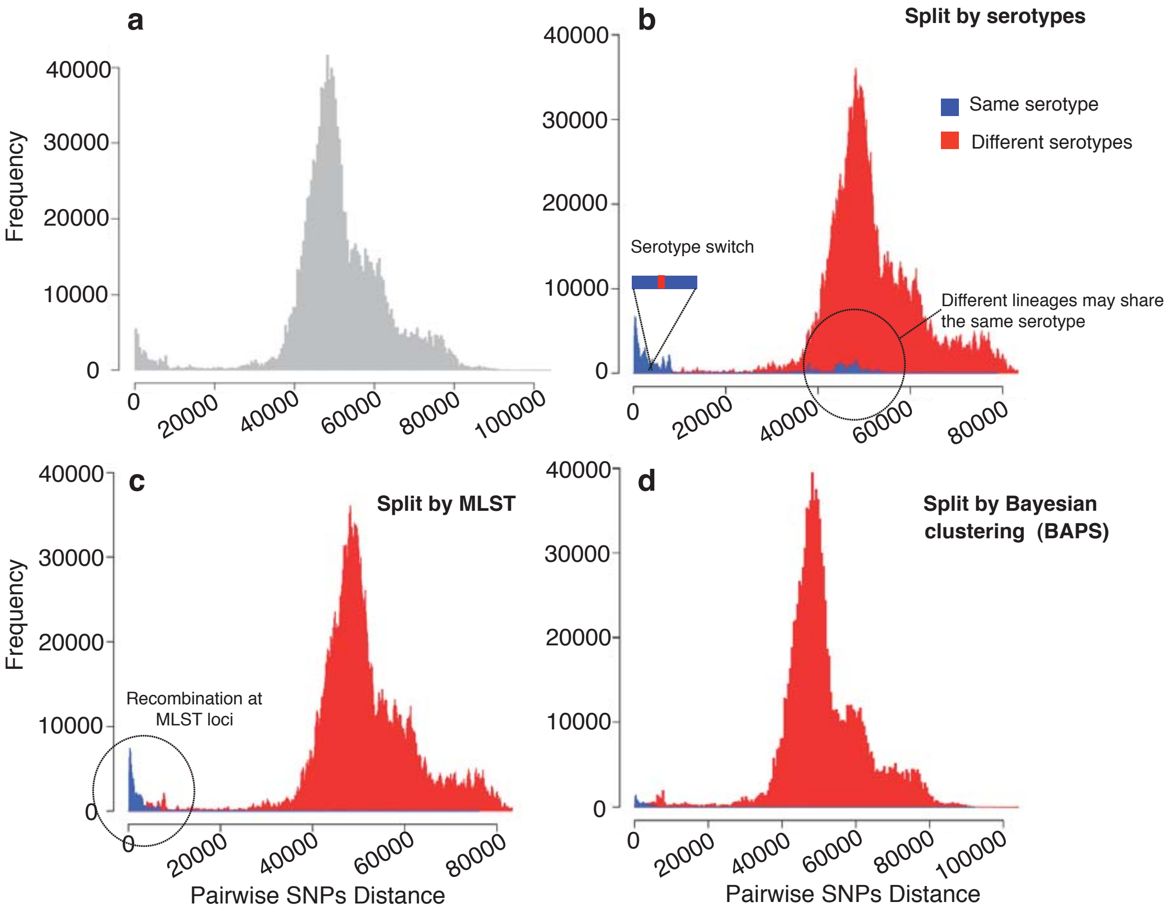
A greater power in detecting variations provides a higher population resolution but also presents a computational challenge to the analysis. For each isolate, total allelic sequences used to determine MLST account for 0.145% of the whole genome sequences, indicating around 690 times increase in the amount 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.