1 Introduction

1.1 The history and biology of the pneumococcus

1.1.1 "*Une maladie nouvelle*"

The first two reports of the isolation of the pneumococcus date from 1881. The US army surgeon George Sternberg subcutaneously inoculated rabbits with his own saliva (Sternberg, 1881), which he described as not presenting "any peculiarity unless it be that it is secreted in unusual abundance" (Sternberg, 1882), typically resulting in the animals' death within 48 hours. He was also able to establish that "the virulence of these fluids depended on the presence of the micrococci" that were observed to proliferate to high densities in the blood of the rabbits. Contemporaneously, Louis Pasteur performed similar experiments with the saliva of a child killed by rabies, resulting in "*une maladie nouvelle*" in the rodent hosts (Pasteur, 1881). The agent responsible was found to be a microscopic organism with "*propriétés sont fort curieuses*", namely the characteristic paired cell, diplococcal morphology and distinct capsular layer around each bacterium. Pasteur appears to have received some acclaim for discovering this 'new disease', whereas the reaction to Sternberg's study was more muted (Salmon, 1884).

Although discovered in cases of asymptomatic carriage in both cases, it did not take long to identify the pneumococcus as a major cause of human morbidity and mortality. It seems likely that the first pneumococci cultured from disease were those isolated by Friedländer, observed by staining exudates from cases of pneumonia (Friedländer, 1882; Friedländer, 1883; White *et al*., 1938), although his later works describe a different aetiological agent of pneumonia (Friedländer, 1886) thought to be *Klebsiella pneumoniae* (White *et al*., 1938; Austrian, 1960). This lead Friedländer to advocate the hypothesis that pneumonia was caused by multiple pathogens, which was confirmed by one of the first applications of Gram staining (Gram, 1884) and a survey of 129 cases of pneumonia (Weichselbaum, 1886), which found that the pneumococcus was the principal bacterial agent responsible for the disease.

The first transthoracic lung aspirations from patients with pneumonia had also recovered pneumococci (Leyden, 1882; Talamon, 1883). In 1884, Fränkel isolated a pneumococcus from a fatal case of pneumonia and was able to produce an infection in a rabbit inoculated with this bacterium, providing the first experimental link between human disease and the rodent experimental model (Fränkel, 1884). Two years later, he reported the first experiments that appear to have fulfilled Koch's postulates (Koch, 1893) and thereby demonstrated the pathogenicity of the pneumococcus (Fränkel, 1886; White *et al*., 1938). All patients in the study with fibrous pneumonia were found to have pneumococci in the lungs. These bacteria were then cultured and found to be able to cause a fatal infection in an animal model; the pneumococci could then be cultured again from the dead animals, and were still capable of producing disease in the same model.

Simultaneously, the role of the pneumococcus in other diseases was being uncovered. While studying patients with extra-pulmonary infections associated with pneumonia, Senger was able to isolate diplococcal bacteria from the cerebrospinal fluid of patients with meningitis, from cardiac lesions of patients with endocarditis and pericarditis, pleura of patients with pleurisy, and from the kidneys of patients with nephritis (Lancereaux and Besançon, 1886; Netter, 1886; Senger, 1886). These disseminated infections were hypothesised to have metastasised from the lungs. Pneumococcal colonisation of the inner ear, acute otitis media and meningitis, each without an accompanying lung infection, were reported the following year (Netter, 1887; Zaufal, 1887). Hence, in less than a decade, the pneumococcus had been identified as an agent carried asymptomatically by humans that was capable of causing disease in a number of mammalian hosts, including mouse, rat and rabbit (Gamaléia, 1888).

1.1.2 Taxonomic classification

Many names have been bestowed upon the pneumococcus since its discovery (White *et al*., 1938), but relatively few were sufficiently widely used to make a measurable impact on the literature (Figure 1.1). The term 'pneumococcus' itself originates with Fränkel (Fränkel, 1886), whereas Sternberg proposed *Micrococcus pasteuri* in honour of the bacterium's co-discoverer (Sternberg, 1885). *Diplococcus lanceolatus* was

widely used in the early 20th century, referring to the morphology of the pathogen (Foa and Bordoni-Uffreduzzi, 1888), but the most common binomial name at this time was *Diplococcus pneumoniae* (Weichselbaum, 1886). This persisted until superseded by the term *Streptococcus pneumoniae* following the reclassification of the bacterium into the streptococcal genus (Wannamaker and Matsen, 1972; Buchanan and Gibbons, 1974).

Streptococci are named for the distinctive chains they form during *in vitro* growth (from the Greek *streptos*, twisted, and *kokkos*, berry or grain), although with *S. pneumoniae* these are frequently short enough to only comprise pairs of cells. As part of the Firmicutes phylum, they are Gram-positive cells containing genomes with high AT content and a strong bias for protein coding genes to be on the leading strand (Rocha, 2008). Members of the streptococcus genus are further characterised as being catalase-negative facultative anaerobes with a number of auxotrophies (Wood and Holzapfel, 1995). On the basis of its 16S rRNA sequence, *S. pneumoniae* is deemed part of the 'mitis group', although DNA-DNA hybridisation data indicate that even this comparatively precise taxonomical grouping contains a great deal of genetic diversity (Kawamura *et al*., 1995). The most closely related species to *S. pneumoniae* appear to be *S. mitis* and *S. pseudopneumoniae* (Fraser *et al*., 2007; Kilian *et al*., 2008); both are nasopharyngeal commensals, with the former sometimes found to cause endocarditis (van der Meer *et al*., 1991) and the latter seemingly associated with some cases of pulmonary disease (Keith *et al*., 2006).

1.1.3 Metabolic and microbiological characteristics

Streptococci lack a functional tricarboxylic acid (TCA) cycle, hence their metabolism is driven by fermentation of carbohydrates, usually to lactate (Wood and Holzapfel, 1995). Under aerobic conditions, it appears that lactate oxidase and pyruvate oxidase are able to convert lactate to acetate, thereby producing another molecule of ATP from the fermentation process (Taniai *et al*., 2008), with both enzymes generating hydrogen peroxide as a by-product. The catalase-negative nature of the pneumococcus (Yesilkaya *et al*., 2000) means extracellular hydrogen peroxide can accumulate to concentrations up to ~1.1 mM (Pericone *et al*., 2000). The production

of lactate and acetate by streptococci lowers the pH of the environment to between 4.5 and 5.0, but the cells are able to maintain their cytosolic pH between 7.6 and 5.7 (Kashket, 1987). In the absence of an electron transport chain, the primary role of the streptococcal F_0F_1-ATPase may be the extrusion of protons, powered by ATP hydrolysis, to enable the bacteria to tolerate such acidic conditions.

Pneumococci are unable to synthesise a number of amino acids. Although missing a complete TCA cycle, oxaloacetate (the precursor required for aspartate, threonine, isoleucine, cysteine and methionine biosynthesis) is produced through an alternative route, the action of phosphoenolpyruvate carboxylase (Yamada and Carlsson, 1973). The other TCA cycle intermediate required for amino acid biosynthesis, 2-oxoglutarate (used for the production of glutamate and glutamine), is produced by the sequential activities of citrate synthase, aconitase and isocitrate dehydrogenase in *S. mutans* (Cvitkovitch *et al*., 1997) and the mitis group species *S. sanguinis* (Xu *et al*., 2007), but there is no evidence for these enzymes being present in *S. pneumoniae*. Furthermore, the complete biosynthetic pathways for histidine and arginine, again present in *S. mutans* (Ajdic *et al*., 2002) and *S. sanguinis* (Xu *et al*., 2007), are absent in pneumococci (Hoskins *et al*., 2001). Hence a defined medium for growing pneumococci requires the inclusion of glutamate, arginine and histidine (Rane and Subbarow, 1940).

Another molecule required for pneumococcal growth is choline (Rane and Subbarow, 1940), which is phosphorylated on uptake into the cell (Whiting and Gillespie, 1996). This is incorporated into the outer surface of the cell as part of a polymeric teichoic acid, the pneumococcal monomeric unit of which consists of ribitol phosphate linked to a pair of *N*-acetylgalactosamine residues (either one or both of which are connected to phosphocholine moieties through a phosphodiester bond), in turn joined to 2-acetamido-4-amino-2,4,6-trideoxy-D-galactose and a glucose residue (Jennings *et al*., 1980; Karlsson *et al*., 1999). It is the phosphoribitol group, which may carry *N*-acetylgalactosamine or D-alanine substitutions in place of hydroxyl groups, which links to peptidoglycan through a phosphodiester bond to form cell wall-associated teichoic acid (Brundish and Baddiley, 1968). Lipoteichoic acid is instead attached to a lipid anchor through the glucose residue in *S. pneumoniae* (Seo *et al*., 2008).

Although it is thought that all Gram positive bacteria have a teichoic acid or functionally analogous compound, the phosphocholine content of such structures seems to be limited to the mitis group streptococci (Neuhaus and Baddiley, 2003; Kilian *et al*., 2008). This appears to be an adaptation to the nasopharyngeal niche, as the Gram negative commensals and respiratory pathogens *Neisseria meningitidis* and *Haemophilus influenzae* both have phosphocholine decorations on their exteriors: the former adds these moieties to its pili (Weiser *et al*., 1998), while the latter includes the group within its lipopolysaccharide (Weiser *et al*., 1997), as do commensal *Neisseria*, such as *N. lactamica* (Serino and Virji, 2000).

1.1.4 Microbiological identification

The conventional microbiological approach for identifying pneumococci is based upon some readily testable characteristics (Tuomanen *et al*., 2004). *S. pneumoniae*, like many streptococci, is α-haemolytic when grown aerobically on blood plates due to the hydrogen peroxide produced as a by-product of fermentation (Facklam, 2002); this reflects the lack of a secreted cytolytic exotoxin capable of causing significant β-haemolysis around the cells, as observed with group A and B streptococci (Nizet, 2002). A distinctive feature that allows pneumococci to be differentiated from other α-haemolytic bacteria is the triggering of cell lysis by bile (Neufeld, 1900), which results from the activation of the major autolytic enzyme, LytA, by the bile salt deoxycholate (Mosser and Tomasz, 1970). An alternative test is susceptibility to optochin (also known as ethylhydrocupreine), which inhibits the pneumococcal F_0F_1-ATPase (Fenoll *et al*., 1994). However, resistance to deoxycholate has been observed in some *S. pneumoniae* isolates (Obregon *et al*., 2002), as has optochin resistance, which can arise through a single point mutation in the genes encoding the F_0F_1-ATPase (Pikis *et al*., 2001); hence neither of these phenotypic tests are completely accurate.

1.1.5 Pneumococcal serology

The first demonstration that antiserum from infected rabbits could be used to agglutinate *S. pneumoniae* cells *in vitro* was reported in 1902 (Neufeld, 1902). Subsequently, this approach was used to group some pneumococcal isolates into two

types; each was incapable of infecting a rabbit immunised with an isolate of the same type, but able to cause disease if the animal had been inoculated with strains of the other type (Neufeld and Händel, 1910). Further work on those strains that did not fall into either of these groups divided them into group 3, if they had a mucoid colony phenotype (at the time referred to as the separate species *Pneumococcus mucosus*), or group 4 if they superficially resembled types 1 or 2 (Dochez and Gillespie, 1913). This study found group 4 to be a heterogenous collection of strains, as each isolate produced an antiserum only effective against itself; such an analysis of group 3 was not possible, due to the difficulty in producing appropriate antisera. Over the following decades, groups 3 and 4 have been categorised into a large number of distinct types (Lund, 1960); the accepted nomenclature puts antigenically-related types into serogroups, denoted with a number, with a letter added to distinguish between individual serotypes if necessary.

A soluble component of the pneumococcus, present in the blood and urine of infected rabbits and clinical cases of disease, was found to be capable of causing agglutination when combined with the antisera appropriate for the type of the infecting bacteria (Dochez and Avery, 1917). Chemical analysis of this substance provoking the immune reaction showed it was a polysaccharide (Heidelberger and Avery, 1923). That this material constituted the capsule was demonstrated through the inability of antisera to agglutinate pneumococci once their capsule had been removed through either acid hydrolysis or enzymatic activity (Dubos and Avery, 1931). The enzymatic removal of the capsule also eliminated the ability of a group 3 pneumococcus to infect mice, showing that this polysaccharide layer is crucial for the ability of *S. pneumoniae* to cause systemic infections (Avery and Dubos, 1931). Examination of the exudates produced by mice inoculated intraperitoneally with *S. pneumoniae*, with and without the decapsulating enzyme, revealed only encapsulated bacteria were able to avoid phagocytosis by leucocytes, suggesting that the capsule's role was in evading the host immune response.

Southern blot analysis of genomic DNA digests revealed that the locus encoding the genes for capsule biosynthesis (*cps* locus) in a type 3 strain is present in between the penicillin-binding protein genes *pbp2X* and *pbp1A* (Arrecubieta *et al*., 1994);

sequencing of cloned regions of this *cps* locus revealed the flanking genes were *dexB* (encoding a dextran glucosidase) and *aliA* (encoding an extracellular oligopeptide-binding protein) (Arrecubieta *et al*., 1995; Dillard *et al*., 1995). All capsule biosynthesis gene clusters are found in this region (Dillard *et al*., 1995), except that of mucoid serotype 37; this simple capsule, a homopolymer of sophorosyl units, is synthesised by the single gene *tts* found elsewhere in the chromosome (Llull *et al*., 1999). Sequencing of the known *cps* loci has revealed almost all share a similar genetic structure, despite varying between ~10 kb and ~30 kb in size (Bentley *et al*., 2006; Park *et al*., 2007; Bratcher *et al*., 2010). A conserved set of regulatory genes are found at the 5' end of all the loci, followed by the glycosyl transferases and genes for the biosynthesis and modification of the nucleotide sugars in the downstream region. Also present in the gene cluster are the flippase, responsible for moving the structure out across the cell membrane, and the polymerase. All the non-mucoid capsules use a conserved polymerase, which operates using lipid-linked repeat unit intermediates, whereas serotypes 3 and 37 are composed of long, simple polysaccharide chains generated by a processive transferase activity (Waite *et al*., 2003).

As well as the genetic variation underlying the different serotypes, each capsule itself is expressed in a phase variable manner. Pneumococci of serotypes 3, 8 and 37 produce acapsular variants (termed 'rough', as opposed to 'smooth' encapsulated strains) during *in vitro* growth under conditions resembling those of a biofilm (Waite *et al*., 2001; Waite *et al*., 2003). This is the result of high-frequency frameshift mutations that disrupt genes necessary for capsule production. Spontaneous changes in the 'opacity' of colonies, representing a quantitative variation in the amount of capsule produced by *S. pneumoniae* cells, results from the variable expression of a putative regulatory gene (Saluja and Weiser, 1995). The 'opaque' variants produce higher levels of capsule and are more virulent in a mouse model when administered intraperitoneally (Kim and Weiser, 1998), whereas the 'transparent' variants have higher levels of teichoic acid and show a greater ability to colonise the nasopharynx in a rat model of carriage (Weiser *et al*., 1994; Kim and Weiser, 1998), likely due to their increased ability to adhere to host cells (Cundell *et al*., 1995b). The display of phosphocholine-containing structures in *N. meningitidis* and *H. influenzae* is also phase variable; in both cases, the switching mechanism is a variable length intragenic

repeat sequence within a phosphocholine processing gene (Weiser *et al*., 1997; Warren and Jennings, 2003).

1.1.6 A transformable pathogen

As the capsule is such an important virulence factor, rough strains must be administered in much larger doses than encapsulated strains in order to be able to cause systematic disease in mice. However, following subcutaneous inoculation of mice with a sublethal dose of rough pneumococci, accompanied by a sample of heat-killed smooth pneumococci, Griffith was able to select for, and recover, live encapsulated pneumococci from bacteraemia in the animal (Griffith, 1928). This 'transformation' of live cells by dead cells allowed rough isolates derived from type 2 strains to revert to their previous capsule type, or be converted to types 1 or 3, depending on the serotype of the killed cells. Subsequent work, converting a rough isolate derived from type 2 with material extracted from a type 3 pneumococcus *in vitro*, found that the transformation could occur when protein, polysaccharide and lipid were chemically removed from the extract, but that the process was inhibited by digestion with DNase (Avery *et al*., 1944). This demonstrated that DNA was the basis of bacterial genetics.

Like *S. pneumoniae*, a number of other mitis group streptococci have been found to be naturally transformable, and it seems likely that almost all members of the genus are able to incorporate exogenous DNA into their chromosome (Havarstein, 2010). There is also evidence that pneumococci are able to exchange sequence not just within the species, but also with *S. pseudopneumoniae* and *S. mitis* (Hanage *et al*., 2006; Kilian *et al*., 2008) and possible even more distantly related streptococci (Sibold *et al*., 1994).

1.2 Interactions with the host and microbiota

1.2.1 The pneumococcus as a human commensal

The main site of carriage of the pneumococcus is the human nasopharynx. Pneumococcal carriage rates are highest in the young: a study in Germany found 30 of 52 infants acquired a pneumococcus in the first two weeks of life (Gundel and Schwarz, 1932), and in Papua New Guinea all children had carried the pneumococcus at least once in the first three months of life (Gratten *et al*., 1986). The peak rate of carriage in healthy infants is in the first three years of life (Gray *et al*., 1982; Aniansson *et al*., 1992; Coles *et al*., 2001; Syrjanen *et al*., 2001; Bogaert *et al*., 2004b), with estimates of colonisation levels at this stage typically about 20% or greater (Bogaert *et al*., 2004a), then decline again as the individual approaches adulthood (Parry *et al*., 2000; Bogaert *et al*., 2004b). Evidence has been found that young children living together, such as those attending a day care centre (Bogaert *et al*., 2001; Dunais *et al*., 2003; Bogaert *et al*., 2004b; Regev-Yochay *et al*., 2004b) or living at home with siblings (Principi *et al*., 1999; Petrosillo *et al*., 2002; Regev-Yochay *et al*., 2004b), have an increased level of pneumococcal carriage. Exposure to cigarette smoke (Greenberg *et al*., 2006; Cardozo *et al*., 2008) and recent use of macrolide antibiotics (Principi *et al*., 1999; Petrosillo *et al*., 2002) have also been found to be risk factors for pneumococcal carriage by healthy individuals. There may also be a role for host genetics, as Australian Aboriginal children have a higher pneumococcal carriage rate than non-Aboriginal children (Watson *et al*., 2006), and very high rates of carriage are seen in some native American communities in the USA (Millar *et al*., 2006). However, these may alternatively be linked to socio-economic factors, which also affect carriage rates (Huang *et al*., 2004). Increased colonisation is seen in children with viral respiratory infections (Smith *et al*., 1976), likely as a result of increased adhesion between the bacterium and epithelium (Peltola and McCullers, 2004; Avadhanula *et al*., 2006), but it is not clear whether HIV-1 infection causes an increase in carriage (Janoff *et al*., 1993; Polack *et al*., 2000; McNally *et al*., 2006; Madhi *et al*., 2007).

Colonisation of an individual by a pneumococcal lineage may persist for a period ranging from a few days to several months (Gratten *et al*., 1986; Raymond *et al*., 2000), with the duration of carriage associated with the serotype of the bacterium (Gray *et al*., 1980; Smith *et al*., 1993; Sleeman *et al*., 2006). When such carriage

periods overlap, multiple colonisation is observed, a situation crucial to the horizontal exchange of DNA within the species. Surveys of carriage have typically found that between 8 and 30% of individuals colonised with pneumococci carry multiple strains (Gratten *et al*., 1989; Hare *et al*., 2008; Kaltoft *et al*., 2008; Brugger *et al*., 2010). However, such estimates are greatly affected by the techniques used, as approaches based on typing individual colonies underestimate the diversity of *S. pneumoniae* within a single nasopharynx (Huebner *et al*., 2000). As technological improvements lead to more sensitive methods for typing pneumococci, estimates of the rates of co-colonisation between pneumococcal strains will grow more precise (Turner *et al*., 2011).

1.2.2 Competition within the nasopharynx

The nasopharynx is also a reservoir for a number of other bacteria, including other respiratory pathogens. *In vitro* experiments have demonstrated that the levels of hydrogen peroxide produced by *S. pneumoniae* are sufficient to inhibit the growth of *Staphylococcus aureus*, *H. influenzae*, *N. meningitis* and *Moraxella catarrhalis* in culture (Pericone *et al*., 2000; Regev-Yochay *et al*., 2006). Furthermore, *S. pneumoniae* expresses a neuraminidase that is capable of removing sialic acid from the capsules of *N. meningitidis* and *H. influenzae*, thereby reducing the protection this moiety gives these bacteria from complement-mediated opsonophagocytosis by the host immune system (Shakhnovich *et al*., 2002). However, in a mouse model of colonisation, *H. influenzae* stimulated the clearance of *S. pneumoniae*, apparently through *H. influenzae* stimulating neutrophil-mediated killing of *S. pneumoniae* (Lysenko *et al*., 2005). Despite these mechanisms, surveys have found *S. pneumoniae* and *H. influenzae* colonising individuals (both HIV positive and negative) together more frequently than expected from their respective prevalences (Jacoby *et al*., 2007; Madhi *et al*., 2007), although this finding is not universal (Luotonen, 1982). A positive correlation has also been found between *S. pneumoniae* and *N. meningitidis* carriage (Bakir *et al*., 2001; Bogaert *et al*., 2005). However, an antagonistic relationship between pneumococci and *Staph. aureus* has been found in healthy children (Bogaert *et al*., 2004b; Regev-Yochay *et al*., 2004a; Madhi *et al*., 2007) although not in those infected with HIV (McNally *et al*., 2006; Madhi *et al*., 2007).

Intra-specific competition primarily appears to be mediated by small peptide bacteriocins, narrow spectrum bactericidal peptides secreted by cells. Both of the well-characterised systems in *S. pneumoniae*, the *blp* and *cibAB* bacteriocins, are regulated by similar, simple extracellular signalling mechanisms; an unmodified peptide signal, secreted into the medium, is detected by the extra-cellular surface of a two-component system. The *blp* locus encodes a specific signalling pathway along with the structural bacteriocin genes (de Saizieu *et al*., 2000), which were shown to be crucial in mediating the elimination of one pneumococcal strain by another in a mouse model of co-colonisation (Dawid *et al*., 2007). By contrast, the *cibAB* system is controlled by the same peptide pheromone as the competence system, and hence is partially responsible for the release of genomic DNA into the environment, through causing cell lysis, making it available for uptake (Guiral *et al*., 2005). At least two alleles of each of these quorum sensing systems are present in the pneumococcal population (Pozzi *et al*., 1996; Reichmann and Hakenbeck, 2000), leading to variation in the interactions between strains. Furthermore, assays based on the bacteriocin-induced lysis of indicator strains suggests there are other loci, not present in all pneumococci, that are also responsible for bacteriocin production (Lux *et al*., 2007).

1.2.3 The pneumococcus as a respiratory pathogen

The pneumococcus is able to cause a number of 'primary infections' through escaping its nasopharyngeal niche to other anatomical locations (Bogaert *et al*., 2004a). *S. pneumoniae* infections of the sinuses (sinusitis), conjunctiva (conjunctivitis) and inner ear (otitis media) are not usually life-threatening, but they have a large socioeconomic cost (Stool and Field, 1989; Klein, 2000) and there is evidence that the antibiotics used to treat these diseases increases the proportion of drug-resistant *S. pneumoniae* isolates being carried in the circulating population (Cohen *et al*., 1997). Pneumonia results when pneumococci descend into the lungs and inflame the alveoli, leading to fluid entering the air space and inhibiting oxygenation of the blood. Empyema, a complication involving the infection of the pleura or pericardium, arises in between 0.6 and 30% of pneumonias (Ravitch and Fein, 1961; Ferguson *et al*., 1996; Hardie *et al*., 1996; Byington *et al*., 2002), with a strong association with serotype 1 pneumococcal infections (Byington *et al*., 2002; Eltringham *et al*., 2003; Eastham *et al*., 2004). Penetration of the alveolar wall, allowing *S. pneumoniae* to enter the

bloodstream, results in bacteraemia (Marrie, 1992). This also allows the bacterium to metastasise and cause 'secondary infections'. These include bones (osteomylitis) and joints (arthritis) (Jacobs, 1991), the abdominal cavity (peritonitis) (Capdevila *et al*., 2001) and the kidneys, where the pathology can either be defined as nephritis or haemolytic-uraemic syndrome (Corriere and Lipshultz, 1974; Brandt *et al*., 2002). The valves of the heart can also be colonised (endocarditis) in some cases. The most severe threat to health is when the bacteria penetrate the blood-brain barrier and cause meningitis (Koedel *et al*., 2002). Other infections are seen more rarely, but include pancreatic abscesses and necrotising fasciitis (Taylor and Sanders, 1999).

Combined, these diseases kill over a million individuals annually (WHO, 2003). Detailed estimates of the global burden of pneumococcal disease in 2000 indicated around 14.5 million cases occurred in children under five, resulting in approximately 826,000 deaths (O'Brien *et al*., 2009). *S. pneumoniae* is frequently found to be the most common cause of bacterial otitis media (Klein, 1994; Bluestone and Klein, 2007) and pneumonia (Fang *et al*., 1990; Burman *et al*., 1991; Macfarlane, 1994; Ruiz *et al*., 1999; Almirall *et al*., 2000; Niederman *et al*., 2001). It is also one of the principle aetiological agents of bacteraemia (Gordon *et al*., 2001; Siegman-Igra *et al*., 2002; Valles *et al*., 2003) and meningitis (Bryan *et al*., 1990; Chotpitayasunondh, 1994; Schuchat *et al*., 1997). Despite the effectiveness of antibiotic therapies, the mortality rate for such diseases remains high: in adults in developed countries, the mortality rate for pneumococcal pneumonia is 10-15% (Feikin *et al*., 2000; Lujan *et al*., 2004; Aspa *et al*., 2006) while for meningitis it is 24-30% (although lower in infants), with a high proportion of survivors suffering neurological sequelae as a result of infection (Kalin *et al*., 2000; Auburtin *et al*., 2002; Kastenbauer and Pfister, 2003; Weisfelt *et al*., 2006; Johnson *et al*., 2007).

1.2.4 Risk factors for disease

A number of risk factors have been identified as predisposing individuals to pneumococcal disease. Some represent factors that make individuals more susceptible to colonisation, such as young age (Robinson *et al*., 2001; Tuomanen *et al*., 2004), attendance at a children's day care centre (Takala *et al*., 1995) and relatively low

socioeconomic status (Chen *et al*., 1998; Pastor *et al*., 1998). As with carriage, there is evidence that there are differences between ethnicities: black individuals have higher rates of disease relative to whites, even when correcting for levels of income (Chen *et al*., 1998; Pastor *et al*., 1998). High levels of pneumococcal disease have also been observed in native Americans (Cortese *et al*., 1992; Rudolph *et al*., 2000) and Australian Aborigines (Torzillo *et al*., 1995; Trotman *et al*., 1995), but these studies do not correct for socioeconomic factors. Once colonised, compromised immune status increases the risk of progression to pneumococcal disease: hence infections are seen disproportionately frequently in the elderly (Robinson *et al*., 2001; Tuomanen *et al*., 2004), who are not colonised at a high rate as young children are (Flamaing *et al*., 2010; Ridda *et al*., 2010), in those with HIV (Frankel *et al*., 1996; Nuorti *et al*., 2000b; Kyaw *et al*., 2005), asplenia (Chilcote *et al*., 1976; Donaldson *et al*., 1978; Foss Abrahamsen *et al*., 1997), sickle cell anaemia (Barrett-Connor, 1971; Powars *et al*., 1981) or taking immunosuppressive medications (Lipsky *et al*., 1986; Calverley *et al*., 2007; Ernst *et al*., 2007). Reduced clearance of pneumococci from the airways also increases the risk of disease: smoking (Lipsky *et al*., 1986; Nuorti *et al*., 2000a), chronic obstructive pulmonary disease (Lipsky *et al*., 1986; Kalin *et al*., 2000; Kyaw *et al*., 2005) and asthma (Talbot *et al*., 2005; Juhn *et al*., 2008) have all been found to increase the risk of pneumococcal infection. Heavy alcohol consumption is also a risk factor (Burman *et al*., 1985; Kyaw *et al*., 2005), but the importance of diabetes (Lipsky *et al*., 1986; Koivula *et al*., 1994; Kyaw *et al*., 2005) and heart disease (Lipsky *et al*., 1986; Kalin *et al*., 2000; Kyaw *et al*., 2005) remains unclear.

1.2.5 Epidemiology of pneumococcal disease

Progression from carriage to disease is usually sporadic, but there are a number of reports of pneumococcal disease outbreaks in densely populated environments. Multiple infections resulting from transmission within an establishment have been recorded in an overcrowded Texan prison (Hoge *et al*., 1994), nursing homes (Quick *et al*., 1993; Nuorti *et al*., 1998) and on an oncology ward (Berk *et al*., 1985). Non-encapsulated strains have also been recorded as causing outbreaks of conjunctivitis (Ertugrul *et al*., 1997; Martin *et al*., 2003). Most seriously, in the ‘meningitis belt’ of

sub-Saharan Africa, epidemics of serotype 1 pneumococci causing high frequencies of meningitis cases have been recorded (Leimkugel *et al*., 2005). Outbreaks of this serotype have also been recorded on South Pacific islands (Le Hello *et al*., 2010), among Aboriginal individuals in central Australia (Gratten *et al*., 1993) and in a shelter for the homeless in France (Mercat *et al*., 1991).

A number of studies have investigated the differing proclivities of *S. pneumoniae* serotypes to cause human disease. These have generally used 'odds ratios' to indicate the rates at which different serogroups caused invasive disease relative to their presence in the asymptomatically carried population. Such analyses have been performed in children in Papua New Guinea (Smith *et al*., 1993), Toronto (Kellner *et al*., 1998), Oxford (Brueggemann *et al*., 2003) and Massachusetts (Yildirim *et al*., 2010), with UK-wide disease prevalence relative to carriage frequency in Oxford (Sleeman *et al*., 2006) and by using primarily adult infection isolates relative to the population carried by children in Stockholm (Sandgren *et al*., 2004). A meta-analysis of seven studies suggested that calculated odds ratios for common serogroups were quite consistent (Brueggemann *et al*., 2004); furthermore, different lineages of the same serotype were found to have similar odds ratios, while one genetic background present as two variants with different serotypes appeared to have differing odds ratios (Brueggemann *et al*., 2003). These results indicate that serotype is more important in determining the invasiveness of an *S. pneumoniae* isolate than the rest of the genotype.

In these studies serotypes 1, 7F and 14 are often found to be invasive, whereas types 19F and serogroups 6 and 15 are more prevalent in carriage. The level of invasiveness of a serotype appears to inversely correlate with the duration of colonisation, with those serotypes carried for short periods (and therefore presumably transmitting between hosts more frequently) causing a disproportionately high level of disease (Sleeman *et al*., 2006). This may be related to the observation that progression to disease is often associated with the acquisition of new serotype in the nasopharynx (Gray *et al*., 1980). Another clinically important association is the observation that serotype 3 infections are consistently associated with a higher rate of mortality than

other capsule types (Gransden *et al*., 1985; Henriques *et al*., 2000; Martens *et al*., 2004; Harboe *et al*., 2009; Ruckinger *et al*., 2009b).

1.2.6 Structures involved in pathogenesis

A number of protein and polysaccharide structures produced by the pneumococcus have been found to be important in facilitating immune evasion, invasion across epithelial and endothelial barriers, and adhesion to tissues other than the nasopharynx.

1.2.6.1 Capsule

The polysaccharide capsule, usually negatively charged with the exception of the zwitterionic serotype 1 polymer (Tzianabos, 2000), reduces the opsonophagocytosis of pneumococci by neutrophils through inhibiting the deposition of complement on the cell surface. This is achieved through two effects: firstly, the binding of acquired immunoglobulins to subcapsular target antigens is restricted by the capsule, and secondly, it prevents the recognition of phosphoylcholine residues by C-reactive protein, thereby preventing another route that results in complement deposition (Hyams *et al*., 2010a). Although regarded as the major pneumococcal virulence factor, some *S. mitis* isolates are encapsulated (Kilian *et al*., 2008).

1.2.6.2 Teichoic acid

Teichoic acid is important in promoting adherence to, and potentially invasion of, host cells. Human Platelet Activating Factor receptor (PAFr), widely expressed by many tissues in mammals (Bito *et al*., 1994), binds its normal ligand, Platelet Activating Factor, through a phosphorylcholine moiety; hence the protein also binds pneumococci via interactions with teichoic acid (Cundell *et al*., 1995a). This interaction leads to the internalisation of the receptor and associated bacterium into the eukaryotic cell, which can lead to *S. pneumoniae* being trafficked across epithelial and endothelial barriers (Ring *et al*., 1998). This may be important for passage across the lung epithelium and the blood-brain barrier, into the CSF, during pathogenesis.

Phosphorylcholine is also important in anchoring a family of proteins, known as

choline-binding proteins (CBPs), to the surface of mitis group bacteria (Garcia *et al*., 1988). The interaction is mediated through a multiple tandem repeats of a ~20 amino acids (aa) domain found at the C terminal end of such proteins.

1.2.6.3 Pneumococcal surface protein A (PspA)

PspA is a highly variable CBP. It appears to have two roles: it prevents the binding of complement component C3 to the bacterial surface, thereby inhibiting complement-mediated opsonophagocytosis (Tu *et al*., 1999), and also binds lactoferrin, thereby ameliorating the bacteriocidal effects of the iron-depleted form of this chelator, apolactoferrin (Shaper *et al*., 2004).

1.2.6.4 Pneumococcal surface protein C (PspC)

There are two different classes of PspC alleles (Kadioglu *et al*., 2008). One is also known as Choline Binding Protein A (CbpA), which attaches to the cell surface through non-covalent interactions with phosphorylcholine residues. The other is H-binding Inhibitor of Complement (Hic), which contains an LPXTG motif and is attached to the cell surface via a sortase-dependent mechanism. Both versions are able to prevent the binding of factor H to pneumococci, again leading to the inhibition of complement-mediated opsonophagocytosis (Janulczyk *et al*., 2000; Dave *et al*., 2004; Quin *et al*., 2005).

CbpA has also been found to bind polymeric Immunoglobulin Receptor (pIgR), a host protein highly expressed in the nasopharyngeal epithelium important in secreting polymeric immunoglobulins across mucosal surfaces (Zhang *et al*., 2000). This interaction leads to the internalisation of the receptor-ligand complex by the host cell, facilitating invasion by the pneumococcus and thereby allowing transcytosis of the bacteria across mucosal epithelia. Aside from the adherence to the host surface, the importance of this pathway in colonisation is not clear.

1.2.6.5 Immunoglobulin A1 metalloprotease (ZmpA)

This integral membrane protein is a zinc metalloprotease that cleaves immunoglobulin A1 molecules, the most common form of immunoglobulin secreted into the nasopharynx, attached to the surface of the bacterium (Wani *et al*., 1996). This prevents the triggering of the inflammatory response following antibody binding.

1.2.6.6 Pneumococcal serine-rich repeat protein (PsrP)

PsrP is a large, typically 4,000-5,000 aa, integral membrane serine-rich repeat glycoprotein. Not present in all pneumococcal genomes, it is encoded on an island along with a series of glycosyl transferases, likely to post-translationally modify the protein, and a secretory apparatus. The protein has been found to bind keratin 10, expressed by lung cells but not those lining the nasopharynx, and, via a separate domain, mediate pneumococcal aggregation in biofilms (Shivshankar *et al*., 2009; Sanchez *et al*., 2010).

1.2.6.7 Pneumococcal collagen-like protein A (PclA)

PclA is large, typically ~2,000 aa, sortase-anchored protein encoded by an island not found in all pneumococcal genomes. Its presence leads to increased adherence of pneumococci to both nasopharyngeal and lung epithelial cells (Paterson *et al*., 2008).

1.2.6.8 Pneumolysin

Pneumolysin is a cholesterol-activated cytolysin, which undergoes substantial structural rearrangements and oligomerises into a complex of around 40 subunits on contact with a cholesterol-containing membrane of a eukaryotic cell, forming a pore with a diameter of around 260 Å (Tilley *et al*., 2005). At sublytic levels, the protein is reported to inhibit ciliary beating, reduce the level of bactericidal free radicals produced by human monocytes and bind complement (Nandoskar *et al*., 1986; Mitchell *et al*., 1991; Hirst *et al*., 2004). However, unlike related toxins, pneumolysin has no signal sequence, hence is not secreted but instead confined to the pneumococcal cytosol until the bacteria lyse (Walker *et al*., 1987).

1.2.6.9 Autolysin (LytA)

This CBP lyses the *N*-acetyl-muramoyl-L-alanine bonds of the peptidoglycan bacterial cell wall, in order to allow for cell growth and remodelling (Howard and Gooder, 1974). However, this activity causes the release of pneumolysin, as well as inflammatory peptidoglycan and teichoic acid fragments, leading to increased damage to host tissues (Canvin *et al*., 1995; Berry and Paton, 2000).

1.3 Pneumococcal genomics and epidemiology

1.3.1 The genome of *S. pneumoniae* TIGR4

Dideoxy terminator sequencing of Firmicutes is relatively difficult due to the high proportion of the genome that cannot be cloned into *Escherichia coli* (Sorek *et al*., 2007). The sequence of *S. pneumoniae* TIGR4, a serotype 4 isolate from the blood of a 30 year old male patient in Denmark, was published in 2001 (Tettelin *et al*., 2001), six years after the first bacterial genome (Fleischmann *et al*., 1995). The annotation of the 2,160,837 bp sequence, which had a GC content of 39.7%, included 2,236 protein coding sequences (CDSs), of which a high proportion (3.7%), relative to previously published genomes, were IS elements. Small interspersed repeats were also identified at a high density in the chromosome: 127 BOX elements (Martin *et al*., 1992), modular repeats consisting of variable arrangements of boxA, boxB and boxC subsequences, and 108 RUP (Repeat Unit of Pneumococcus) elements (Oggioni and Claverys, 1999), approximately palindromic ~107 bp sequences. Four rRNA operons were present, associated with a total of 46 tRNAs, with a further 12 tRNAs elsewhere in the genome. Another notable feature, uncovered by the shotgun sequencing, is the *hsdS* type I restriction-modification system locus. Along with the functional *hsdS* gene, there are two incomplete versions comprising just the C terminal sequence of the enzyme; all three CDSs are associated with inverted repeats, which are acted upon by a site-specific recombinase encoded by an adjacent gene. Recombinations between the sequences result in the generation of a different functional, expressed protein. Hence the specificity of the *hsdS* system can be altered over very short timescales, which was hypothesised to inhibit the transfer of DNA between clonally related strains.

The genome contained a large number of transporters for the uptake of fermentable sugars, along with a range of enzymes likely to be important for the degradation of host polymers, such as mucins and glycolipids, to release monosaccharides. Most of these transporters are of the ATP-binding cassette (ABC) or phosphoenolpyruvate (PEP)-dependent phosphotransferase system (PTS) types. The prevalence of these transporters, driven by the hydrolysis of ATP or PEP respectively, likely reflects the reliance of pneumococci on substrate-level phosphorylation for the production of ATP, rather than the maintenance of proton motive force that can power transporters that use ion gradients (Paulsen *et al*., 2000). ABC transporters, in particular, are energetically inefficient compared to such ion-driven systems, but have the advantage of very high affinities for their substrates (Poolman, 1993).

1.3.2 Functional genomics experiments

The genome sequence was crucial for advancing functional genetics studies of the pneumococcus. Signature tagged mutagenesis experiments, which screened libraries of randomly generated mutants for their ability to cause disease in the mouse model of infection, highlighted a number of loci whose function was important for pathogenesis (Polissi *et al*., 1998; Lau *et al*., 2001). By performing such a screen in *S. pneumoniae* TIGR4, which was chosen on the basis of its high virulence in the mouse model of disease and susceptibility to all antibiotics in order to make it as genetically tractable as possible, greater interpretation of positive results relating to CDSs of unknown function was possible (Hava and Camilli, 2002). For instance, two previously unidentified genes were found to be important in maintaining nasopharyngeal carriage and causing pneumonia, but not bacteraemia; these delineated a gene cluster encoding of a transcriptional regulator, three sortases (which attach secreted proteins to the external surface of the cell wall) and their putative substrate proteins. The structure produced by these CDSs was later found to be a pilus (Barocchi *et al*., 2006), which has subsequently emerged as the prime candidate for mediating the antagonistic relationship with *Staph. aureus* (Regev-Yochay *et al*., 2009) and proved to be a promising vaccine target (Gianfaldoni *et al*., 2007). Oddly, none of these three screens identified the capsule genes as a virulence factor (Hava and Camilli, 2002).

The availability of a complete sequence allowed the construction of microarrays, with oligonucleotide probes corresponding to each CDS in the chromosome, for genome-wide expression analyses. This system has again been used to study pneumococcal virulence by comparing RNA extracts from a rough derivative of *S. pneumoniae* TIGR4 co-cultured *in vitro* with a human pharyngeal cell line and the encapsulated version of the strain growing in rabbit CSF in an animal model of meningitis (Orihuela *et al*., 2004). This showed that LytA, pneumolysin and the pyruvate oxidase *spxB* were downregulated in the CSF; all three of these genes have been implicated in causing inflammation in meningitis, through releasing immunogenic peptidoglycan fragments, lysing host cells and causing oxidative damage through hydrogen peroxide generation, respectively. However, because the microarray is based on TIGR4, it is difficult to perform an equivalent study on samples from humans due to the genetically heterogeneous nature of clinical isolates.

1.3.3 Variation uncovered by dideoxy terminator sequencing

Microarrays have also proved useful in looking at the variation between strains through comparative genomic hybridisation (CGH). Two studies have taken different approaches to using this technique to identify loci contributing to the virulence of invasive pneumococcal lineages. Obert *et al* compared strains of serotypes 6A, 6B and 14 isolated from nasopharyngeal carriage and disease; within each serotype, the hierarchical clusterings derived from CGH were used to divide strains into 'invasive' or 'noninvasive' clades, and the differences in genome content between these groups analysed (Obert *et al*., 2006). This identified two loci as being associated with disease isolates: one locus encoding a V-type sodium ion-driven ATPase and neuraminidase, and a second corresponding to the *psrP* island. Instead of looking for differences in virulence while controlling for serotype, Blomberg *et al* analysed accessory genomic loci differing between 13 serotypes with varying propensities to cause invasive disease (Blomberg *et al*., 2009). This study also found the *psrP* island to be associated with more invasive serotypes, along with a locus encoding a 6-phospho-β glucosidase; however, while knock out of *psrP* was observed to reduce the virulence of *S. pneumoniae* TIGR4 in the mouse model, no such effect was observed following

the disruption of the 6-phospho-β glucosidase locus (Obert *et al*., 2006; Blomberg *et al*., 2009).

However, CGH is always subject to the limitation of the microarray design; divergent loci and novel, previously unsequenced, regions of the chromosome cannot be detected. Hence there is a need for multiple strains to be sequenced to represent such a diverse species as *S. pneumoniae*. Two further genomes were published in the same year as TIGR4. One was a draft sequence of the serotype 19F multidrug resistant isolate *S. pneumoniae* G54 (Dopazo *et al*., 2001), which appeared to have a highly disrupted synteny relative to that of TIGR4. However, this was an artefact of the assembly, and a corrected sequence has since been completed that has a similar chromosomal structure to other pneumococci [EMBL accession code CP001015].

The other was the sequence of an extant descendent of the rough strain used in Avery's transformation experiment, *S. pneumoniae* R6 (Hoskins *et al*., 2001); six years later, the genomes of two laboratories' versions of the serotype 2 progenitor, *S. pneumoniae* D39, were published (Lanie *et al*., 2007). Just two single base differences in sequence length (indels) and ten single nucleotide polymorphisms (SNPs) differentiated the two isolates of *S. pneumoniae* D39. In addition to the 7,505 bp deletion in *S. pneumoniae* R6 that removes the *cps* locus, the loss of the cryptic plasmid pDP1, nine indels and 71 SNPs distinguished R6 from the D39 strains. Expression analysis using a microarray based on the genome of *S. pneumoniae* R6 revealed a number of transcriptional differences between the rough and smooth strains, with the former showing increased transcription of a number of genes involved in competence for DNA transformation; however, there was no clear relationship between many of the changes in expression and sequence polymorphisms. None of the observed mutations involved the transposition of any of the IS, BOX or RUP elements, indicating these are stable features of the genome.

1.3.4 Epidemiological typing techniques

Given the difficulties and expense of sequencing pneumococcal chromosomes, other techniques have been used for characterising the relationships between isolates in large collections.

1.3.4.1 Serotyping

The oldest method for typing pneumococci is using the capsule, traditionally performed using the *Quellung* (German for 'swelling') reaction first described by Neufeld (Neufeld, 1902). Antisera are sequentially mixed with the bacteria until a positive reaction, indicated by the agglutination of the mixture, is observed. This method can be used either on single strains, or on a sweep of colonies off an agar plate (for instance, from a nasopharyngeal swab), resuspended in saline and mixed with antisera attached to latex beads, in order to detect the presence of multiple serotypes within a mixture of strains (Hill *et al*., 2008). The publication of the *cps* locus sequences (Bentley *et al*., 2006) allowed the development of a multiplex PCR scheme for differentiating serotypes (Pai *et al*., 2006) and the implementation of a typing microarray, which is capable of detecting and quantifying different serotypes present within a mixed sample (Turner *et al*., 2011).

Serotyping is a poor method for ascertaining the relationships between strains, because it is based on a single genetic locus, hence it easily confounded by recombination events. However, it still provides important information regarding the likely invasiveness of a strain, and its susceptibility to capsule-based vaccines.

1.3.4.2 Multilocus enzyme electrophoresis (MLEE)

MLEE was developed for studying polymorphism in the human population (Harris, 1966). It requires electrophoretic separation of cell lysates, followed by multiple assays, each specific for a particular enzyme, using chromatogenic substrates (Selander *et al*., 1986). The position of the resultant staining on the gel reflects the properties of the enzyme, allowing distinguishing polymorphisms in the protein to be observed. This technique was applied to multidrug-resistant serotype 23F and serogroup 19 isolates of *S. pneumoniae*, demonstrating that they formed a single lineage that had spread from Europe to the USA (Coffey *et al*., 1991; Munoz *et al*., 1991).

1.3.4.3 Pulsed field gel electrophoresis (PFGE)

PFGE was first used to characterise populations of *Saccharomyces cerevisiae* (Schwartz and Cantor, 1984). For bacteria, this method involves digestion of genomic DNA with an infrequently cutting restriction enzyme (for instance, *Sma*I or *Apa*I produce an appropriate number of fragments with *S. pneumoniae*), followed by electrophoretic separation of the digest fragments in an alternating electric field (McClelland *et al*., 1987). This method, which does not require the range of biochemical assays needed for MLEE, was originally applied to pneumococci in order to track the spread of antibiotic-resistant isolates in Europe (Figueiredo *et al*., 1995; Tarasi *et al*., 1995) and America (Barnes *et al*., 1995; Moreno *et al*., 1995).

1.3.4.4 BOX PCR

BOX PCR uses a primer that binds within the boxA module of BOX elements to amplify segments of the chromosome that lie between closely spaced, inverted BOX repeats; the products of this reaction can then be electrophoretically separated in order to produce a pattern of bands characteristic of a strain (van Belkum *et al*., 1996). This method was first applied to following an outbreak of non-typeable strains causing conjunctivitis (Ertugrul *et al*., 1997) and to classifying penicillin-resistant isolates in Texas (Rodriguez-Barradas *et al*., 1997) and the Netherlands (Hermans *et al*., 1997).

1.3.4.5 Multilocus sequence typing (MLST)

MLST was originally applied to *N. meningitidis* (Maiden *et al*., 1998). The method involves sequencing ~450 bp loci within multiple unlinked housekeeping genes around the chromosome; each different sequence at a given locus is assigned a unique allele number, with the overall ‘sequence type’ (ST) referring to a specific combination of alleles. The scheme developed for *S. pneumoniae* uses seven loci (*aroE*, *gdh*, *gki*, *recP*, *spi*, *xpt* and *ddl*) (Enright and Spratt, 1998), although the *ddl* locus is often omitted from analyses as it is linked to the penicillin binding protein *pbp2B*, resulting in the ‘hitchhiking’ of divergent *ddl* sequences as *pbp2B* alleles causing penicillin resistance are imported from other species (Enright and Spratt, 1999).

There are considerable advantages to using sequences for epidemiology. Firstly, unlike all the electrophoresis-based methods, results can be directly compared between studies; online facilities have been created for such data collation (Aanensen and Spratt, 2005). Secondly, the information is appropriate both for short term classification of strains during epidemics, for which PFGE or BOX PCR have sufficient resolution, and for ascertaining more distant relationships between isolates, where MLEE is more appropriate, due to the slower evolution of protein sequences relative to restriction enzyme cut sites or repeat elements (Maiden *et al*., 1998). Thirdly, more sophisticated evolutionary analyses can be performed on sequence data than on the type of discrete data produced by MLEE, PFGE or VNTR. For instance, when comparing closely related isolates, the precise level of divergence between each of the loci is known, hence transformation events that introduce high densities of polymorphisms can be detected. A study of the divergent alleles distinguishing single locus variants (SLVs), classifying those that differed by three or more SNPs as having arisen through a recombination event, estimated that the ratio of polymorphisms introduced through recombination to those generated by point mutation (the *r/m* ratio) was ~66 (although this dropped to 45 when *ddl* was excluded) (Feil *et al*., 2000).

1.4 The emergence of antibiotic-resistant pneumococci

As indicated above, the development of the epidemiological typing techniques was primarily motivated by the development of antibiotic resistance that became prevalent among pneumococci in the late 1970s. This marked the end of a transition from the period in which the species was universally susceptible to a number of highly effective drugs, to one in which a number of multidrug-resistant clones became prevalent worldwide.

1.4.1 Early observations of resistance

The first antibiotic to be used to treat pneumococcal infections was optochin, employed against conjunctivitis (Reber, 1917), pneumonia (Moore and Chesney, 1917) and empyema (Lowenburg, 1929), but it was a poor antimicrobial due to the specificity of its action against *S. pneumoniae* and the frequency with which it caused

loss of vision as a side-effect. Furthermore, resistant bacteria were found to evolve very rapidly on exposure to the chemical *in vitro*, during experimental infection or clinical treatment (Morgenroth and Kaufmann, 1912; Moore and Chesney, 1917; Ash and Solis-Cohen, 1929). Unfortunately, sulphanilamide, an early sulphonamide drug and hence one of the first broad-spectrum antibiotics, proved relatively ineffective against pneumococci compared to other streptococci (Long and Bliss, 1937). Hence the need remained for an effective anti-pneumococcal drug, leading to the development of an alternative sulphonamide, sulphapyridine (Whitby, 1938), which was successfully applied in treatment of pneumonia (Evans and Gaisford, 1938). However, resistance was again readily observed in laboratory settings (Ross, 1939), and it was not long before insensitive pneumococci were observed in patients being treated with such antibiotics (Lowell *et al*., 1940; Hamburger *et al*., 1943).

1.4.2 Resistance to β lactams

The first patient to show any benefit from treatment with penicillin by Fleming, shortly after its initial discovery in 1929, appears to have been an individual with pneumococcal conjunctivitis (Watson *et al*., 1993), although the drug was not widely available for larger scale clinical applications until 1943 (Keefer *et al*., 1943). Pneumococci were found to be highly sensitive to β lactams, which proved very effective in treating such infections, although once more, tolerance of the antibiotic was soon demonstrated to be possible *in vitro* (McKee and Houck, 1943; Schmidt and Sesler, 1943). However, no clinically important resistance was observed for decades after the introduction of the penicillins, during which time another class of β lactam antibiotics, the cephalosporins, were introduced as another effective treatment for pneumococcal disease (Murdoch *et al*., 1964; Thornton and Andriole, 1966). This resulted in a significant drop in interest in the pneumococcus as a pathogen throughout the middle decades of the 20th century (Figure 1.1) (Powel, 2004).

The first report of clinically relevant penicillin resistance in *S. pneumoniae* concerned a serogroup 23 strain from Australia (Hansman and Bullen, 1967), although there is evidence there may have been some isolates with increased tolerance of β lactams previously (Kislak *et al*., 1965). In the same year, a penicillin-resistant serogroup 6

strain was isolated from the same region of Australia, and shortly afterwards a number of resistant serotype 4 isolates were isolated in Papua New Guinea, where penicillin was being used as prophylaxis against the high rates of pneumococcal infection (Hansman *et al*., 1971). By 1978, one-third of clinical isolates from Papua New Guinea were found to be penicillin-resistant (Gratten *et al*., 1980). Meanwhile, resistant strains were being found all over North America: Canada (Dixon *et al*., 1977), Boston (Finland *et al*., 1976), Pittsburgh (Ahronheim *et al*., 1979) and Wisconsin (Maki *et al*., 1980) all reported penicillin-insensitive pneumococci, while in New Mexico they were found to comprise over 14% of *S. pneumoniae* isolated from native Americans (Tempest *et al*., 1974) and in Oklahoma they accounted for 15% of clinical isolates (Saah *et al*., 1980). Other foci of emerging resistance were South Africa, where penicillin-resistant isolates were found being carried by 29% of paediatric patients in Johannesburg hospitals (Jacobs *et al*., 1978), and certain countries in Europe: in Hungary in the late 1980s 70% of *S. pneumoniae* clinical isolates from children were insensitive to β lactams (Marton *et al*., 1991), while around the same time in Spain up to 44% of pneumococcal clinical isolates were penicillin-resistant strains (Fenoll *et al*., 1991; Pallares *et al*., 1995).

Penicillin acts on bacteria through behaving as an inhibitor of multiple penicillin-binding proteins (PBPs), which together function to remodel the peptidoglycan of the cell wall to allow for cell growth, remodelling and division (Spratt, 1975; Spratt and Pardee, 1975). *S. pneumoniae* has six PBPs: PBP1A, PBP1B, PBP2A, PBP2B, PBP2X and PBP3 (Hakenbeck *et al*., 1986), and resistance occurs following the acquisition of alleles that are still able to function while having a decreased affinity for the relevant antibiotic (Hakenbeck *et al*., 1980; Zighelboim and Tomasz, 1980). Different β lactams target the PBPs to varying extents: resistance to oxacillin only requires changes to PBP2X (Dowson *et al*., 1994), but mutations in this gene only confer low-level resistance to cephalosporins (Laible *et al*., 1989; Laible and Hakenbeck, 1991), with higher resistance following additional changes in PBP1A (Munoz *et al*., 1992). Alterations in both of these genes leads to low-level penicillin resistance, with a more highly resistant phenotype developing with changes to PBP2B (Williamson *et al*., 1980; Barcus *et al*., 1995). The MurM protein, involved in synthesising cross-links between peptidoglycan chains, also appears to have been

involved in the high-level penicillin resistance of some isolates from South Africa and Eastern Europe in the 1970s and 1980s (Filipe and Tomasz, 2000; Smith and Klugman, 2001). The alteration of this gene causes a dramatic restructuring of the cell wall (Garcia-Bustos *et al*., 1988; Garcia-Bustos and Tomasz, 1990), hence extant penicillin-resistant isolates typically lack altered *murM* genes and the associated altered peptidoglycan structures (Filipe *et al*., 2000).

Sequence comparisons indicate that these resistant PBP forms have been generated through the recombination of fragments from *S. mitis* and *S. oralis* into the original *S. pneumoniae* version to yield a new 'mosaic' form of the protein (Dowson *et al*., 1989; Dowson *et al*., 1993; Sibold *et al*., 1994). The analogous situation was also observed among *Neiserria*, where *N. meningitidis* developed resistance to penicillin following the acquisition of PBP sequence from the related commensal *N. flavescens* (Spratt *et al*., 1989). Although within each penicillin-resistant pneumococcal lineage the PBP alleles were conserved (Jabes *et al*., 1989; Munoz *et al*., 1991), the diversity between such lineages relative to the conservation of these proteins in pencillin-sensitive strains suggested multiple, independent acquisitions of the resistant phenotype (Markiewicz and Tomasz, 1989; Hakenbeck *et al*., 1991b; Hakenbeck *et al*., 1991a). These alleles have also been found to be spreading from pneumococci into other nasopharyngeal streptococci (Dowson *et al*., 1990; Coffey *et al*., 1993).

1.4.3 Resistance to other antibiotics

Shortly after the introduction of penicillin, three other types of antibiotics, each of which targeted the bacterial translational machinery, were reported as being highly effective in the treatment of pneumococcal disease: the tetracycline aureomycin in 1948 (Collins *et al*., 1948), chloramphenicol in 1950 (Riley, 1950) and the macrolide erythromycin in 1953 (Austrian and Rosenblum, 1953). However, universal susceptibility to these three drugs among pneumococci lasted only a decade. Tetracycline resistance was observed in Australia in 1963 (Evans and Hansman, 1963), while erythromycin-resistant *S. pneumoniae* was reported the same year as penicillin-resistant strains were isolated (Kislak, 1967; Weisblum, 1967). Chloramphenicol resistance took longer to emerge, with the first observation, from

France, not until 1973 (Dang-Van *et al*., 1978), although a survey in 1970 detected some loss of sensitivity (Cybulska *et al*., 1970).

In the late 1960s the combination of sulphonamides with the synthetic anti-folate trimethoprim, a mixture named co-trimoxazole, was used to treat pneumococcal infections (Hughes, 1969). Again, resistance was detected within a few years (Howe and Wilson, 1972). As with optochin and penicillin, resistance to co-trimoxazole results from changes in the sequence of the drug targets leading to them having decreased affinity for the antibiotics: dihydrofolate synthase in the case of sulphonamides (Wolf and Hotchkiss, 1963; Ortiz, 1970), and dihydrofolate reductase in the case of trimethoprim (Pikis *et al*., 1998). Another example is rifampicin: not typically used to treat *S. pneumoniae* infections, throughout the 1980s pneumococcal resistance was seen to increase in South Africa, where the drug is widely administered to individuals with tuberculosis (Klugman and Koornhof, 1988a; Klugman and Koornhof, 1988b). Rifampicin insensitivity results from base substitutions in the *rpoB* gene, which encodes a subunit of the target of the drug, RNA polymerase (Enright *et al*., 1998).

By contrast, the other resistances result not from drug target sequence changes, but instead from the acquisition of specific resistance genes. Chloramphenicol is inactivated on entry into the cell by the *cat* acetyltransferase in resistant *S. pneumoniae* (Dang-Van *et al*., 1978), and similarly phosphotransferases that modify and inactivate aminoglycosides have been found in some pneumococci (Collatz *et al*., 1984), despite their high intrinsic tolerance of these antibiotics (Ward, 1981). Protection against tetracycline is afforded by the *tet* genes that associate with the ribosome and it prevent binding the antibiotic (Sanchez-Pescador *et al*., 1988; Connell *et al*., 2003). Macrolide resistance is a consequence of the *mel*/*mef* efflux pump or the *erm* methylases that modify the target rRNA; this latter mechanism also provides resistance against lincomycin and streptogramin B, structurally distinct antibiotics with the same target (Courvalin *et al*., 1985; Sutcliffe *et al*., 1996; Gay and Stephens, 2001).

1.4.4 The spread of the PMEN clones

By the 1970s, resistance to all the major anti-pneumococcal chemotherapies had been observed. A fresh treatment option became available the 1980s, when fluoroquinolone antibiotics were introduced for the treatment of bacterial infections. The first generation of fluoroquinolones, such as ciprofloxacin, exhibited relatively poor activity against *S. pneumoniae* (Wijnands *et al*., 1986; Thys *et al*., 1989), but the second generation introduced in the 1990s, such as sparfloxacin, were a more effective treatment of pneumococcal infections (Pankuch *et al*., 1995; Thornsberry *et al*., 1999). However, reports of resistance emerging during treatment rapidly emerged (Mehtar *et al*., 1990; Perez-Trallero *et al*., 1990; Ball, 1994), followed by surveys finding high prevalence of resistance in pneumococcal populations (Goldstein and Acar, 1996; Goldsmith *et al*., 1998). These were consequences of resistance resulting from single base changes in the genes encoding target topoisomerases: first-step mutations to give low level resistance occur in *parC* or *parE*, with subsequent mutations in *gyrA* resulting in more highly resistant phenotype (Janoir *et al*., 1996; Munoz and De La Campa, 1996; Tankovic *et al*., 1996; Perichon *et al*., 1997). Resistance can also result from the upregulation of the PmrA and PatAB efflux pumps (Gill *et al*., 1999; Marrer *et al*., 2006). Hence the only antibiotic that remains effective against all pneumococci is vancomycin, the first reported use of which to treat a pneumococcal infection in an adult was not until 1981 (Garau *et al*., 1981). However, tolerance can be selected *in vitro* (Novak *et al*., 1999) and has been observed in clinical isolates (McCullers *et al*., 2000; Henriques Normark *et al*., 2001; Moscoso *et al*., 2010).

Of further clinical concern was the accumulation of multiple resistances in single strains. In 1977, a serotype 19A strain resistant to penicillin, tetracycline, erythromycin, clindamycin, chloramphenicol and co-trimoxazole was observed in South Africa (Jacobs *et al*., 1978). Over the next few years, strains with similarly extensive resistance profiles were found in the UK, USA and continental Europe (Dublanchet and Durieux, 1979; Radetsky *et al*., 1981; Williams *et al*., 1981). The Pneumococcal Molecular Epidemiology Network (PMEN) was set up to track the epidemiology of these lineages meeting the criteria of being antibiotic resistant and having a wide geographical distribution (Klugman, 1998; McGee *et al*., 2001). The

first three clones, PMEN1 (Spain^{23F}-1), PMEN2 (Spain^{6B}-2), and PMEN3 (Spain^{9V}-3), were all thought to have originated in Spain, where they were found to be the most common lineages causing meningitis in the late 1990s (Enright *et al*., 1999). As of 2011, there are 43 PMEN clones, with the expanded remit that some globally disseminated strains are included despite commonly being entirely antibiotic sensitive (McGee and Klugman, 2011).

1.5 Horizontal sequence exchange in the pneumococcus

The majority of pneumococcal resistance mechanisms involve horizontal sequence transfers. There are three main mechanisms by which DNA passes between bacteria: transformation, the uptake of exogenous DNA from the environment; transduction, the phage-mediated transfer of DNA between bacteria; and conjugation, the movement of DNA between cells in direct contact driven by mobile elements.

1.5.1 The pneumococcal competence system

S. pneumoniae has a dedicated system for the uptake of exogenous DNA. A competence pseudopilus appears to mediate the first interaction with exogenous dsDNA (Campbell *et al*., 1998; Pestova and Morrison, 1998). The DNA then permeates the cell wall, perhaps driven by pseudopilus retraction, and contacts the uptake pore complex (Chen *et al*., 2005). This contains nucleases that degrade one strand of the dsDNA, while nicking the backbone of the other strand as it is imported in a 3'→5' direction (Morrison and Guild, 1972; Lacks and Neuberger, 1975; Mejean and Claverys, 1988). The ssDNA-binding protein RecA is loaded onto the strand in the cytosol with the aid of DprA (Mortier-Barriere *et al*., 2007), and the resulting nucleoprotein filament is able to invade the cell's dsDNA (Chen *et al*., 2008). Strand exchange events then lead to incorporation of the imported DNA into the genome, which appears to take about 15 mins on the evidence of pulse-chase experiments with radiolabelled DNA (Mejean and Claverys, 1984; Berge *et al*., 2003).

The competence state is tightly regulated, primarily by an extracellular signalling mechanism (Tomasz and Mosser, 1966). The signal is generated by the ComAB

transporter, which cleaves the 41 aa ComC peptide at a Gly-Gly bond as it is exported to yield the 17 aa extracellular pheromone Competence Stimulating Peptide (CSP) (Havarstein *et al*., 1995; Claverys and Havarstein, 2002). CSP is detected by the ComDE two-component system (Pestova *et al*., 1996), which initiates a variety of transcriptional responses when stimulated, including activation of the competence genes and a number of stress response systems, that lead to a physiological state termed the 'X state' (Claverys *et al*., 2006). There are two phases to the response: 'early' genes, which are preceded by a direct repeat bound by ComE, and 'late' genes, which have a 'combox' in their promoter recognised by one of the early genes, the alternative σ factor ComX (Alloing *et al*., 1998; Peterson *et al*., 2000). Positive feedback maintains the X state, as *comABCDE* are among the early genes, whereas the genes for the competence pseudopilus, DNA uptake pore and ssDNA binding are in the late class (Dagkessamanskaia *et al*., 2004; Peterson *et al*., 2004). Another four late genes, the bacteriocin *cibAB* and the murein hydrolases *lytA* and *cbpD*, are involved in the lysis of cells to release DNA: the autolytic activity of these proteins is prevented by the transcription of *cibC*, providing immunity from the bacteriocin, and the early gene *comM*, which prevents *lytA* and *cbpD* acting on the host cell, respectively (Guiral *et al*., 2005; Havarstein *et al*., 2006). Hence cells in the X state are able to lyse nearby related cells not exhibiting the same response to CSP, a process termed 'fratricide' (Claverys *et al*., 2007). Two distinct alleles of CSP system (and the cognate receptor) have been detected in the *S. pneumoniae* population (Pozzi *et al*., 1996), with the consequence that mixed pneumococcal populations containing different 'pherotypes' may lead to one strain eliminating the other.

Other signals also appear to be involved in regulating competence. An upward shift in the pH of the growth medium has been found to trigger the development of competence (Tomasz, 1966). DNA damaging agents, such as mitomycin C and fluoroquinolones, also promote the development of the competent state (Prudhomme *et al*., 2006), although this response is not seen in *H. influenzae* or *B. subtilis* (Redfield, 1993a). By contrast, the CiaRH two-component system, which seems to be regulated by the integrity of the cell wall, inhibits the development of competence via an unknown mechanism involving the HtrA serine protease (Guenzi *et al*., 1994; Sebert *et al*., 2005). Disruption of the oligopeptide transporter lipoproteins Ami-

AliAB appears to promote the development of competence, leading to the suggestion that high levels of free amino acids within the cell indicate adequate nutrient availability, and hence repress the X state (Claverys *et al*., 2000). Similarly, elevated intracellular purine levels have been suggested to repress the X state, as competence is upregulated by the disruption of purine biosynthesis genes such as *purA*, *guaA* and *guaB* (Claverys and Havarstein, 2002).

The size of the DNA fragments incorporated into the genome via this system was originally measured using linked markers separated by a known distance, which estimated a mean length of ~2 kb (Lacks, 1966), and through the mass of isotopically labelled integrated donor DNA; experiments with ^{32}P labelled DNA suggested the mean lay in the range of 3-6 kb (Fox and Allen, 1964), while later work with ^{3}H and ^{15}N labelled DNA resulted in a more precise estimate towards the lower end of this range (Gurney and Fox, 1968). Subsequent estimates from MLST data suggest recombinations have a mean size of ~4.4 kb (Feil *et al*., 2000). These were reassuringly similar to, or at least smaller than, the size of the donor ssDNA entering the cell, which had a median size of ~6.7 kb prior to integration into the chromosome (Morrison and Guild, 1972). Much larger events are possible, however: transfer of ~39 kb, involving the *cps* locus and both flanking *pbp* genes, has been detected in a clinical isolate. This recombination conferred a change of serotype, resulting in vaccine escape, and penicillin resistance in a single recombination, and therefore is likely to be a rare event of unusual magnitude preserved by high levels of selection (Brueggemann *et al*., 2007).

1.5.2 Transduction and pneumophage

Surveys of clinical isolates have found that a large proportion of the pneumococcal population are lysogenic (Ramirez *et al*., 1999; Romero *et al*., 2009). The first reported isolations of prophage infecting *S. pneumoniae* date to 1975 (McDonnell *et al*., 1975; Tiraby *et al*., 1975). The former study also demonstrated that pneumococci with altered teichoic acid containing ethanolamine in place of choline were resistant to infection by 'Diplophage-1' (ϕDp-1), and subsequent work found adhesion of ϕDp-1 to *S. pneumoniae* was inhibited by the presence of free choline in the medium (Lopez *et al*., 1982), suggesting the phage used choline or a CBP as a receptor. As a

consequence, the capsule appears to inhibit phage adsorption (Bernheimer and Tiraby, 1976).

Four independent complete pneumophage genomes have been sequenced: the temperate phage ϕMM1 (Obregon *et al*., 2003) and ϕMM1-1998 (Loeffler and Fischetti, 2006), and the lytic phage ϕCp-1 (Martin *et al*., 1996) and ϕDp-1 (Sabri *et al*., 2011). While the two lytic phage were diverse, differing in size by about 37 kb, the temperate phage displayed a similar genetic organisation to other lysogenic streptococcal viruses (Lopez and Garcia, 2004), with no evidence of antibiotic resistance or virulence factor transduction. This involves division of the 30-40 kb genome into five modules: in order, the clusters of genes for maintaining lysogeny (including the integrase), DNA replication, DNA packaging, phage construction and host cell lysis. This last module includes an autolysin, which is a CBP in sequenced pneumophage and appears to exchange sequence with the host *lytA* gene (Lopez *et al*., 1992; Sheehan *et al*., 1997; Whatmore and Dowson, 1999). The prophage ϕMM1-1998 also appears to affect bacterial surface structures as it promotes increased adhesion to a human pharyngeal cell line (Loeffler and Fischetti, 2006).

1.5.3 Conjugative elements

Two types of conjugative element are found in pneumococci: plasmids, typically circular, extra-chromosomal elements; and integrative and conjugative elements (ICEs), which are inserted into the chromosome. Just two types of *S. pneumoniae* plasmid have been characterised, both of them cryptic: pDP1 (3,160 bp) (Smith and Guild, 1979) and pSpnP1 (5,413 bp) (Romero *et al*., 2007). Several surveys have identified a small number of plasmids that might represent novel types, but pDP1 is commonly re-isolated, indicating a genuine paucity of diversity in the pneumococcal plasmid population (Berry *et al*., 1989; Sibold *et al*., 1991). Consequently there are now five almost identical *S. pneumoniae* pDP1-like sequences available (Cortaza *et al*., 1983; Schuster *et al*., 1998; Munoz *et al*., 1999; Oggioni *et al*., 1999).

By contrast, ICEs are more diverse and heavily associated with antibiotic resistance. Using strain *S. pneumoniae* BM6001, resistant to both chloramphenicol and

tetracycline, it was shown that the clinically important resistances were co-transferred with a known novobiocin-resistance marker in transformation experiments, indicating they were integrated into the chromosome (Shoemaker *et al*., 1979). Subsequent filter-mating experiments demonstrated that the tetracycline and chloramphenicol resistance markers could be transferred between pneumococci in a DNase-insensitive manner, showing they were carried by a conjugative element (Shoemaker *et al*., 1980). A contemporaneous survey of resistant *S. pneumoniae* found that resistance determinants to macrolides, lincosamides and aminoglycosides, but not those to penicillin, trimethoprim or sulphonamides, were also ICE-borne (Buu-Hoi and Horodniceanu, 1980). In the transformation experiments, the chloramphenicol and tetracycline markers were linked, but in an asymmetrical manner whereby the tetracycline resistance marker was almost always associated with chloroamphenicol resistance but not *vice versa* (Shoemaker *et al*., 1979), suggesting they were on proximate but distinct elements. Mapping of the transposon Tn*5253* in *S. pneumoniae* BM6001 showed these resistances were carried about ~25 kb apart on an ICE ~70 kb in total length (Vijayakumar *et al*., 1986). Sequencing fragments of this element subsequently showed Tn*5253* was a composite of Tn*5251*, which carried the tetracycline resistance gene and was later shown to be almost identical to Tn*916* (Provvedi *et al*., 1996), inside a larger transposon, Tn*5252*, which carried chloramphenicol resistance (Ayoubi *et al*., 1991). Both elements retained their independent conjugative abilities. The Tn*916* family of elements, first discovered in *Enterococcus faecalis* in the 1970s (Roberts and Mullany, 2009), are known to exhibit considerable plasticity; examples found in pneumococci have subsequently been found to sometimes harbour genes causing resistance to kanamycin (Poyart-Salmeron *et al*., 1991) and macrolides (McDougal *et al*., 1998; Seral *et al*., 2001; Cochetti *et al*., 2007; Cochetti *et al*., 2008). Chloramphenicol resistance remains associated with the Tn*5252* family of ICEs, which acquired the requisite acetyltransferase through integration and linearisation of the staphylococcal pC194 resistance plasmid (Widdowson *et al*., 2000).

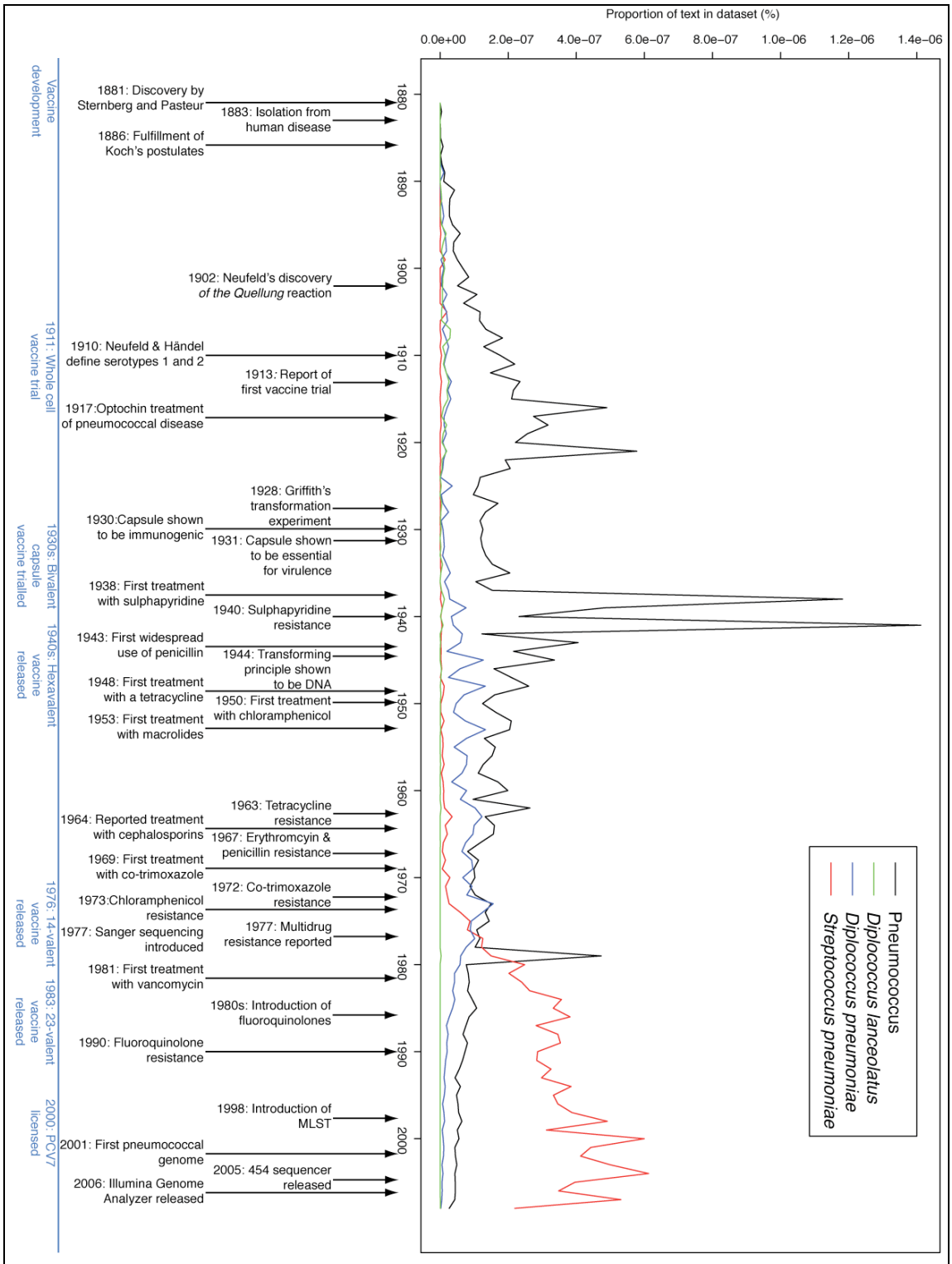

Figure 1.1 A timeline of pneumococcal research. The graph shows the proportion of English literature, as defined by Michel *et al* (Michel *et al*., 2010), in each year containing the listed descriptive terms for *S. pneumoniae*. Major developments in the study and treatment of the pneumococcus are listed chronologically; the introductions of anti-pneumococcal vaccines are detailed in blue on the far left hand side.

1.6 Anti-pneumococcal vaccines

Development of anti-pneumococcal vaccines was initially motivated by the lack of chemotherapeutic treatment options. However, interest in different formulations successively waned as new antibiotics were introduced. Recently, the advent of multidrug-resistant lineages has focussed efforts on improved vaccine designs that afford protection even to young children without a fully matured immune system.

1.6.1 Early whole cell vaccines

Sternberg himself appears to have been the first to immunise animals against the pneumococcus: he describes rabbits appearing to be protected against infection following inoculation with bacteria killed with alcohol or quinine (Sternberg, 1882). The first vaccine to be administered to humans was tested in South Africa, in response to the high morbidity and mortality from pneumococcal pneumonia among workers in the gold mining industry (Austrian, 1978). It consisted of dead pneumococci of one, or more, undefined serotypes; a trial involving 50,000 recipients found a reduced rate of pneumonia in the recipients during the four months after administration, but protection was lost over time (Maynard, 1913; Wright *et al*., 1914). Later trials, in the same population, of a formulation including five different, defined serotypes found a 20% decrease in the rate of pneumonia, but no work was done to evaluate whether the disease still occurring in the vaccinated population was due to serotypes included in the vaccine or not (Maynard, 1915). Refinement of these prophylactic measures continued until efforts were largely discontinued once sulphonamides were introduced (Austrian, 1978).

1.6.2 Polysaccharide vaccines

The demonstration that the capsule triggered the development of an adaptive immune response during disease in humans (Dochez and Avery, 1917; Dubos and Avery, 1931) lead to tests showing that an immune response was also possible when the polysaccharide was injected intradermally (Francis and Tillett, 1930). The first anti-pneumococcal capsule-based vaccine was a bivalent formulation, containing type 1 and 2 polysaccharides, trialled in the USA in the 1930s (Ekwurzel *et al*., 1938);

however, the study again failed to distinguish between vaccine and non-vaccine serotype disease in the evaluation (Austrian, 2000). It was not until the Second World War when a tetravalent vaccine, comprising capsule types 1, 2, 5 and 7, was more thoroughly tested in a US Air Force base; this reduced the incidence of vaccine-type pneumococcal pneumonia by 85%, and also protected against the nasopharyngeal acquisition of these serotypes (Macleod *et al*., 1945). As a consequence, two hexavalent vaccines were made commercially available in the late 1940s: an adult version containing serotypes 1, 2, 3, 5, 7 and 8, and one for children containing serotypes 1, 4, 6, 14, 18 and 19. However, the introduction of penicillin meant that these were withdrawn due to lack of interest in 1954 (Kemp, 1979).

Interest in vaccines was revived in the 1970s, when a 13 valent formulation was tested in 12,000 South African miners; the observed burden of vaccine-serotype pneumococcal pneumonia in the vaccinated group was reduced by almost 80% (Austrian *et al*., 1976). A 14 valent version was licensed for use in the USA in 1977, and subsequently expanded to encompass 23 serotypes (1, 2, 3, 4, 5, 6B, 7F, 8, 9N, 9V, 10A, 11A, 12F, 14, 15B, 17F, 18C, 19A, 19F, 20, 22F, 23F, and 33F) in 1983; this proved successful in causing a 60-70% decline in pneumococcal disease in immunocompetent adults (Shapiro and Clemens, 1984; Bolan *et al*., 1986; Sims *et al*., 1988; Shapiro *et al*., 1991). However, rates of paediatric disease were largely unaffected (Makela *et al*., 1981; Sloyer *et al*., 1981; Teele *et al*., 1981), as children under five were not capable of producing a strong immune response to most types of the polysaccharide capsule (Sell *et al*., 1981; Douglas *et al*., 1983; Lawrence *et al*., 1983). This is because, as a T cell-independent antigen, the immune response to such stimuli is not fully functional in infants (Stein, 1992).

1.6.3 Conjugate polysaccharide vaccines

The first pneumococcal capsule polysaccharide to be conjugated to a protein was that of serotype 3, attached to a horse serum globulin; this was used to immunise and protect rabbits against experimental infection with pneumococci of the same type (Avery and Goebel, 1931; Goebel and Avery, 1931). The first application of this technology to a vaccine in humans was the linking of the *H. influenzae* type b (Hib) capsule, which alone fails to trigger a strong immune response in infants, to either to

the inactivated diptheria toxin CRM_{197} or a meningococcal protein, to generate a T cell-dependent antigen that provoked an immune response in young children (Stein, 1992; Adams *et al*., 1993). The widespread use of this vaccine lead to a significant decrease in Hib carriage and disease in infants (Adams *et al*., 1993; Barbour *et al*., 1995), triggering an interest in developing an equivalent vaccine for *S. pneumoniae.*

The first pneumococcal conjugate polysaccharide vaccine was PCV7, a heptavalent formulation comprising seven capsule polysaccharides (4, 6B, 9V, 14, 18C, 19F, 23F) attached to CRM_{197} (Rennels *et al*., 1998). However, prior to it being licensed in the USA in 2000, a number of potential problems were identified (Lipsitch, 1999; Spratt and Greenwood, 2000). Despite constituting only a small proportion of the carried population (Barbour *et al*., 1995), Hib was responsible for the vast majority of disease caused by the species before the introduction of the vaccine, with infections caused by other serotypes largely opportunistic in their nature (Peltola, 2000). Furthermore, experiments in animal models clearly linked virulence to the capsule type (Moxon and Vaughn, 1981). By contrast, *S. pneumoniae* disease is caused by a wider range of serotypes in humans, and animal infections investigating the link between capsule type and genetic background have not conclusively shown that serotype controls virulence in the same way as for *H. influenzae* (Kelly *et al*., 1994; Hyams *et al*., 2010b). Additionally, the seven serotypes constituting PCV7 are commonly carried, hence their elimination would lead to the opening of a large ecological niche. Therefore, following the pneumococcal vaccine, there was predicted to be scope for serotype replacement (Lipsitch, 1997), whereby non-vaccine type strains increase in prevalence to replace those protected against by immunisation, and serotype switching (Spratt and Greenwood, 2000), involving successful genetic lineages changing serotype to evade the vaccine.

Following the introduction of the vaccine in the USA, decreases in total invasive pneumococcal disease (IPD; defined as isolation of pneumococci from a normally sterile site) of around 60-70% relative to pre-vaccination levels were reported in young children, for whom the vaccine was recommended (Lin *et al*., 2003; Whitney *et al*., 2003; Kaplan *et al*., 2004; Hsu *et al*., 2005). The proportion of acute otitis media caused by pneumococci also declined by 30-40% following PCV7 introduction

(Block *et al*., 2004; Casey and Pichichero, 2004), suggesting the immune response to the vaccine was strong enough to prevent mucosal disease. Furthermore, there was a significant decline in the carriage of vaccine-type pneumococci, although the overall level of pneumococcal carriage did not decrease due to replacement by non-vaccine serotypes (Moore *et al*., 2004; Huang *et al*., 2005; Park *et al*., 2008). However, this replacement of vaccine serotypes by less invasive non-vaccine serotypes has resulted in a herd immunity effect (Weinberger *et al*., 2011), with a decline in the level of pneumococcal disease in infants too young to receive the vaccine (Poehling *et al*., 2006) and adults (Whitney *et al*., 2003; Lexau *et al*., 2005), including those with HIV (Flannery *et al*., 2006), observed in the USA.

Also, as five of the vaccine serotypes were strongly associated with clinically-relevant antibiotic resistance (Dagan and Klugman, 2008), early surveillance data suggested PCV7 had been at least partially effective in reducing the problem of non-susceptible pneumococci. Following the introduction of the vaccine, studies in the USA found decreased proportions of IPD caused by strains resistant to penicillin (Kaplan *et al*., 2004; Talbot *et al*., 2004) and macrolides (Stephens *et al*., 2005). However, by 2004 the proportion of resistant strains among clinical isolates from children under two had largely rebounded to their previous levels (Kyaw *et al*., 2006). It remains unclear as to whether PCV7 caused a reduction in the proportion of acute otitis media due to penicillin-resistant strains (McEllistrem *et al*., 2003; Block *et al*., 2004; Casey and Pichichero, 2004). Surveys of carriage in the USA have found no change in the proportion of antibiotic resistant *S. pneumoniae* isolates in the carried population, with the exception of a consistent fall in co-trimoxazole resistance (Moore *et al*., 2004; Huang *et al*., 2005; Park *et al*., 2008). However, there is an argument that by decreasing the amount of antibiotics prescribed for pneumococcal disease, PCV7 will reduce the selection pressure for the evolution of resistance (Dagan and Klugman, 2008), although without a more general decline in prescriptions it is unclear how strong the selection pressure will be on the carried population.

In the USA, the main problem associated with the introduction of PCV7 was the rise of serotype 19A isolates, which have emerged as the major type causing paediatric disease and are increasingly associated with multidrug resistance (Pai *et al*., 2005;

Hicks *et al*., 2007; Messina *et al*., 2007; Pelton *et al*., 2007; Singleton *et al*., 2007). MLST analysis of the emergent 19A isolates revealed they were quite diverse: along with some instances of known 19A lineages expanding to fill the vaccine-generated niche, the multidrug-resistant strains generally represented 19A variants of extant PMEN1 lineages, such as PMEN1, PMEN3 and PMEN14 (Taiwan^{19F}-14) (Moore *et al*., 2008). The increase in serotype 19A disease has also been seen outside of the USA: rises have been reported in France (Mahjoub-Messai *et al*., 2009), the UK (Gladstone *et al*., 2011) and Spain, where the most common 19A lineages are variants of the PMEN1 and PMEN2 lineages (Munoz-Almagro *et al*., 2008; Ardanuy *et al*., 2009).

Increases in the level of serotype 7F disease have been even more widespread in Europe, occurring in Spain (Munoz-Almagro *et al*., 2008; Ardanuy *et al*., 2009), Portugal (Sa-Leao *et al*., 2009; Aguiar *et al*., 2010), Germany (Ruckinger *et al*., 2009a), France (Lepoutre *et al*., 2008) and the UK (Gladstone *et al*., 2011). Increases in the prevalence of serotype 1 have been manifest as the dramatic increases observed in pneumococcal empyema in some regions (Byington *et al*., 2006; Hendrickson *et al*., 2008; Munoz-Almagro *et al*., 2008). The high level of serotype replacement seen in countries such as France (Doit *et al*., 2010), the Netherlands (Rodenburg *et al*., 2010), Spain (Guevara *et al*., 2009), Australia (Hanna *et al*., 2008) and the UK (Gladstone *et al*., 2011) has resulted in a smaller decline in IPD affecting vaccinated individuals, often leading to a negligible herd immunity effect for the rest of the population. It seems likely that the composition of the resident pneumococcal population is a crucial factor in determining the success of the national PCV7 vaccination programmes.

1.7 The impact of second-generation sequencing technologies

Following the inception of DNA sequencing, throughput was originally increased through modifications of the original dideoxy terminator sequencing method. Although this approach was successful in producing many complete bacterial and eukaryotic genomes, the development of entirely new techniques for DNA sequencing have vastly increased the rate at which such data can be produced.

1.7.1 Dideoxy terminator sequencing

For many years, the most common method of DNA sequencing was the dideoxy terminator method used to sequence the first DNA genome, that of ϕX174 (Sanger *et al*., 1977); this followed a year after the genome of the first RNA bacteriophage, MS2 (Fiers *et al*., 1976). The method relies on using four separate reactions, each containing a different radioactively or fluorescently-labelled dideoxy terminator corresponding to one of the bases, which is randomly incorporated into a strand synthesised from the template and, at that point, curtails further extension of the strand. Parallel electrophoretic separation of the products of four reactions on the basis of their size allows, based on the pattern of bands, the sequence of bases in the template to be determined. Modern capillary-based sequencing approaches, using a highly refined version of the original method, produce reads with a mean length around 800 bp, with typically around one error per read (Metzker, 2005). However, there are intrinsic limitations: large amounts of template are required for each sequencing run, necessitating the generation of libraries of cloned DNA fragments for most large-scale projects. This introduces a bias, as some DNA inserts are lethal for *E. coli*, which is pronounced for Firmicutes. Furthermore, the space and coordination required for four electrophoretic separations per template limits the throughput of the technique. Hence, although the mean read length and error rate have yet to be bettered, the bulk of sequence production has now moved to the second generation sequencing technologies (SGSTs).

1.7.2 Second-generation sequencing technologies

The SGSTs share some common differences with dideoxy terminator sequencing. Firstly, their primary advantage, they all sequence large numbers of templates in parallel, allowing the generation of vast quantities of data. Secondly, they all use the emission of light as the signal that indicates the sequence of the template. This necessitates charge-coupled devices (CCDs) to detect light in a sensitive and precise manner, such that the photon emissions can be assigned to one of the many strands being sequenced concurrently. Thirdly, in order to generate a detectable signal, the sequencing reactions must act on a spatially clustered set of identical sequences,

rather than an individual strand; the amplification of the target DNA in each case occurs *in vitro*, avoiding the biases resulting from cloning the sequences into *E. coli*.

1.7.2.1 454 sequencing

The first SGST to become commercially available was the 454 system, which is based on an approach termed 'pyrosequencing' (Margulies *et al*., 2005). Rather than terminate the sequencing strand, extension is controlled through the management of deoxynucleotide triphosphate (dNTP) concentrations (Figure 1.2). During the sequencing cycle, each of the four dNTPs is added in turn, and when a base is incorporated, pyrophosphate is released and used to generate ATP from adenosine 5' phosphosulphate (APS) by a sulphurylase. The ATP is then used to generate light through the action of luciferase on luciferin. The magnitude of the light pulse from the strand throughout the sequencing cycle indicates the number of bases of the same type that have been incorporated in each step of the cycle; from this information, the sequence of the template can be deduced.

The template strands are distinguished from one another through being separated into different wells on a plate. This is achieved through first ligating adapters to fragments of the template DNA in solution, then annealing these to beads coated in oligonucleotides complementary to the adaptors, under conditions that favour no more than one strand attaching to each bead. An emulsion PCR is then used to coat each bead in multiple copies of the adhered DNA construct (Figure 1.2): an oil-aqueous mix is created, such that each bead is isolated in its own aqueous droplet, thereby preventing cross-contamination of sequences between beads, before the PCR is performed using generic primers that bind the adaptor sequences.

The 454 platform is currently capable of producing around 0.5 Gb per run, composed of reads with a mean length between 300-400 bp (Metzker, 2010). Although the rate of substitutions errors is comparatively low, the reads contain a high density of insertion or deletion errors concentrated in homopolymeric tracts. This is a result of the difficulty of correctly estimating the number of bases incorporated at once, during a single sequencing step, on the basis of the luminescence pulse magnitude. Hence

assemblies based on 454 alone are liable to contain large numbers of false frameshift mutations.

1.7.2.2 SOLiD sequencing

The SOLiD platform operates through a 'sequencing by ligation' approach (Valouev *et al*., 2008). The templates are generated through an emulsion PCR, as for 454, but instead of being deposited in wells the beads are covalently linked to an amino-coated glass surface. These are then exposed to sixteen different probes, corresponding to all possible dinucleotide combinations; each probe consists of two specific bases at the 5' end attached to six degenerate bases and one of four fluorophores (Figure 1.2). Excess probes are then washed away and those annealed to the template are attached using DNA ligase. Following laser excitation, the wavelength of fluorescence indicates the fluorophore that has been attached; three of the degenerate bases and the fluorophore are then cleaved off, allowing a further probe to hybridise to the base five nucleotides from the 5' end of the previous probe. Cycles of ligation give a sequence of 'colours' that relate to the template nucleotide sequence. The ligated strand is then removed, and a different primer, which anneals to the adapter one base offset from the first primer, is used to initiate sequencing. This process is repeated five times in total, each with a primer binding at a position offset by one base from the previous primer, allowing an unambiguous translation of the colour sequence into a nucleotide sequence.

SOLiD sequencing is currently capable of producing 30 Gb (single end) or 50 Gb (paired end) of sequence data per run, although the reads, at 50 bp, are very short (Metzker, 2010). The substitution error rate is low, due to the redundancy of translating the sequence of colours into bases, and the controlled stepwise addition of probes avoids the problem of insertion or deletion errors that plagues 454 sequencing. However, the need for multiple analyses of the same template strand to recover the exact sequence does mean each SOLiD sequencing run takes considerably longer than those of other SGSTs.

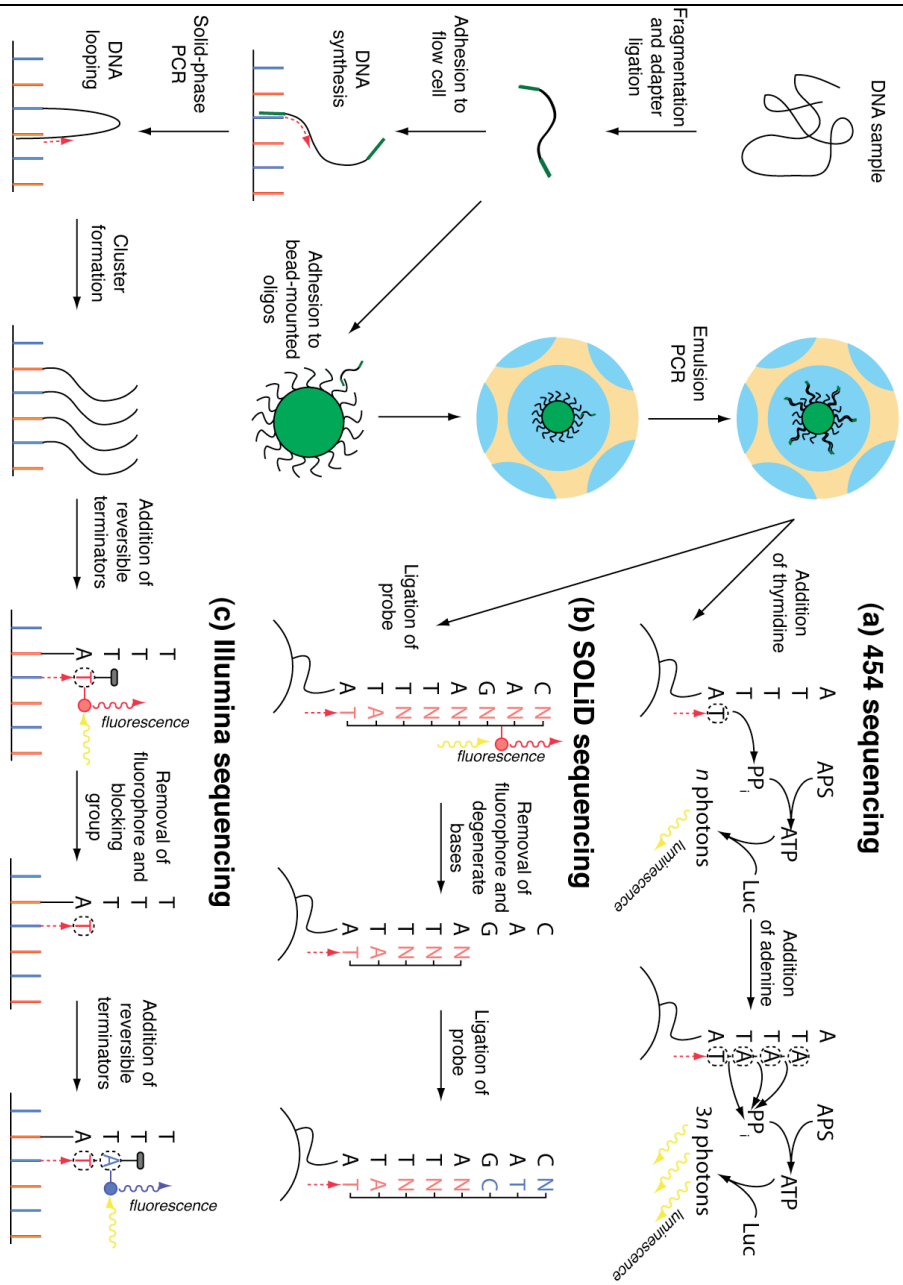

Figure 1.2 A summary of the SGST methods. All three require DNA constructs composed of adapters ligated to small, randomly fragmented pieces of the sample DNA. Multiple copies of these fragments are then produced through emulsion PCR (454 and SOLiD) or solid-phase PCR (Illumina). (A) In 454, the beads are deposited into individual wells and then separately exposed to each of the four nucleotides in a cycle. A series of enzyme-catalysed reactions lead to chemiluminescence of a magnitude proportional to the number of nucleotides of the same type that are added at once. (B) In SOLiD, the beads are attached to a surface, and sixteen probes, each attached to one of four different fluorophores, with the different dinucleotide combinations followed by four degenerate bases are ligated to the sequencing primer. Once fluorescence has been detected, the 3' three nucleotide and the fluorophore are removed, and the next probe ligated. Due to the three-fifths of the sequence that is not sequenced on each pass, and the degeneracy of the probe-fluorophore relationship, five different sequencing runs are performed, each primed by an oligonucleotide that binds one base offset from the previous round. (C) Illumina's solid-phase PCR generates clusters of identical sequences attached to a flow cell surface. A modified nucleotide, with a fluorophore and a 3' blocking group, is then incorporated into the sequencing strand, and fluorescence used to ascertain which of the four possibilities has been added. The fluorophore and blocking group are then removed and the next base incorporated.

1.7.2.3 Illumina sequencing

The Illumina, previously Solexa, sequencing approach relies on 'cyclic reversible termination' (Bentley *et al*., 2008). Following the addition of adapters to the template DNA, these constructs are adhered to a flow cell, a surface coated in oligonucleotides complementary to the 5' and 3' adapters. 'Clusters' of identical copies of the same sequence are then generated by a solid-phase PCR: during each round of amplification the DNA constructs loop over to bind nearby oligonucleotides to prime synthesis of further copies of the template DNA. This leads to spatially separated clusters of identical sequences (Figure 1.2).

Each sequencing cycle consists of the simultaneous addition of four different reversible terminator molecules, each derived from one of the four bases and carrying a distinctive fluorophore (Figure 1.2). Following the incorporation of the terminator into the sequencing strand, the wavelength of light emitted from each cluster indicates which base is present in the template DNA. The dye and 3'-*O*-azidomethyl blocking group are then removing, allowing the extension of the sequencing strand by a single base in the following cycle.

Illumina sequencing has become the most widely-used of the SGSTs, currently capable of producing 18 Gb (single end) or 35 Gb (paired end) of data from the Genome Analyzer II platform (Metzker, 2010). Reads are intermediate in length between 454 and SOLiD; depending on the number of cycles, they can be over 100 bp. The substitution error rate is typically about 1% or lower, and as with SOLiD there are few false indels.

1.7.3 Bacterial population genomics

Early applications of SGSTs involved sequencing multiple isolates in order to assess the level of genetic diversity present within species. This concept was originally quantified as the species 'pan-genome', with the chromosome of each isolate divided into a 'core', shared with all other members of the species, and a 'dispensible' or 'accessory' component that varied between strains of the same species (Tettelin *et al*., 2005). Through permuting a set of representative sequences, the mean increase in the size of the pangenome, and decrease in the size of the core genome, on the addition of

the *n*th sequence, when compared to a defined set of (*n*-1) sequences, could be calculated for values of *n* between two and the number of available genomes. The first use of an SGST to study pneumococcal diversity used a different algorithm, the 'finite supragenome model', applied to a set of genomes sequenced using 454 and capillary technologies (Hiller *et al*., 2007). Based on their frequency among genomes, each gene is categorised into a one of a discrete number of classes; a maximum likelihood estimate for the total number of genes found in the species can be derived from the relative sizes of these classes (Hogg *et al*., 2007). This indicated the *S. pneumoniae* core genome was ~1,400 genes, around 75% of a typical genome, with a species 'supragenome' approximately twice this size. However, these results should be treated with caution, as the model assumes the presence of each gene is independent of all others (Hogg *et al*., 2007), and yet a high proportion of the accessory genome was composed of prophage-related genes, inherited together as large coherent units (Hiller *et al*., 2007). Furthermore, a recent application of the original pan-genome model to a set of pneumococcal genomes predicted the pangenome would not be finite, but rather 'open', implying that sequencing of further genomes would always continue to uncover novel genes (Donati *et al*., 2010).

More recent applications of SGSTs have used whole genome sequences for bacterial epidemiology, allowing strains to be tracked at a greatly increased level of resolution. The first species to which this approach was applied was *Salmonella enterica* serovar Typhi, all isolates of which are sufficiently closely related that other typing techniques struggle to differentiate them (Kidgell *et al*., 2002). Sequencing of 19 *Salmonella* Typhi isolates from a global collection using 454 and Illumina technologies identified 1,964 SNP sites in the core genome, excluding plasmids and prophage sequences that showed relatively high levels of variation (Holt *et al*., 2008). These core SNPs defined a phylogeny with little homoplasy, with the notable exception of substitutions in *gyrA* causing fluoroquinolone resistance, taken as evidence for the absence of recombination in the population.

Similar approaches were taken to study a sample of strains from the multidrug-resistant *Staph. aureus* lineage ST239 (Harris *et al*., 2010). The efficiency of sequencing was greatly improved through multiplexing samples on the Illumina

platform; each strain's shotgun library was assigned a specific tag, and then twelve strains sequenced in the same lane, so as to minimise costs per strain. In total, 63 isolates were analysed, allowing 4,310 SNP sites to be identified in a core genome that excluded mobile elements, which again showed high levels of variation. Phylogenetic analysis allowed the global dissemination of the lineage to be followed; again, little homoplasy was evident in the tree, other than mutations that lead to antibiotic resistance. Another study used non-multiplexed Illumina sequencing to identify SNPs in 87 strains of *S. pyogenes* M3 isolated over a 15-year period in Ontario (Beres *et al*., 2010). This identified 801 SNP sites and 193 indel sites, which were then specifically assayed in a collection of 344 strains. Once more, having eliminated the mobile elements from the analysis, there was no sign of recombination within the genome. By contrast, a small number of large putative recombination events could be identified among six *Clostridium difficile* genomes, assembled using a combination of 454 and capillary data, and even within the clinically important ribotype 027 lineage, 25 representatives of which were sequenced as multiplexed libraries on the Illumina platform (He *et al*., 2010). Species-wide, the estimate of *r/m* was 0.63-1.13, compared to a range of 0-0.5 deduced for the species from MLST data (Vos and Didelot, 2009).

1.7.4 Bacterial RNA-seq

The first whole transcriptome studies of pneumococci were performed using microarrays, based on the early complete genomes of *S. pneumoniae* TIGR4 and R6. Second generation technologies allow for a different approach to be taken: the total RNA can be extracted from a cell, reverse transcribed into cDNA and sequenced using a SGST, a technique known as RNA-seq. The sequence data can then be aligned to the appropriate genome sequence in order to obtain a quantitative view of expression at a single base resolution. This method was first applied to the yeast species *Schizosaccharomyces pombe* (Wilhelm *et al*., 2008) and *Saccharomyces cerevisiae* (Nagalakshmi *et al*., 2008). These works demonstrated the key advantages of RNA-seq over microarrays: while most microarrays were designed to assay the expression of particular genes, hence were biased by genome annotations, RNA-seq produces an unbiased view of the transcriptome, hence allowing the discovery of novel genetic features. The resolution is also increased, as the mapping of reads to a

reference sequence *in silico* is more precise and stringent than the hybridisation between oligonucleotide probes and RNA or cDNA (Kane *et al*., 2000). Finally, RNA-seq is not affected by saturation in the same way microarrays, which quantify expression through dye fluorescence, permitting expression to be studied across a greater dynamic range (Cloonan and Grimmond, 2008).

For bacteria, especially genetically variable species, another key advantage offered by RNA-seq was the ability to sequence a transciptome specific to a strain without having to construct an array specific for that genome. The earliest applications of RNA-seq to bacteria were the analyses of the transcriptomes of *Bacillus anthracis* (Passalacqua *et al*., 2009), *Burkholderia cenocepacia* (Yoder-Himes *et al*., 2009) and *Listeria monocytogenes* (Oliver *et al*., 2009); however, these studies suffered from the common limitation that, by constructing conventional libraries from double stranded cDNA, the information on the direction of transcription is lost. Various methods were subsequently used to sequence transcriptomes in a strand-specific manner and applied to *Mycoplasma pneumoniae* (Guell *et al*., 2009), *Helicobacter pylori* (Sharma *et al*., 2010) and *Salmonella* Typhi (Perkins *et al*., 2009). These works have greatly enhanced our understanding of bacterial gene expression, finding that antisense transcription is common throughout the genome, identifying a variety of novel coding and non-coding RNAs and concluding that many genes are transcribed from multiple promoters, and hence are simultaneously part of multiple operons and sub-operons.

1.8 Summary

Pneumococcal research has been largely focussed on the capsule since the discovery of the bacterium, with serology forming the basis of studies of virulence, vaccine design and epidemiological typing. It was only the evolution of antibiotic resistance in *S. pneumoniae*, particularly the advent of multidrug resistance in the 1970s, which motivated an interest in superior, multilocus-based typing schemes and investigations of the resistance determinants, including the mobile genetic elements that carried them. The sequencing of the first pneumococcal genomes, and the microarrays they made possible, provided the opportunity to investigate the rest of the chromosome with similar intensity. However, in a species as genetically heterogenous as *S. pneumoniae*, assaying the genome content of a handful of strains has not proved

sufficient to study pneumococcal diversity in its totality, especially given the lack of antibiotic resistance among sequenced isolates. This can only by achieved through *de novo* sequencing of large numbers of strains, an opportunity now afforded to the scientific community by the advent of the SGSTs. This dissertation describes the applications of these technologies to understanding the evolution of this pathogen.