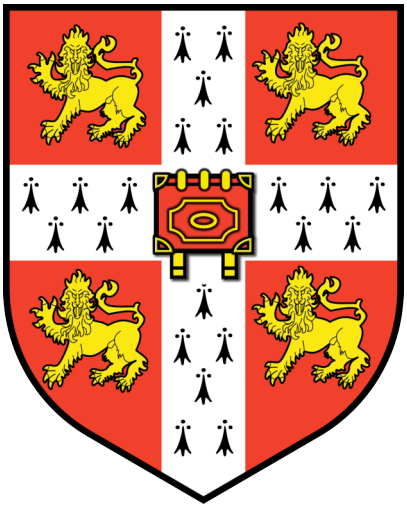
Functional and evolutionary analyses of pneumococcal genome variation

Nicholas Jason Croucher

Downing College

University of Cambridge

2011



This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

I hereby declare that this dissertation is my own work and contains nothing that is the outcome of work done in collaboration with others, except where specifically indicated here or elsewhere in the text.

All sequence data were produced by the sequencing and research and development teams at the Wellcome Trust Sanger Institute. The annotation of complete genome sequences was performed in conjunction with Dr. Stephen Bentley (Wellcome Trust Sanger Institute, Cambridge). Phylogenetic analyses of bacterial genome sequences were performed in collaboration with Dr. Simon Harris (Wellcome Trust Sanger Institute, Cambridge). All microarray experiments and statistical analyses were performed by the BμG@S group (St. George's Hospital, London). All statistical analyses of Omnilog data were performed by Lars Barquist (Wellcome Trust Sanger Institute, Cambridge).

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

Nicholas Jason Croucher
August 2011

Acknowledgements

I thank Julian Parkhill for allowing me the opportunity to undertake this work and for all his help and advice, in particular his ability to patiently correct me when I have been stubbornly wrong. I also owe a huge debt of gratitude to Stephen Bentley, upon whom the burden of my supervision has fallen most heavily; despite this, he has remained unflinchingly positive, understanding and encouraging, and he has proved to be a fantastic, friendly and wise supervisor. Similarly, Nick Thomson has been a perennially enthusiastic and supportive manager; it has been a pleasure to work for all three. I also thank Duncan Maskell for his perceptiveness and advice as my external supervisor, and the other members of my advisory committee: Gordon Dougan, who has also invested much effort in helping with my laboratory work, and Alex Bateman, who has organised the graduate students at the Sanger Institute so effectively.

Much of my informatics work could not have been completed without the help of Simon Harris and Matthew Holden, who have been two of the most friendly, conscientious and helpful people I could possibly hoped to have shared an office with. Thomas Otto also deserves much credit for shouting and singing me into being a better systems user. I am also grateful to Theresa Feltwell and Sally Whitehead for their tolerance of me, and the disruption I inevitably cause, in running their respective laboratories. I thank Maria Fookes and Del Pickard for their vast microbiological expertise and general helpfulness, as well as their unendingly friendly and amusing ways, along with Trevor Lawley, who also taught me a very practical approach to laboratory etiquette. I am also thankful to all of teams 15 and 81 for their patience and understanding, with a special mention of congratulations to the ever entertaining and hospitable Alan ‘Aunt’ Walker. Of course, none of this work could have been completed without the help of the informatics, systems, sequencing and library making teams of the Sanger Institute, which has been a brilliant place to work.

Externally, I have enjoyed many productive discussions with Bill Hanage and Christophe Fraser of Imperial College, London. Gavin Patterson, of the Vetinary School, Cambridge, was very kind in initially providing me with strains and advice on laboratory techniques. Tim Mitchell, of the University of Glasgow, has been a

friendly and invaluable collaborator. In addition, I am greatly appreciative of the kindness of all our other collaborators who have provided us with relevant strains, which have been essential to this project.

I also thank my family, in particular my mum, my dad and my grandparents who have been so supportive and encouraging throughout these seven years at Cambridge, and beyond. The friends I have made during this time have also been central in making the bad times survivable and the good times so enjoyable. My final thanks are to Alina, who has had to demonstrate more patience and understanding than could reasonably be expected of anyone, and has made me so happy.

p.s. Mum, to save you flicking through to the end: no, I still haven't cured anything.

Abstract

Functional and evolutionary analyses of pneumococcal genome variation

Nicholas Jason Croucher

Streptococcus pneumoniae (the pneumococcus) is a human nasopharyngeal commensal and respiratory pathogen responsible for a high burden of morbidity and mortality worldwide. The bacterium's primary virulence factor appears to be its polysaccharide capsule, of which there are more than 90 different serologically-distinguishable types (serotypes). Although this categorisation was originally used for tracing pneumococcal epidemiology, the bacterium is naturally transformable, and hence is able to switch serotypes through horizontal exchange of capsule biosynthesis (*cps*) gene clusters. Therefore, following the emergence of multidrug-resistant lineages in the late 1970s, superior, multilocus-based typing schemes were devised for following pneumococcal evolution. Increasing antibiotic resistance also motivated the development of a heptavalent conjugate polysaccharide vaccine, which targeted seven *S. pneumoniae* serotypes, leading to a decrease in pneumococcal disease. However, this impact has been ameliorated by an increase in disease resulting from replacement by non-vaccine serotypes and switching of *cps* loci by strains previously expressing vaccine serotypes. This thesis describes the application of second-generation sequencing technologies to investigating the mechanisms by which the pneumococcus evolves, especially in response to such clinical interventions.

The first part concerns the Pneumococcal Molecular Epidemiology Network clone 1 (PMEN1) lineage, one of the first multidrug-resistant pneumococcal genotypes to become a worldwide problem. Complete sequencing of the *S. pneumoniae* ATCC 700669 type strain, combined with draft sequencing of a global collection of 240 isolates, quantified the impact of recombination across the chromosome, as well as revealing the diversity of conjugative elements and prophage in the population. The acquisition of antibiotic resistances and the evasion of the conjugate polysaccharide vaccine were both evident in among the strains. *In vitro* transformation experiments, in the same genetic background, were then used to perform a more detailed

investigation of the types of homologous recombination events seen in the global population.

The second part of this dissertation describes the use of RNA sequencing to investigate the functional consequences of genomic variation. A novel method was developed and validated, and, when applied to *S. pneumoniae* ATCC 700669, revealed a family of expressed putative coding sequences that were formed by extended forms of the BOX interspersed repeat. This technique was also applied to two closely related strains of the PMEN31 lineage, both isolated from a single case of disease. This allowed the functional consequences of a small number of distinguishing polymorphisms on the global transcriptome to be ascertained, providing an insight into the level of pneumococcal evolution that can occur within an individual. Sequencing further members of this lineage showed that, although highly successful, this lineage has a much more static genotype than that of PMEN1.

The different mechanisms of pneumococcal genome variation are associated with evolution over different timescales, and in response to different selection pressures, but clearly interact in a number of ways. Hence the use of whole genome sequencing, surveying all the variation throughout the chromosome, will be crucial for greater understanding, and therefore improved control, of this important pathogen.

Contents

List of Figures

List of Tables

Abbreviations

aa ..... Amino acid
ABC ..... ATP-binding cassette
ATCC ..... American type culture collection
ATP ..... Adenosine triphosphate
BAC ..... Bacterial artificial chromosome
BHI ..... Brain heart infusion
BLAST ..... Basic local alignment search tool
bp ..... Base pair
CBP ..... Choline-binding protein
CC ..... Clonal complex
cDNA ..... Complementary DNA
CDS ..... Coding sequences
CGH ..... Comparative genome hybridisation
CSF ..... Cerebrospinal fluid
CSP ..... Competence stimulating peptide
DNA ..... Deoxyribonucleic acid
dNTP ..... Deoxynucleoside triphosphate
ds ..... Double stranded
DUS ..... DNA uptake sequence
EDTA ..... Ethylenediaminetetraacetic acid
EMBL ..... European Molecular Biology Laboratories
FR ..... Flanking region
GC ..... Guanine and cytosine
GMP ..... Guanosine monophosphate
HIV ..... Human immunodeficiency virus
HMM ..... Hidden Markov model
ICE ..... Integrative and conjugative element
IPD ..... Invasive pneumococcal disease
IS ..... Insertion sequence
LB ..... Luria broth
MCS ..... Multiple cloning site

MEPS ..... Minimum efficiently processed segment
MGE ..... Mobile genetic element
MITE ..... Miniature inverted repeat transposable element
MLEE ..... Multilocus enzyme electrophoresis
MLST ..... Multilocus sequence typing
MMR ..... Mismatch repair
MR ..... Mosaic recombination
mRNA ..... Messenger RNA
nt ..... Nucleotide
NTP ..... Nucleoside triphosphate
ORF ..... Open reading frame
PBP ..... Penicillin binding protein
PCR ..... Polymerase chain reaction
PCV ..... Polysaccharide conjugate vaccine
PEP ..... Phosphoenolpyruvate
PFGE ..... Pulsed field gel electrophoresis
PMEN ..... Pneumococcal molecular epidemiology network
PPI-1 ..... Pneumococcal pathogenicity island 1
PTS ..... Phosphotransferase system
PUS ..... *patAB* upregulatory SNP
rDNA ..... Ribosomal DNA
RNA ..... Ribonucleic acid
RPKM ..... Reads per kilobase per million mapped reads
rRNA ..... Ribosomal RNA
RSS ..... Recombined sequence segment
RT-PCR ..... Reverse transcription PCR
RUP ..... Repeat unit of pneumococcus
SGST ..... Second generation sequencing technology
SNP ..... Single nucleotide polymorphism
ss ..... Single stranded
ST ..... Sequence type
TCA ..... Tricarboxylic acid
TIR ..... Terminal inverted repeats
tRNA ..... Transfer RNA

TSD ..... Target sequence duplication
USS ..... Uptake signal sequence
UTR ..... Untranslated region
UV ..... Ultraviolet
VNTR ..... Variable number tandem repeat

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).

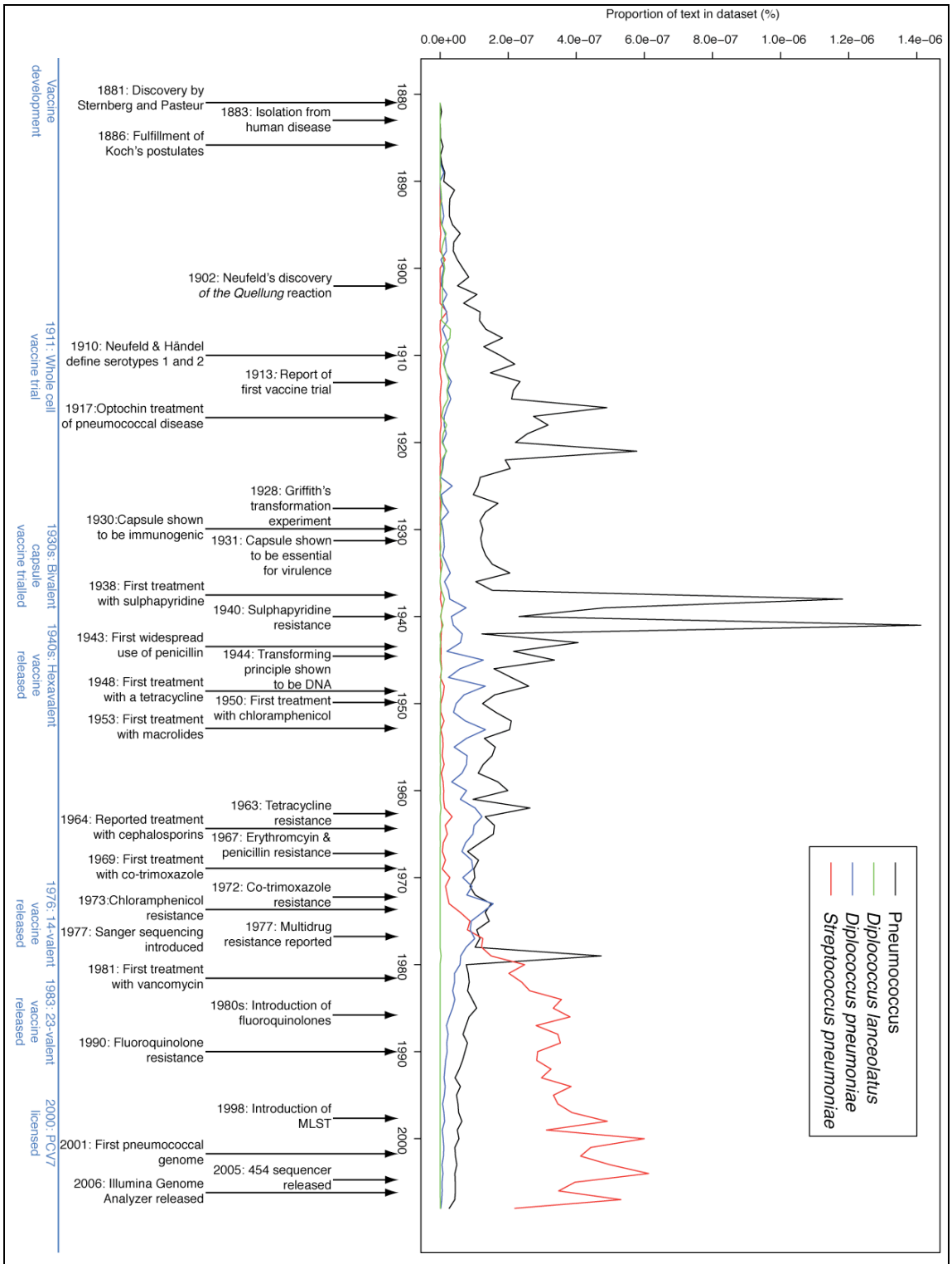


Figure 1.1 A timeline of pneumococcal research. The graph shows the proportion of English literature, as defined by Michel *et al* (Michel *et al*., 2010), in each year containing the listed descriptive terms for *S. pneumoniae*. Major developments in the study and treatment of the pneumococcus are listed 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.

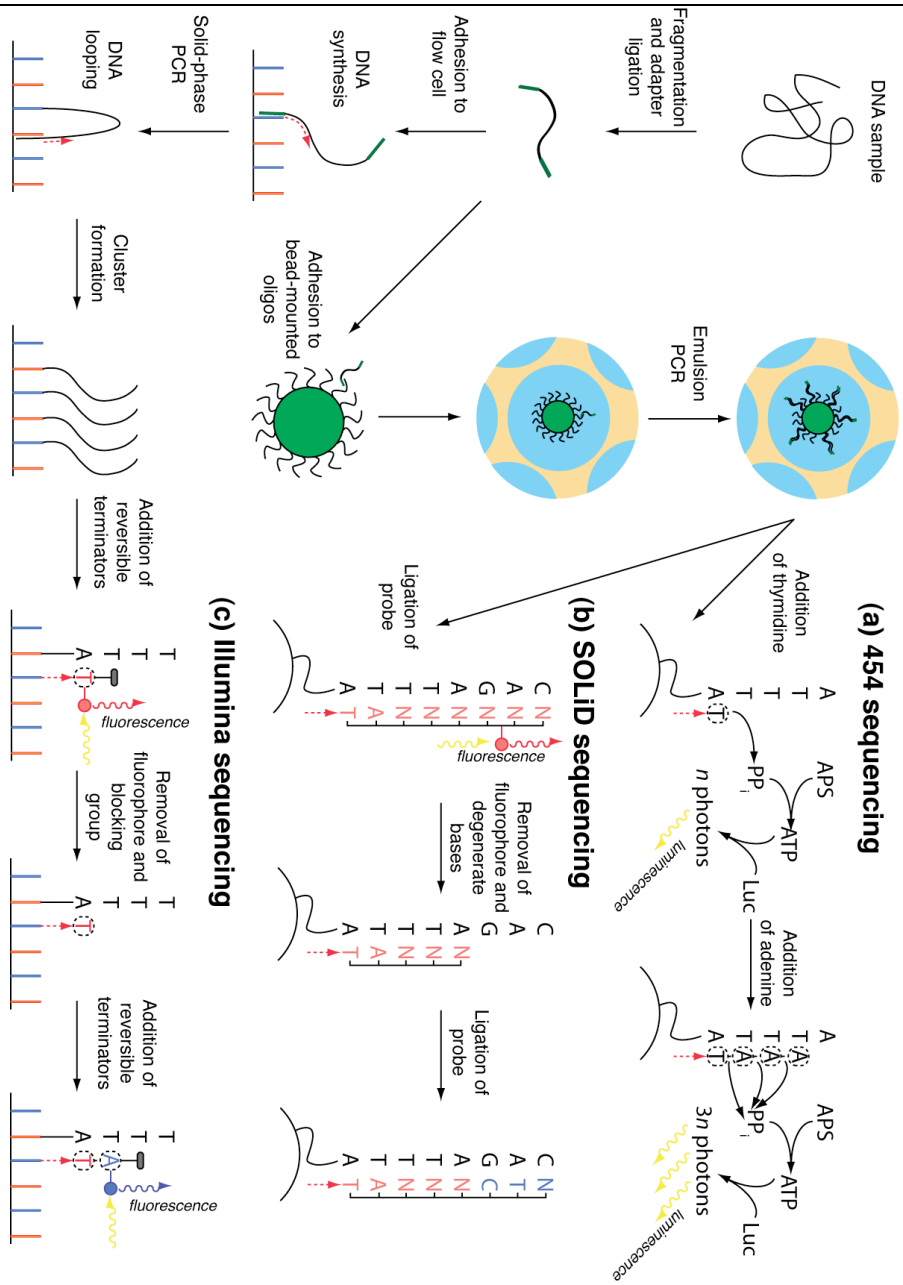


Figure 1.2 A summary of the SGST methods. All three require DNA constructs composed of adapters ligated to small, randomly fragmented pieces of the sample DNA. Multiple copies of these fragments are then produced through emulsion PCR (454 and SOLiD) or solid-phase PCR (Illumina). (A) In 454, the beads 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.

2 Materials and Methods

2.1 Culturing and transforming *S. pneumoniae*

2.1.1 Culturing of strains

Unless otherwise specified, *S. pneumoniae* was cultured on 5% horse blood agar plates (Oxoid), or grown in Brain-Heart infusion (BHI; Oxoid), statically at 37 °C. *S. pneumoniae* ATCC 700669 and TIGR4, and derivatives thereof, were grown under aerobic conditions, whereas *S. pneumoniae* 99-4038 and 99-4039 required a microaerophilic atmosphere, generated by CampyGen Compact sachets (Oxoid), to grow on blood agar plates, but could be grown aerobically in BHI.

2.1.2 Sampling of PMEN1 and ST180 isolates

For the analyses presented in Chapter 4, all available isolates with a sequence type of 81 (or a single locus variant thereof), or a pulsed field gel electrophoresis profile matching that of known PMEN1 strains, were sampled (Appendix II: PMEN1 strains). Where information on serotype and drug resistance profile was available, strains resistant to both penicillin and tetracycline were screened using PCRs, as described below, targeted to detect the presence of the CDS SPN23F00710, which is not commonly found in the pneumococcal chromosome, and the lantibiotic biosynthesis operon carried on ICE*Sp*23FST81, which has not been identified in any other sequenced pneumococcal lineage. No isolate positive for the SPN23F00710 locus was negative for the lantibiotic biosynthesis locus, and only a single isolate selected on the basis of sequence type information lacked the SPN23F00710 gene, so there is no reason to expect that this strategy significantly biased the sample. For the analyses presented in Chapter 8, the samples received from T Mitchell, selected as diverse representatives from a global collection studied using CGH (Inverarity, 2009), were sequenced using 454 and capillary technologies; all samples received from B Henriques-Normark were sequenced using Illumina technology (Appendix IV: Serotype 3 strains).

2.1.3 Transformation experiments

All strains subject to transformation had the same allele of the *comC* gene, hence were all expected to respond to CSP-2. Ten millilitres of Brain Heart Infusion (Oxoid) was inoculated with 150 μL of an overnight culture of the recipient strain. When the culture reached an OD_{600} of between 0.20-0.25, 1 mL was added to the appropriate amount of donor DNA (20 ng unless specified) in 5 μL water, 10 ng CSP-2 in 2 μL water (Sigma) and 5 μL 500 mM calcium chloride. Another 1 mL was added to the same quantity of donor DNA and calcium chloride in the absence of any CSP as a negative control. These reactions were incubated at 37 °C for 2 h. Samples of these cultures were then serially diluted in phosphate-buffered saline solution and total cell population determined by counting colonies in three 20 μL volumes spotted onto 5% blood agar plates from the appropriate dilution. The number of transformants from each reaction was determined by spreading three 50 μL volumes each onto a 5% blood agar plates supplemented with the appropriate antibiotic selection; unless specified, this was 200 μg mL^{-1} kanamycin (Gibco).

2.1.4 Omnilog experiments

Frozen stocks of *S. pneumoniae* 99-4038 and 99-4039 were passaged twice on blood agar plates overnight in order to prevent contamination of assays with glycerol. Colonies were then scraped off plates using sterile cotton swabs and dispensed into IF-0a solution (Biolog) at room temperature to a cell density corresponding to 81% transmittance. For each Omnilog phenotype microarray plate used (PM9-20) (Bochner, 2009), 120 μL of this cell suspension was added to 10 mL IF-10b solution (Biolog). This was then supplemented with 7.5 mM D-ribose (Sigma), 2 mM magnesium chloride, 1 mM calcium chloride, 2 mM sodium pyrophosphate (Sigma), 25 μM L-arginine (Sigma), 25 μM L-methionine (Sigma), 25 μM hypoxanthine (Sigma), 10 μM lipoamide (Sigma), 5 μM nicotine adenine dinucleotide (Sigma), 0.25 μM riboflavin (Sigma), 0.005% by mass yeast extract (Fluka) and 0.005% by mass Tween 80 (Sigma). The solution was then made up to a volume of 12 mL with distilled water, and 100 μL dispensed into each well on the assay plate. Plates were then allowed to equilibrate in an anaerobic atmosphere (80% N_2, 10% CO_2, 10% H_2) for 5 min prior to being sealed in airtight bags and loaded into the Omnilog machine.

Plates were scanned every 10 min for 48 h while incubated at 37 °C. Two paired replicates were performed for the two strains.

2.2 Extraction and analysis of nucleic acids

2.2.1 Genomic DNA extractions

Isolates were grown in 10 mL BHI (Oxoid) and pelleted through centrifugation (2,594 *g*, 10 min). Pellets were washed in 1 mL 50% glycerol and resuspended in 250 μL Tris-EDTA buffer and 50 μL 30 g L^{-1} lysozyme (Roche) in Tris-EDTA buffer. This mixture was vortexed at room temperature for 15 mins and 400 μL 0.1 M EDTA (Gibco) and 250 μL 10% sarkosyl (BDH) were added. Samples were incubated at 4 °C for 2 h, prior to the addition of 50 μL proteinase K (Roche), 30 μl RNase A (Roche) and 3 mL Tris-EDTA buffer. Samples were incubated at 50 °C overnight. Samples were washed with 5 mL of a 25:24:1 mixture of phenol, chloroform and indole-3-acetic acid (IAA; Fluka) and centrifuged (2,594 *g*, 10 min). The aqueous phase was removed, washed with 5 mL chloroform (Sigma) and centrifuged (2,594 *g*, 10 min). DNA was precipitated from the aqueous phase in 7.5 mL isopropanol, washed in 5 mL 70% ethanol and resuspended in 250 μL Tris-EDTA buffer.

2.2.2 RNA sample extractions

For *S. pneumoniae* ATCC 700669, samples were harvested from 10 mL cultures at an OD_{600} of 0.8 through mixing with RNAProtect (Qiagen) in a 1:2 ratio then pelleted through centrifugation (2,594 *g*, 10 min). Cells were resuspended in 1 mg mL^{-1} lysozyme (Roche) in 200 μL Tris-EDTA buffer and lysed at 37 °C for 10 min. The sample volume was then made up to 800 μL with Tris-EDTA and split six equal volumes, each of which was independently processed using the SV Total RNA Extraction System (Promega). The quality of the RNA was then assessed using an Agilent 2100 Bioanalyzer RNA Nano chip (Agilent); any samples with an RNA integrity number below nine were discarded. RNA was then precipitated through mixture with 300% by volume ethanol and 10% by volume 3 M sodium acetate followed by storage at -80 °C overnight. RNA was then resuspended at a

concentration of 0.83 mg mL^{-1} and the 16S and 23S rRNA transcripts depleted through complementary oligonucleotide hybridization (MicrobExpress, Ambion) according to manufacturer's instructions. RNA was then precipitated, then resuspended at a concentration of 0.625 mg mL^{-1} in water and treated with DNase I (Roche) at room temperature for 15 min. Reactions were stopped through washing with an equal volume of phenol, chloroform and IAA mixed in proportions 25:24:1 (Fluka) followed by phase separation through centrifugation at 16,157 *g* for 10 min. A PCR using primers smL and smR, targeting the *rpsL* gene, was performed as described below using the aqueous RNA solution as the template; the DNase I treatment was repeated until no amplification product was detectable. The replicate RNA samples were then pooled and split into two halves, each resuspended in 13.73 μL water and mixed with 1.67 μL of 3 μg μL^{-1} random hexamer oligonucleotides (Sigma), then incubated at 70 °C for 10 min, followed by incubation on ice for 10 min. A reverse transcription reaction was then performed using SuperScript III (Invitrogen), according to manufacturer's instructions, at 42°C for 2 h. Samples were then washed on a G50 spin column (GE Healthcare). For one of the two samples, the second cDNA strand was synthesised through incubating this first strand cDNA with DNA polymerase I (Invitrogen) and RNase H (Invitrogen) in second strand buffer, according to manufacturer's instructions, at 16°C for 2.5 h.

For analysis of *S. pneumoniae* 99-4038 and 99-4039, samples were harvested from 10 mL cultures at an OD_{600} of 0.6 through centrifugation (2,594 *g*, 10 min), then lysed by treatment with 30 mg mL^{-1} lysosyme (Roche) at room temperature for 15 min. RNA was extracted as described above, but no depletion of rRNA was performed. For analysis of *S. pneumoniae* TIGR4 and $TIGR4^{PUS}$, RNA was extracted as described for *S. pneumoniae* 99-4038 and 99-4039 then, following evaluation on the Bioanalyzer, sent to the BμG@S group (St. George's Hospital, London) for microarray analysis.

2.2.3 PCR and RT-PCR

All PCRs were conducted using 1 μL 10 ng μl^{-1} template DNA, when using genomic or plasmid DNA as a template, or 1 μL of the relevant undiluted sample when using RNA or cDNA as a template. Each primer (Appendix I: Primer sequences) was then added as 1 μl of a 10 μM solution (Sigma or IDT), and the reaction made up to a

volume of 50 μl with PCR Platinum Supermix (Invitrogen). The thermocycle in each case used a denaturing temperature of 95 °C (30 s), a hybridization temperature falling from 60 °C to 55 °C over 5 cycles, then remaining at 55 °C for a further 25 cycles (30 s), and an extension temperature of 72 °C (one minute per 1 kb of product length).

2.3 Construction of mutant strains

2.3.1 Construction of *S. pneumoniae* TIGR4^{PUS}

The region upstream of *patAB* in *S. pneumoniae* 99-4038 and 99-4039 was amplified through PCR using primers IntGL and IntGR (Appendix I: Primer sequences). The ~500 bp PCR products were each purified through agarose gel electrophoresis using a QIAquick Gel Extraction Kit (Qiagen) and then ligated into pGEM-T Easy (Promega) using T4 ligase (Promega) in a 10 μL reaction volume, according to manufacturer's instructions. A 1 μL sample of this reaction was then used to transform electrocompetent *E. coli* TOP10 cells (Invitrogen) through electroporation with a 2.5 kV pulse. These cells were then grown in 250 μL SOC medium (Invitrogen), shaken at 37 °C for 2 h. A 50 μL sample of this culture was then spread on Luria broth (LB) agar plates supplemented with 100 μg mL^{-1} ampicillin (Sigma), 300 μg mL^{-1} S-Gal (Sigma) and 30 μg mL^{-1} isopropyl β-thiogalactoside (Sigma). White colonies were then picked, grown in LB supplemented with 100 μg mL^{-1} ampicillin and stored. The sequences of the plasmid inserts were then amplified by PCR and checked through capillary sequencing as described below. Both plasmids were then extracted from their host *E. coli* using the QIAprep Spin Miniprep Kit (Qiagen) and diluted to 25 μg mL^{-1}. These two stocks was then used to transform three *S. pneumoniae* TIGR4 cultures in parallel as described above; after 2 h growth, a 50 μL sample of each transformation reaction was used to inoculate either BHI or BHI supplemented with 2 μg mL^{-1} ciprofloxacin. Colonies were then isolated from the *S. pneumoniae* TIGR4 culture transformed with the region upstream of *patAB* from *S. pneumoniae* 99-4039 after 20 h growth in the presence of ciprofloxacin. PCR amplification and sequencing of the region upstream of *patAB* in these strains revealed they all shared the PUS; one of these was stored and designated *S. pneumoniae* TIGR4^{PUS}.

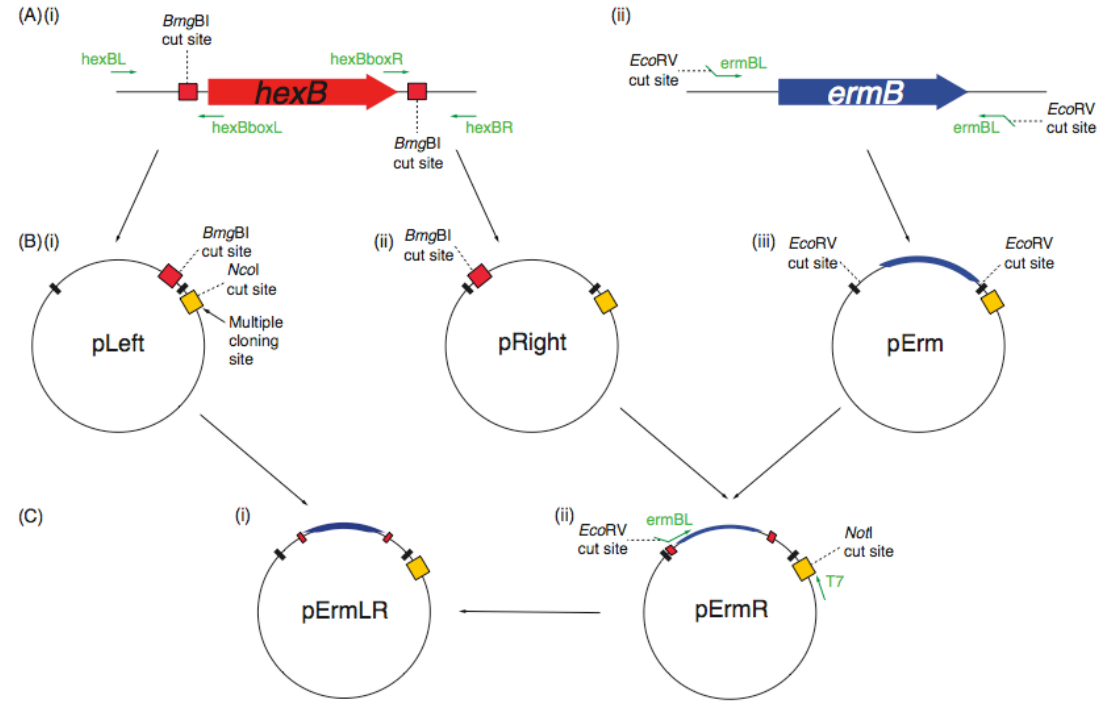


Figure 2.1 Construction of *S. pneumoniae* ATCC 700669Δ*hexB*. See text for details.

2.3.2 Construction of *S. pneumoniae* ATCC 700669Δ*hexB*

The sequences upstream and downstream of *hexB* in *S. pneumoniae* ATCC 700669 each contain similar BOX elements, which encode *Bmg*BI cut sites (Figure 2.1A i). Approximately 500 bp regions flanking *hexB*, both including the *Bmg*BI cut sites, were amplified by PCR and cloned into pGEM-T Easy as described above to give plasmids pLeft and pRight (Figure 2.1A i, B i, ii). The *ermB* erythromycin resistance gene was then amplified from strain *S. pneumoniae* 11930 using the primers ermBL and ermBR (each with an *Eco*RV cut site on the 5' end; Appendix I: Primer sequences), and cloned into pGEM-T Easy as described above to give plasmid pErm (Figure 2.1A ii, B iii). All three plasmids were then transformed into *E. coli* TOP10 cells through electroporation with a 2.5 kV pulse, followed by blue-white selection on ampicillin, as described above. The plasmid carrying the *ermB* gene was then extracted using a QIAprep Spin Miniprep Kit (Qiagen) and digested using *Eco*RV (New England Biolabs), releasing the insert with blunt ends. This fragment was purified through agarose gel electrophoresis as described above. Plasmid pRight was similarly extracted, then digested with *Bmg*BI (New England Biolabs), which also

cuts to give blunt ends. The *ermB* fragment was cloned into this blunt cut site using T4 ligase (Promega) overnight at 4 °C according to manufacturer's instructions (Figure 2.1C ii). A 1 μL sample of this reaction was then used to transform electrocompetent *E. coli* TOP10 cells as described above. A 50 μL sample of this culture was then spread on LB agar plates supplemented with 100 μg mL^{-1} erythromycin. The plasmid pErmR was isolated from a colony, and its composite insert and the adjacent multiple cloning site (MCS) amplified using primers ermBL (which has an *Eco*RV cut site on the 5' end) and T7. This amplicon was purified through agarose gel electrophoresis using a QIAquick Gel Extraction kit (Qiagen) and digested with with *Eco*RV and *Nco*I (New England Biolabs), the latter of which cuts within the MCS of pGEM-T Easy (Figure 2.1C ii). This was ligated into pLeft following its digestion with *Bmg*BI and *Nco*I (Figure 2.1C i) as described above. This ligation reaction was used to transform *E. coli* TOP10 cells as described above, with strains carrying the plasmids again selected on LB agar supplemented with 100 μg mL^{-1} erythromycin. This gave pErmLR, which was used to transform *S. pneumoniae* ATCC 700669[lab]. Pneumococcal transformants with disrupted *hexB* genes were selected on 5% blood agar plates supplemented with 0.1 μg mL^{-1} erythromycin. One of these colonies was picked and the insert checked through PCR amplification with primers hexBL and hexBR and capillary sequencing.

2.4 DNA and RNA sequencing

2.4.1 Genome sequencing

All processing and sequencing of genomic DNA samples was performed by the Wellcome Trust Sanger Institute's core sequencing teams. DNA sample concentrations were determined using the Qubit system (Invitrogen) and subsequently diluted to 2.5 μg in 100 μl Tris-EDTA buffer. Unless specified otherwise, all samples analysed as part of the PMEN1 and ST180 populations or the *in vitro* transformation experiment were sequenced as multiplexed libraries of up to twelve isolates using the Genome Analyzer II (Illumina). Libraries were constructed according to manufacturer's instructions with a specified insert size of 250 bp. Briefly, DNA samples were first sheared by nebulisation (35 psi, 6 min). Duplexes were then blunt

ended through an end repair reaction using large Klenow fragment, T4 polynucleotide kinase and T4 polymerase. A single 3' adenosine moiety was added to the cDNA using Klenow exo$^-$ and dATP. Illumina adapters, containing primer sites for flow cell surface annealing, amplification and sequencing, along with one of the twelve unique tag sequences in multiplexed libraries, were ligated onto the repaired ends of the DNA. Gel electrophoresis was used to select for DNA constructs around 250 bp in size, which were subsequently amplified by 18 cycles of PCR with Phusion polymerase. These libraries were denatured with sodium hydroxide and diluted to 3.5 pM in hybridization buffer for loading onto a single lane of an Illumina Genome Analyzer II flow cell (Illumina). Cluster formation, primer hybridization and sequencing reactions were according to the manufacturer's recommended protocol. Data for all PMEN1 and ST180 samples was in the form of paired end 54 nt reads, while data for all transformant experiments was in the form of 76 nt paired end reads.

The reference genomes of *S. pneumoniae* ATCC 700669 and *S. pneumoniae* OXC141 were generated through capillary sequencing (Appendix III: EMBL accession codes). Briefly, a shotgun sequence with ~8-fold genome coverage was achieved through sequencing of pUC clones with 1.4- to 2.8-kb inserts and pSMART clones with 8- to 12-kb inserts using a BigDye terminator sequencing kit and AB 3700 sequencers (Applied Biosystems). Sequences from 30- to 40-kb pEpiFOS-5 fosmid clones and 12- to 23-kb pBACe3.6 BAC clones were used to scaffold contigs and bridge repeats. The sequence was finished according to standard criteria (Parkhill *et al*., 2000). Sequence assembly, visualization, and finishing were performed by using PHRAP (www.phrap.org) and Gap4 (Bonfield *et al*., 1995). All repeat sequences were independently verified.

In order to ascertain a better view of the accessory genomes of PMEN1 strains *S. pneumoniae* 11876 and 11930, and ST180 strains *S. pneumoniae* 03-4156, 03-4183, 07-2838, 02-1198, 99-4038 and 99-4039, genomic DNA samples were sequenced using the 454 platform (Roche). Assemblies were generated by using Newbler to produce scaffolds from the 454 data that were subsequently improved by iteratively mapping the Illumina data to the draft assembly using IMAGE (Tsai *et al*., 2010).

2.4.2 Transcriptome sequencing

All processing and sequencing of complementary DNA samples was performed by the Wellcome Trust Sanger Institute's core sequencing teams. Complementary DNA samples were sequenced using the Illumina Genome Analyzer II (Illumina) as described for genomic DNA samples, with the exceptions that the specified insert size was 150 bp and each sample was sequenced as a non-multiplexed library to ensure sufficient coverage. Paired RNA-seq samples for differential expression analyses were sequenced on adjacent lanes on the same flow cell, to avoid variation introduced between cells (Marioni *et al*., 2008). All data was in the form of 54 nt paired end reads.

2.4.3 Oligonucleotide sequencing

2.4.3.1 DNA and RNA synthetic oligonucleotide mixtures

All processing and sequencing of DNA and RNA oligonucleotide mixture samples, used for validation of the RNA-seq methodology, was performed by the Wellcome Trust Sanger Institute's research and development sequencing team. A DNA oligonucleotide with the sequence AACATCTGCAAG$(N)_{19}$CAGCGACGCATC$(N)_5$ (Sigma), either alone or in the presence of an equimolar amount of a 3' phosphorylated RNA oligonucleotide of sequence GAUGCGUCGCUG (Sigma), was diluted to a concentration of 120 nM in Tris–EDTA buffer and subjected to standard Illumina library preparation reactions. Following Illumina library construction, as described above, the DNA and RNA oligonucleotides were subjected to both 36 nt paired end sequencing and 54 nt single end sequencing as non-multiplexed libraries on the Illumina Genome Analyzer II.

2.4.3.2 PCR product and plasmid sequencing

All capillary sequencing of PCR products and plasmids was performed by the Wellcome Trust Sanger Institute's faculty sequencing team. Prior to submission, all PCR products were purified through agarose gel electrophoresis and all plasmids purified using the QIAprep Spin Miniprep Kit (Qiagen). These samples were then diluted to a concentration of 10 ng mL^{-1} in 50 μL, measured through the Qubit system

(Invitrogen), in preparation for 3.1 Bigdye sequencing on AB 3730 capillary sequencing machines (Applied Biosystems).

2.5 Alignment and assembly of short sequence reads

2.5.1 Generation of whole genome alignments for phylogenetic analyses

Illumina sequence data were mapped to the appropriate complete reference genome as paired end reads with an insert size between 50 and 400 bp using either SSAHA2 (Ning *et al*., 2001) for PMEN1, or otherwise SMALT (Ponstingl, 2011). The reference sequences used were *S. pneumoniae* ATCC 700669 for PMEN1 (Chapter 4), *S. pneumoniae* ATCC 700669^{lab} for the *in vitro* transformation experiments (Chapter 5) or *S. pneumoniae* OXC141 for ST180 (Chapter 8). For those ST180 strains that had only 454 or capillary data, paired end reads of the appropriate length and insert size were simulated from the best available assembly and analysed in parallel with those isolates sequenced using the Illumina platform. SNPs were identified as described in Harris *et al.* (Harris *et al*., 2010). Briefly, only reads aligning to the reference sequence with a quality score of greater than 30 were considered. For each position, a base was only called if the Phred quality score of the site was above 50 (theoretically equating to an accuracy of 99.999%; www.phrap.org), and the call was supported by 75% of at least four reads, with at least two on each strand. Otherwise, the position was recorded as unknown.

For PMEN1, small indels were identified from the SSAHA2 output. Indels were called where more than 75% of the reads (corresponding to at least five reads) spanning the site supported the change in sequence length. If between 25% and 75% of the reads supported an indel identical to one confidently identified in another strain, the indel was marked as missing data. The sequence of insertions was called as a base if 75% or more of the reads supporting the change agreed on a consensus; otherwise, the position was treated as an unknown base. For the *in vitro* transformation and ST180 samples, indels were identified using bcftools (Danecek *et al*., 2011) and filtered using the same criteria as other sites in the genome.

2.5.2 Generation of a genome alignment for *in vitro* transformation analysis

A draft genome assembly for strain TIGR4Δ*cps* was generated by splicing together the sequence of the disrupted capsule biosynthesis locus (Pearce *et al*., 2002) [EMBL accession code AF160759] with the rest of the TIGR4 genome (Tettelin *et al*., 2001) [EMBL accession code AE005672]. Illumina sequence data generated from the DNA used for transformation was then used to correct this sequence using ICORN (Otto *et al*., 2010). The sequence of the *S. pneumoniae* ATCC 700669 line used in this experiment (*S. pneumoniae* ATCC 700669^{lab}) was derived by correcting the reference sequence with ICORN (Otto *et al*., 2010) using resequencing data; this revealed the presence of four substitutions and the loss of prophage ΦMM1-2008. The finalized genomes were aligned using MUGSY (Angiuoli and Salzberg, 2011) to allow marker polymorphisms to be identified. To avoid the false positive identification of polymorphisms, the hypervariable *hsdS* locus (Tettelin *et al*., 2001) and highly repetitive *psrP* gene were excluded from all analyses.

2.5.3 Assembly of Illumina data

De novo assemblies of prophage, *cps* loci and partially deleted ICE*Sp*23FST81 sequences were produced from multiplexed Illumina sequence data. Assemblies were initially generated from EDENA (Hernandez *et al*., 2008), an overlap graph-based assembler. The overlap parameter was iteratively increased to obtain the highest N_{50} value. Any contigs over 1.5 kb in length were then used to generate a FASTQ file of simulated reads 350 bp in length separated by an insert size of 800 bp. These were then used in conjunction with the original data as an input for Velvet (Zerbino and Birney, 2008), a de Bruijn graph-based assembler. Velvet was optimized to run with the longest k-mer that gave an expected coverage value above 20. Contigs were then ordered against the appropriate reference sequence using ABACAS (Assefa *et al*., 2009) and ACT (Carver *et al*., 2005). Assemblies displayed in Chapter 4 were then annotated and submitted to the EMBL database (Appendix III: EMBL accession codes).

2.5.4 Detecting accessory genome components through mapping

Following extraction and assembly, accessory genome loci detected in the PMEN1 and ST180 populations were concatenated into a multiFASTA file. Illumina sequence read data were then mapped against this reference using BWA (Li and Durbin, 2010) to produce an alignment, including redundant mapping of sequences aligned to repeats, that was processed to give a coverage plot using bcftools (Danecek *et al*., 2011). These were then displayed as a heatmap relative to the phylogeny using Biopython (Mangalam, 2002).

2.5.5 Analysis of RNA-seq data

Sequence reads were mapped as paired end data using BWA (Li and Durbin, 2010). The orientation of the second read in correctly mapped pairs was reversed using Samtools (Li *et al*., 2009) before producing coverage plots, in order to maintain the directional fidelity of the data. The 'XA' note in the alignment file was used to identify alternative mapping locations. All reads were used to generated the fully redundant plot in Figures 7.6 and 7.7; reads with alternative mapping loci only within the displayed region were maintained in the set used to generate the 'locally redundant' plot, whilst those that mapped equally well to sequences outside of the displayed region were excluded. This allows reads that come from a specific BOX element, but cannot be unambiguously assigned to a particular boxB module therein, to be retained within the 'locally redundant' plot.

2.6 Bioinformatic analyses

2.6.1 Mathematical analyses

All statistical test *p* values and graphical representations were generated using R (R Development Core Team, 2011).

2.6.2 Annotation of sequences

Coding sequences were initially identified by using Glimmer3 (Delcher *et al*., 2007) and then manually curated using Frameplot (Bibb *et al*., 1984) and Artemis (Carver *et*

al., 2008). All genes were annotated in Artemis using standard criteria (Berriman and Rutherford, 2003). Genome comparisons were performed using BLAST (Altschul *et al*., 1997) and visualised using ACT (Carver *et al*., 2005).

2.6.3 Determination of serotype and sequence type from Illumina data

The sequences of 91 pneumococcal *cps* loci (Bentley *et al*., 2006; Park *et al*., 2007) were concatenated and Illumina sequence reads redundantly aligned against this reference using BWA (Li and Durbin, 2010). The locus with the highest proportion of its length covered by mapped sequence reads was taken to be that encoding the capsule.

The sequences of the seven loci used for sequence typing, along with several hundred base pairs of flanking sequence, were extracted either from the genome of *S. pneumoniae* ATCC 700669 or *S. pneumoniae* OXC141. Five rounds of Illumina read mapping were then used to iteratively transform the reference sequences into those of the sequenced isolate using ICORN (Otto *et al*., 2010). The sequences were then analysed using www.mlst.net (Aanensen and Spratt, 2005). The sequence type of *S. pneumoniae* BM4200 was independently verified from the *de novo* whole genome assembly of the strain.

2.6.4 Recombination and phylogenetic analyses

The algorithm described in Chapter 4 was implemented in order to produce a maximum likelihood phylogeny based on vertically-inherited substitutions occurring outside of recombinations. When applied to PMEN1, the algorithm was run for five iterations, but failed to converge. This was a consequence of the low probability of resolving the short, poorly supported branches at the base of the phylogeny in the same manner in subsequent iterations. Convergence was instead assessed through comparing the Robinson-Foulds distance between the trees (Felsenstein, 1989) produced by each of the iterations, which showed that the phylogenies were highly similar in the three preceding, and following, iterations. By contrast, when applied to the ST180 dataset, the algorithm converged on a stable topology by the third iteration.

2.6.5 Analysis of gene disruption events

Frameshift mutations and premature stop codons that reduced the length of CDSs relative to their annotation in the reference genome were defined as 'disruptive events'. These were reconstructed onto the phylogeny using parsimony, as PAML (Yang, 2007) is unable to reconstruct changes in sequence length. For PMEN1, 537 disruptive events were estimated to occur, giving a mean incidence of 0.278 kb^{-1} disruptive events across the 1,934,819 bp of coding sequence in the reference genome. Modelling the occurrence of these disruptions as a Poisson distributed process occurring at a rate proportional to the length of the gene, 11 CDSs exceeded a *p* value threshold of 0.05 after a Bonferroni correction for multiple testing of 2,135 CDSs. For ST180, 230 disruptive events were identified across 1,756,252 bp of coding sequence, giving a mean incidence of 0.131 kb^{-1}. Assuming the same Poisson model, only two functional CDSs exceeded the Bonferroni corrected 0.05 *p* value threshold: SPNOXC10420 and SPNOXC12950.

2.6.6 Bayesian phylogenetic analyses

BEAST (Drummond and Rambaut, 2007) was used to date the most recent common ancestors of the two 19A clades in PMEN1 (clade 'U' from the USA and clade 'S' from Spain) and the most recent common ancestors of ST180 and clade I in the analysis of the serotype 3 isolates. The program was used to analyse the final maximum likelihood tree, the topology of which was fixed, and the alignment of base substitutions occurring outside of putative recombinations using an uncorrelated lognormal relaxed molecular clock (Drummond *et al*., 2006). The tree was calibrated using the strains' dates of isolation. The ages of strains for which no precise date of isolation was available were estimated using a uniform distribution spanning the range of years from which the sample could have been isolated. The non-parametric Bayesian skyline plot was used as the tree prior to allow for fluctuation in population size (Drummond *et al*., 2005). A general time reversible model of substitution was used, but no evidence of a requirement of different rate categories was found. Data were combined from multiple runs, with the appropriate 'burn in' removed from each on the basis of parameter traces, such that all values had an effective sample size greater than 200. As validation of the application to PMEN1, the analysis estimated

that the lineage originated around 1969 (95% credibility interval 1958-1977), in concordance with the date calculated from root-to-tip distances (about 1970).

2.6.7 Alignment of assembled sequences

Phylogenies were constructed for the serotype 19A and 19F *cps* loci using RAxML (Stamatakis *et al*., 2005) by extracting these regions from the *de novo* strain assemblies and aligning them as co-linear sequences with progressiveMauve (Darling *et al*., 2010).

2.6.8 Clustering of prophage elements

Prophage sequences were extracted from the host genomes based on the positions of the autolysin and integrase genes at the two ends of such elements. Gene prediction was performed on all the prophage concatenated together using Glimmer 3 (Delcher *et al*., 2007); the putative protein sequences were then extracted and compared using BLASTP (Altschul *et al*., 1997) with an E value cutoff of 10^{-50}. Orthologue clusters were then defined by analysing these sequences with TribeMCL (Enright *et al*., 2002), using an inflation parameter value of 1.5. A Jaccard distance matrix was then constructed on the basis of the number of shared and unique orthologue clusters in each pairwise comparison between phage. This was used to produce a neighbour joining tree using Neighbor in the Phylip package (Felsenstein, 1989).

2.6.9 Analysis of repeat sequences

2.6.9.1 Identification of repeat sequences

The sequence of *S. pneumoniae* ATCC 700669 was searched for repeats longer than 50 bp using RepeatScout (Price *et al*., 2005). For each of the three families identified, multiple sequence alignments were produced with MUSCLE (Edgar, 2004), which were used to generate Hidden Markov Models (HMM) using HMMER1.8 (more recent versions of HMMER have not been optimised for searching long nucleic acid sequences for short motifs) (Eddy, 2008). In order to define the modular nature of BOX elements, HMMs representing boxA, B and C sequences individually were

produced using available sequence data (Martin *et al*., 1992; Koeuth *et al*., 1995). Sequences identified with these initial models were then aligned and used to produce the final HMMs used in this study; cutoff score thresholds were determined empirically from the distribution of scores for all hits throughout the genome. HMM logos were produced using LogoMat-M (Schuster-Bockler *et al*., 2004). Composite BOX elements were defined as two or more adjacent boxA, B or C modules. The same approach was used to generate HMMs for the repeats identified in *S. suis*. Thorough *de novo* searches for novel interspersed repeats in other species were not conducted.

In a number of cases where annotated repeat sequences overlapped, it was evident that one element had inserted into another. In such cases, for each repeat in the pair, a realignment of one repeat with the appropriate HMM was attempted using the concatenated flanking sequences of the other repeat, effectively excluding the sequence of the other element. If one of the elements had a greater bit score when realigned in such a manner, it was reannotated as a split feature into which the other repeat had inserted. The HMMs for the *S. pneumoniae* and *S. suis* repeats, and a program to automate their annotation for viewing in Artemis (Carver *et al*., 2008), are freely available from ftp://ftp.sanger.ac.uk/pub/pathogens/strep_repeats/. The annotation of repeat elements in complete *S. pneumoniae*, *S, mitis and S. suis* genomes is also available from this site.

2.6.9.2 Definition of orthologous repeat sequences

Of the 14 available complete pneumococcal genomes in the EMBL database, all except *S. pneumoniae* R6 (a laboratory derivative of *S. pneumoniae* D39, the sequence of which is also available in the database) were analysed. For each annotated repeat element, 250 bp of upstream and downstream flanking sequence were concatenated into a single 500 bp string. All pairwise sequence comparisons between strings corresponding to repeats of the same type were performed using BLASTN (Altschul *et al*., 1997). The alignments with an E value smaller than 10^{-25} were then used to cluster the strings into groups, corresponding to orthologous repeat insertions, using OrthoMCL (Li *et al*., 2003). The inflationary parameter used in clustering was set to 3, the smallest integral value that did not cluster a pair of

insertions within the same genome together (*i.e.* identify 'paralogous' insertions). The IS elements in Figures 7.3 and 7.4 correspond to all the annotated IS element transposase CDSs in the *S. pneumoniae* ATCC 700669 genome; however, in order to identify orthologous IS elements in different pneumococcal genomes, a consistent annotation across the genomes was required. To automate this, such repeats were identified as BLASTN (Altschul *et al*., 1997) matches to defined elements in the IS database (Siguier *et al*., 2006) with a nucleotide identity >95% and a length >90% of that of the reference sequence. Insertions of the same IS element type were then clustered as described for the small interspersed repeats, and the results for all IS elements subsequently combined to generate the data used in the graph.

2.6.10 Prediction of RNA secondary structures

2.6.10.1 Hypothetical structures of interspersed repeat sequences

Secondary structure predictions were produced from a multiple alignment of 30 repeat sequence examples (a random sample in the base of RUP elements; only BOX elements with the canonical $A_1B_1C_1$ structure were used) using RNAalifold (Bernhart *et al*., 2008).

2.6.10.2 Hypothetical structure of *patAB* leader sequence

Starting at the putative position of the *patAB* transcript's initiation, the downstream sequence, of a length extended by one base at a time, was extracted from the genomes of *S. pneumoniae* 99-4038 and 99-4039. For each sequence length extracted, the most stable folded structure, and its corresponding free energy at 37 °C, were calculated using the Vienna RNA package (Hofacker, 2009). The stability of these structures were then plotted against the length of the sequence.

2.6.11 Analysis of *in vitro* transformation data

2.6.11.1 Identification of recombinant sequences

For each transformant, all sites identified as being polymorphic from the whole genome alignment with a base quality greater than 50 were used to identify sequence characteristic of the donor or recipient in the transformation experiment. Recombinations were initially defined as regions containing donor alleles at polymorphic sites with no intervening recipient alleles. The ambiguous flanking regions around each recombination extended between the outermost donor SNPs identified in the transformant and the nearest flanking sites that were found to have recipient allele SNPs.

Many of the 112 secondary recombinations in the same strain were positioned very close to one another. A bootstrapping approach was used to link recombinant sequences likely to have arisen through the same recombination event. The shortest distance between all pairs of non-overlapping recombinant segments outside the primary recombination locus from all strains, in terms of the donor genome, was calculated to produce the population of test values. For each strain in turn, the shortest distance between two secondary recombinations (d_{test}) was used as the test statistic. Ten thousand distances were then randomly sampled with replacement from the test population, and this distribution then used to test the hypothesis that d_{test} was significantly shorter than expected under the null hypothesis (H_0) of recombinant sequences being positioned at random relative to one another. Multiple testing was accounted for by using a Holm-Bonferroni correction to alter the one-tailed threshold p value according to the number of secondary recombinations in the transformed strain. This test was performed 100 times for each d_{test}, with distances rejecting H_0 on 95 or more of these trials considered to be significantly close to one another and therefore likely to have arisen from the mosaic incorporation of the same strand of donor DNA. If the first d_{test} in a strain was accepted, then the second smallest distance was tested with the appropriate corrected p value threshold; this process continued until d_{test} was accepted under H_0.

2.6.11.2 Sliding window analysis

There are a number of biases that must be taken into account when investigating the impact of sequence identity on recombination distributions. The non-uniform

distribution of SNPs between the donor and recipient sequences mean that there are large numbers of low-identity polymorphisms, and hence short regions of identity between SNPs, concentrated in small regions of the genome. This structuring of the sequence divergence is not commensurate with the requirement of many statistical tests that observations should be independent. However, regions of high identity are also problematic, as recombinations can only be identified and analysed where it is possible to detect them through the transfer of polymorphisms. Hence a sliding window approach was used to ascertain the effect of sequence identity, and infer the proportion of recombinations of a given length that occurred but could not be detected. Each secondary recombinant segment was analysed independently; the $L50_R$ was used as the window size, which was moved along each base of the recipient genome with an orthologous nucleotide in the donor. Using the set of SNPs that could be identified using the Illumina data, at each position it was recorded as to whether the recombination could be detected or not. If it could be identified, then the size of the ambiguous flanking sequences, and the mean identity of the SNPs within, were recorded. This allowed the proportion of possible recombinations of the same size with the same, or greater, mean SNP identity and flanking regions of equal, or greater, length. Fisher's exact test was then used to compare the null hypotheses, that half the 112 recombinations should have a greater than expected mean SNP identity, and half the recombinations should have longer than expected flanking regions, with the observed distribution.

2.6.12 Differential gene expression and phenotype analyses

2.6.12.1 RNA-seq data

Read count and RPKM values for each CDS were calculated using a custom Perl script. Differential expression analysis with the three paired replicate samples from *S. pneumoniae* 99-4038 and 99-4039 was performed using DEseq (Anders and Huber, 2010) to analyse the read count values. The *p* values presented have been corrected through the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

2.6.12.2 Expression microarray data

Microarray analyses were performed by the BμG@S group (St. George's Hospital, London). Three replicates of paired RNA samples from *S. pneumoniae* TIGR4 and TIGR4^{PUS} were analysed using the SPv1.1 PCR product microarray (Hinds *et al*., 2002a; Hinds *et al*., 2002b). Each sample was hybridised independently against a common genomic DNA control sample and the three sets of paired replicates compared using LIMMA (Smyth, 2004; Inverarity, 2009). The *p* values presented have been corrected through the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

2.6.12.3 Phenotype microarray data

The level of respiration occurring in each well of the phenotype microarray plates was quantified as the area under the curve calculated by the Biolog analysis software. Significant differences in respiration rates between strains were assessed using LIMMA (Smyth, 2004). The *p* values presented have been corrected through the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

3 The genome of *S. pneumoniae* ATCC 700669

3.1 Introduction

3.1.1 The spread of the PMEN1 lineage

The PMEN1, or Spain^{23F}-1 lineage, was one of the first multidrug-resistant pneumococcal clones to be recognised as having spread across the globe. Since the early 1980s, serogroup 23 isolates resistant to penicillin, chloramphenicol and tetracycline had been observed in Spain (Linares *et al*., 1983; Latorre *et al*., 1985) and the UK (George *et al*., 1981). The clonal nature of a number of these European isolates, along with a serogroup 19 strain with an identical resistance profile, was initially demonstrated using MLEE (Coffey *et al*., 1991), and, using the same technique, it was found that this lineage had spread to the USA by the late 1980s (Munoz *et al*., 1991). In the 1990s, both MLEE and PFGE were used to identify PMEN1 isolates in South America (Camou *et al*., 1998; Castaneda *et al*., 1998; Echaniz-Aviles *et al*., 1998), Asia (Tarasi *et al*., 1997) and South Africa (Sibold *et al*., 1992). Soon after its introduction, MLST was used to identify this clone, with an ST of 81, in a hospital in Taiwan (Shi *et al*., 1998).

By the late 1990s, PMEN1 was responsible for almost 40% of penicillin-resistant pneumococcal disease in the USA (Corso *et al*., 1998), and about 30% of penicillin-resistant paediatric disease in Latin America (Tomasz *et al*., 1998). During its spread, it expanded its repertoire of antibiotic resistances, being frequently associated with the acquisition of macrolide (Reinert *et al*., 2005b) and fluoroquinolone (Pletz *et al*., 2004) resistance. The lineage has also proved proficient at acquiring different serotypes: serotype 19F (Coffey *et al*., 1991; Coffey *et al*., 1998a), 19A (Coffey *et al*., 1998b) and 14 (Barnes *et al*., 1995) variants were detected in the 1990s. Although serotypes 23F, 19F and 14 have decreased in prevalence since the introduction of PCV7, the increasing incidence of multidrug-resistant serotype 19A disease that has followed the vaccine's use has included switched PMEN1 isolates (Moore *et al*., 2008; Munoz-Almagro *et al*., 2008; Ardanuy *et al*., 2009).

While serotype 23F strains have consistently been found to cause disease at a low level relative to their carriage prevalence, the opposite is true of serotype 4 (Brueggemann *et al*., 2004), and serotype 2, which, although no longer common in most countries, was recently found to be the most common cause of pneumococcal meningitis in Bangladesh (Saha *et al*., 2009). Hence if it is correct that serotype more strongly determines the proclivity of a strain to cause invasive disease than the rest of the bacterial genotype, then differences between the PMEN1 genome and those of TIGR4 (serotype 4) and D39 (serotype 2) may represent adaptations to causing disease at different frequencies, and being carried for different lengths of time (Sleeman *et al*., 2006). However, this may be unlikely, given that isolates with the PMEN1 genotype have been observed expressing serotypes 14 and 19A, both associated with high odds ratios for causing disease (Brueggemann *et al*., 2003; Brueggemann *et al*., 2004).

3.1.2 *S. pneumoniae* ATCC 700669

S. pneumoniae 264 was a serotype 23F isolate from the Hospital de Bellvitge, Barcelona, in 1984 (Coffey *et al*., 1991). Found to be resistant to benzylpenicillin, chloramphenicol and tetracycline, it was submitted to the American Type Culture Collection (ATCC) as the representative type strain of the PMEN1 lineage, thereafter designated *S. pneumoniae* ATCC 700669.

3.2 Description of the genome

3.2.1 Features of the chromosome

The genome of *S. pneumoniae* ATCC 700669 was sequenced using dideoxy terminator technology. The complete circular chromosome of 2,221,315 bp (39.49% GC content; Figure 3.1) contains four rRNA operons and 58 tRNA genes (all but 12 of which are adjacent to rRNA genes). An unusual asymmetry in the GC skew of the chromosome, resulting from the recent integration of a prophage and an ICE into the same replichore, is evident. There are 2,135 predicted coding sequences (CDSs), 144 (6.7%) of which appear to be pseudogenes. In common with other Firmicutes, there is a strong coding bias, with 76% of the CDSs on the leading strand. Overall, 197

(9.2%) of the predicted CDSs in ATCC 700669 are not present in TIGR4 or D39, the majority of these being present on an ICE or prophage-derived sequences. Seventy-nine IS elements can be identified, of which 73% appear to be non-functional due to disruptive mutations. Twenty-one of the IS insertions are not present in the TIGR4 or D39 genomes, the majority of which are due to IS*1167* or IS*1167A*-type elements. Both of these IS families exhibit relatively little sequence diversity when compared with others present in the chromosome, in accordance with the previously proposed hypothesis that IS isotypes spread through short bursts of transposition (Tettelin *et al*., 2001).

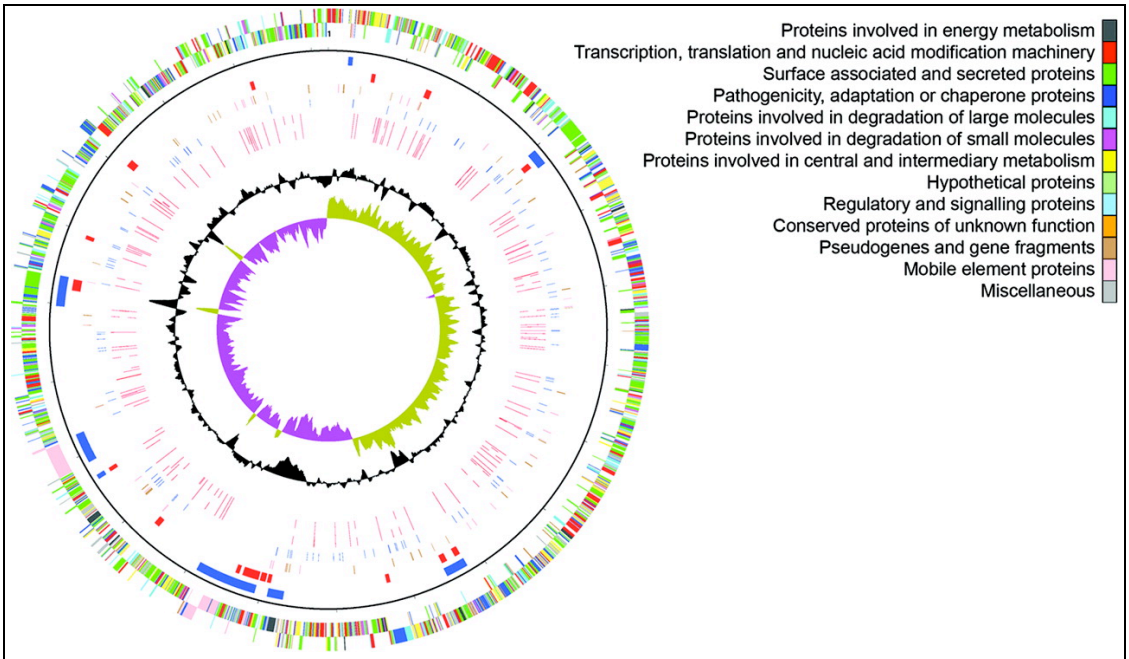


Figure 3.1 Circular diagram representing the *S. pneumoniae* ATCC 700669 chromosome, arranged to have the origin of replication at the top, as indicated by the GC deviation (innermost graph). The outer rings show the arrangement of coding sequences on the two strands of the genome, coloured according to annotated function (see key). The first inner ring indicates the major variable regions as blue blocks: moving clockwise from the origin of replication, these represent the prophage remnant, *cps* locus, PPI-1, the Na^+-dependent ATP synthase island, ICE*Sp*23FST81, the *pclA* gene cluster, ϕMM1-2008, and the *psrP* gene cluster. The red blocks demarcate loci identified as having atypical nucleotide composition by the Alien Hunter algorithm (Vernikos and Parkhill, 2006). Moving inward, the rings show the position of IS elements (pink if intact, brown if pseudogene), RUP repeats (blue), and BOX A, B, and C repeat modules (red), respectively (see Chapter 7). The black graph indicates sequence GC content, and the innermost green and purple graph shows the GC skew of the sequence.

Of the annotated genes, 29% are predicted to encode surface exposed or secreted proteins. These include 18 phosphotransferase system (PTS) transporters, 17 of which are shared with both TIGR4 and D39, which have four and three PTS transporters not present in ATCC 700669, respectively. The one system unique to ATCC 700669

(SPN23F18210-18250) forms part of a ~10-kb insertion that also includes a putative choline sulphatase. Given the importance of choline to pneumococcal metabolism and pathogenesis, this transporter could represent a novel means of acquiring this nutrient from the host environment. There are also two ABC transporter systems within the ATCC 700669 genome that are absent from both TIGR4 and D39. One of these, present on a ~4-kb insertion along with two putative secreted peptides (SPN23F07060-07090), is similar to systems present in *S. pyogenes* MGAS6180, *S. pyogenes* MGAS10270, and *S. sanguinis* SK36.

3.2.2 Genes implicated in pathogenesis

It has been shown that the *pspA* gene of a serotype 23F strain contained a frameshift mutation that truncated the encoded protein to a secreted, rather than surface-associated, form (McCool *et al*., 2002). A similar frameshift mutation, caused by variation in the length of the same polyadenosine tract, is observed in this genome. Furthermore, ATCC 700669 lacks the *zmpC* gene, which encodes a metalloprotease that cleaves and activates human matrix metalloprotease 9 and has been implicated in the pulmonary invasion process (Oggioni *et al*., 2003). However, like *S. pneumoniae* G54, *S. pneumoniae* ATCC 700669 encodes an additional surface-exposed zinc metalloprotease, *zmpD* (Camilli *et al*., 2006), adjacent to the immunoglobulin A protease gene *zmpA*. Notably absent from the ATCC 700669 genome are the genes encoding two more surface-exposed proteins, choline-binding protein *cbpC* and histidine triad protein *phtB*, which appear to be required for full virulence of TIGR4 in the mouse model, based on signature-tagged mutagenesis data (Hava and Camilli, 2002).

Notably present in the genome are both loci associated with invasive clones by Obert *et al* (Obert *et al*., 2006): the V-type Na^+-driven ATPase gene cluster and the island encoding *psrP*. Though ATCC 700669 lacks both identified pneumococcal pilus synthesis gene clusters (Barocchi *et al*., 2006; Bagnoli *et al*., 2008), *pclA* is present, although the D39 orthologue is ~44% longer due to an expansion in the internal repetitive region of the protein.

3.2.3 Prophage

Although the *S. pneumoniae* G54 and D39 genomes are devoid of prophage, a 10.5-kb phage remnant can be found between the *eno* and *rexB* genes in the chromosome of TIGR4 (Obregon *et al*., 2003). The ATCC 700669 genome contains an intact 39.1-kb prophage, ϕMM1-2008, as well as a smaller prophage remnant. ϕMM1-2008 is more than 97% identical, at the nucleotide sequence level, to both ϕMM1 and ϕMM1-1998. Since ϕMM1 and ϕMM1-2008 are from hosts of the same serotype and sequence type (Obregon *et al*., 2003), they are probably descended from the same insertion event, while the host of ϕMM1-1998 is a penicillin-sensitive serotype 24 strain (Loeffler and Fischetti, 2006); hence, this prophage is likely to be the result of a different infection. All three phage are present in the same locus, between a pyridine nucleotide-disulfide oxidoreductase gene and a CDS of unknown function. The exact insertion site of the phage appears to be within the 3' region of the downstream hypothetical gene; duplication of this 15-bp *att* sequence (Gindreau *et al*., 2000) upon integration maintains the full-length target gene sequence and generates the tandem repeats either side of the prophage.

The 6.4-kb prophage remnant is flanked by genes encoding a putative cytidine deaminase and a deoxyuridine 5' triphosphate nucleotidohydrolase. Along with CDS of phage origin, including integrase and amidase pseudogenes, the prophage appears to carry ‘cargo’ genes that have been retained as the replicative machinery of the virus has degenerated. One of these CDSs encodes a type I restriction endonuclease domain and seems to be a member of a family of uncharacterized genes found in a range of bacterial species. Another is an addiction system toxin gene, which may have inhibited the clearance of the remainder of the prophage from the genome. Although a cognate antitoxin cannot be reliably identified, the overlapping upstream gene is a good candidate, as alignments with intact prophage indicate these CDSs have been acquired as a pair. A complete 37.5-kb long prophage can be found at this locus in the draft genome of *S. pneumoniae* 18-BS74 (Hiller *et al*., 2007), suggesting it may be a common target insertion site for temperate pneumophage.

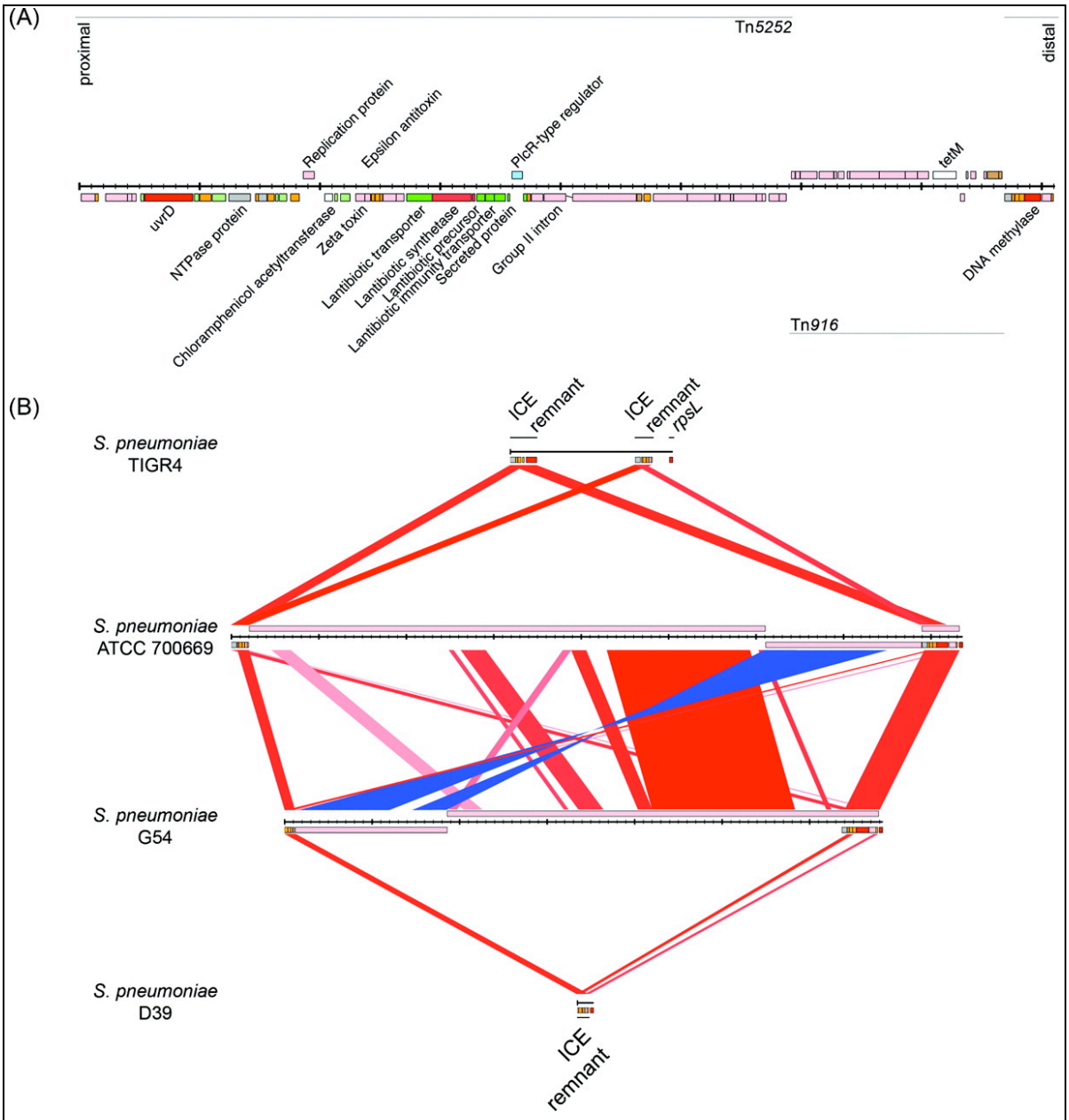


Figure 3.2 (A) Representation of ICE*Sp*23FST81. 'Cargo' genes are coloured according to the scheme detailed in Fig. 3.1 and are labelled with their putative function, where one can be assigned. The division of the ICE into Tn*5252* and Tn*916*-type elements is indicated by the bars at the top and bottom of the diagram. (B) Comparison of the ICE insertion sites in the ATCC 700669 and G54 genomes with the corresponding loci in the TIGR4 and D39 genomes, which lack intact ICEs. The ICEs are represented as pink blocks: Tn*5252*-type elements are represented by blocks above the scale line, while Tn*916*-type elements are represented by blocks below the scale line. Red bands indicate BLAST matches between genomes in the same direction, whereas blue bands indicate matches in opposite directions. The intensity of the band represents the strength of the match. The region of each genome shown is bounded at the 5' end by ICE-derived sequences and at the 3' end by *rplL*, which streptococcal ICEs frequently insert directly upstream of. The ICE remnants in the TIGR4 and D39 genomes, apparently derived from the distal region of the Tn*5252*-type element, are marked. The alignment shows that there is a remnant in D39 directly adjacent to *rplL*, at the point where the elements usually insert, but the two remnants in the TIGR4 genome are further removed upstream.

3.2.4 ICE*Spn*23FST81

The ~81-kb ICE of ATCC 700669, ICE*Sp*23FST81, is a composite element comprising a Tn*916*-like component inserted into a Tn*5252*-like transposon, with the latter consequently being split into a larger proximal region and a smaller distal region (Figure 3.2). This combination of conjugative transposons is common in streptococcal ICEs, although the variation observed in the arrangement of these two elements implies it has arisen independently on a number of occasions (Figure 3.3). This suggests a potential symbiotic advantage between these different transposon types, perhaps resulting from a synergistic combination of the two sets of conjugative machinery or 'cargo' genes. Flanked by a 16-bp tandem duplication, ICE*Sp*23FST81 is inserted near the 3' end of *rplL*, in the same position as that of *S. pneumoniae* G54 and many other streptococcal ICEs, although Tn*5253*, the partially sequenced ICE of *S. pneumoniae* BM6001 (Ayoubi *et al*., 1991), appears to have integrated elsewhere.

Shortly upstream of *rplL* in the TIGR4 and D39 genomes are ~1.8-kb long ICE remnants that are >80% identical, at the nucleotide level, to the distal Tn*5252*-type region of ICE*Sp*23FST81 (Figure 3.2). Similar remnants are also seen in the ATCC 700669 and G54 genomes immediately upstream of their ICE insertions. A second, larger, ICE remnant is also evident in the TIGR4 genome, ~15 kb upstream of *rplL*. This includes a cytosine methyltransferase gene very similar to homologues in the distal region of ICE*Sp*23FST81, on Tn*5253* and in the ϕMM1 phage (85%, 85% and 45% protein sequence identity, respectively). The presence of this gene on conjugative transposons and prophage suggests it may aid the horizontal transfer of both between pneumococci, perhaps through methylating DNA prior to transfer between cells and hence avoiding the recipient's restriction systems.

ICE*Sp*23FST81 is clinically important due to its genetic 'cargo'. The Tn*916*-type component carries a *tetM* gene, responsible for the strain's tetracycline resistance. A similar Tn*916* element, also carrying the *mef(A)* macrolide resistance gene, has been detected in the gammaproteobacterial commensal and emerging nosocomial pathogen *Acinetobacter junii* (Ojo *et al*., 2006). The ~1.2-kb flanking sequences that were determined on either side of the *A. junii* Tn*916* transposon are 99% identical, at the nucleotide level, to the Tn*5252*-type sequences surrounding the Tn*916* transposon on ICE*Sp*23FST81, suggesting these composite elements can transfer between distantly

related bacteria. The Tn*5252*-like component carries a gene for chloramphenicol acetyltransferase, which appears to have been acquired through wholesale integration of the pC194 plasmid (Widdowson *et al*., 2000) originally identified in chloramphenicol-resistant *Staph. aureus*.

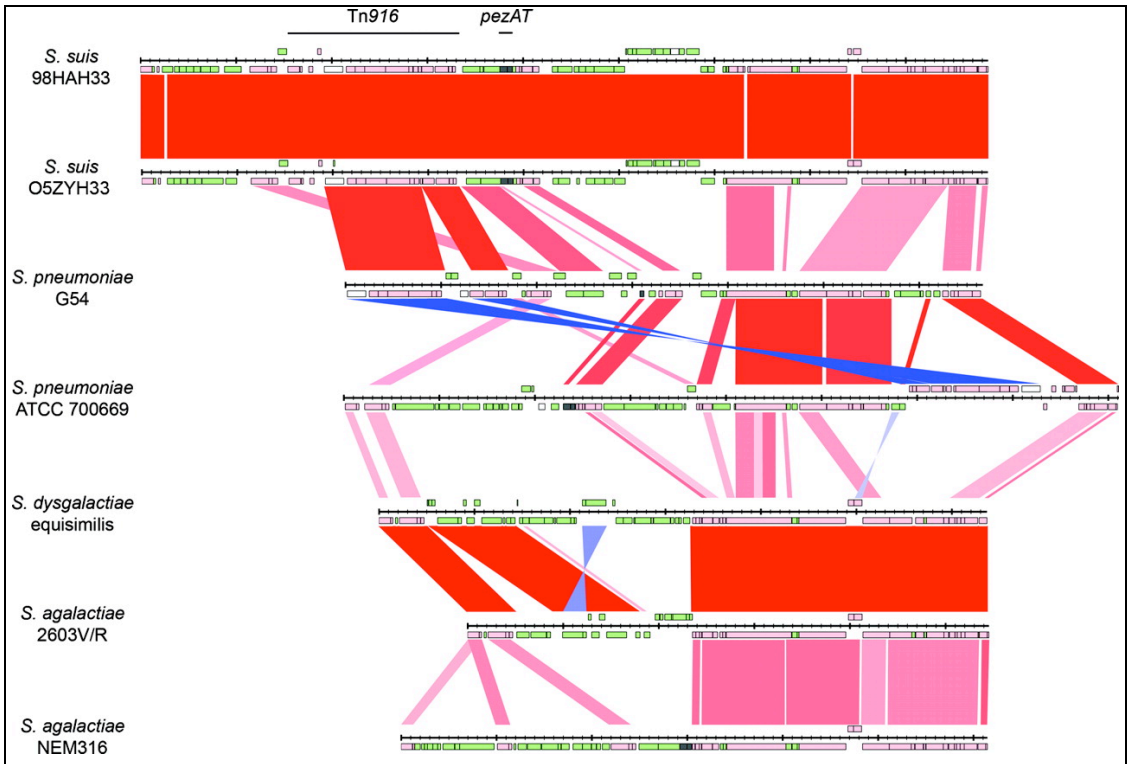


Figure 3.3 Comparison of streptococcal integrative and conjugative elements. Genes likely to be part of the conjugative machinery of the element, on the basis of conservation or functional assignment, are coloured pink. Zeta toxin-epsilon antitoxin systems are coloured grey, antibiotic resistance genes are white, and other 'cargo' genes are green. Bands are coloured as in Fig. 3.2. Unlike the *S. suis* and *S. pneumoniae* elements, the *S. agalactiae* and *S. dysgalactiae* Tn*5252*-like elements do not contain a Tn*916*-type component, which carries the *tetM* gene responsible for tetracycline resistance in the other ICE.

One of the 'cargo' genes found toward the 5' end of the element is a *uvrD* helicase gene, with the closest sequenced homologue being that of the deltaproteobacterium *Geobacter lovleyi* (26% protein sequence identity). A different *uvrD* gene, with a dissimilar sequence (15% protein sequence identity), is present at the equivalent position on the G54 ICE. Streptococci lack an SOS response (Erill *et al*., 2007), so consequently *uvrD* is absent from the *S. pneumoniae* core genome, but horizontal acquisition of this gene could potentially reconstitute the nucleotide excision repair pathway if it were able to act in concert with the *uvrABC* genes shared by all pneumococci. This pathway is important in the repair of peroxidative damage to DNA

(Moller and Wallin, 1998); therefore, given that *S. pneumoniae* is catalase negative and produces hydrogen peroxide, which can function as an antimicrobial (Pericone *et al*., 2000), the gain of the *uvrD* gene may increase the tolerance of ATCC 700669 to reactive oxygen species and hence aid nasopharyngeal colonization, while also resulting in the ICE maintaining its sequence integrity within the host. The other major branch of the SOS response, also absent from the core genome of pneumococci, is mutagenic lesion repair, which requires a reduction in DNA polymerase III replication fidelity caused by an interaction with the UmuC-UmuD complex. Correspondingly, Tn*5252* carries a *umuCD*-containing operon that was demonstrated to increase the UV tolerance of the host bacterium (Munoz-Najar and Vijayakumar, 1999), suggesting ICE-carried genes can functionally restore at least one aspect of the SOS response.

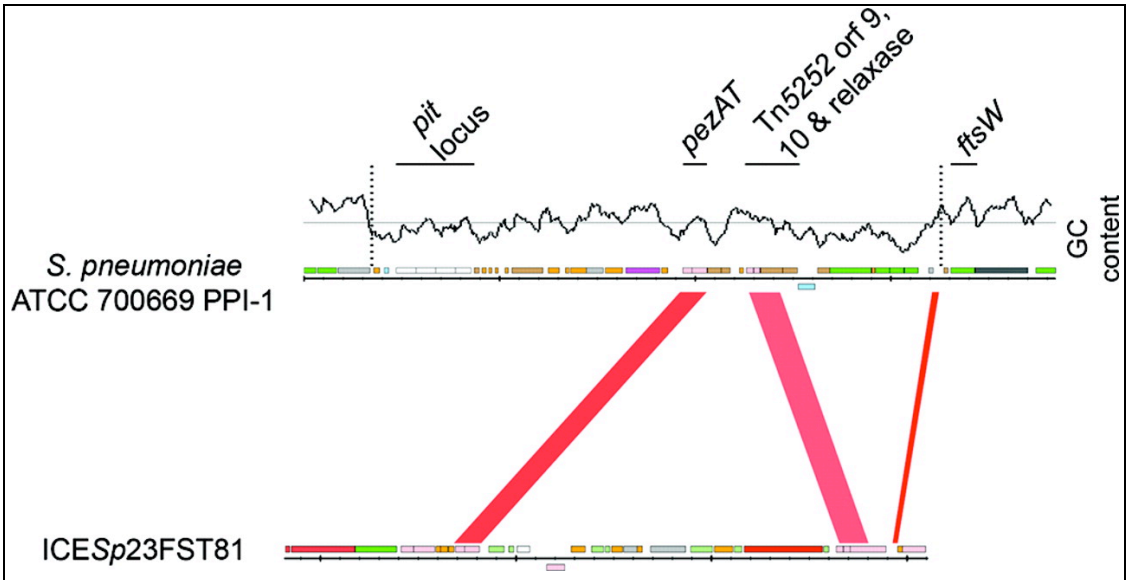


Figure 3.4 Comparison of PPI-1 of ATCC 700669 with ICE*Sp*23FST81 found in the same strain. Three regions of apparently conserved similarity are observed: the zeta toxin-epsilon antitoxin system, a group of Tn*5252* conjugative transfer genes (encoding ORF9, ORF10, and relaxase proteins), and a short stretch of DNA adjacent that forms the 3' border of PPI-1, shortly upstream of the cell division gene *ftsW* in the pneumococcal chromosome. The GC content of this region is shown, with the line across the graph indicating the average for the region (33.71% GC). The vertical dotted lines on the graph delimit the extent of PPI-1.

3.2.5 An ICE-derived genomic island

Genes characteristic of streptococcal ICE are also found in another region of the genome, within the putative Pnuemococcal Pathogenicity Island 1 (PPI-1), as described by Brown *et al*. (Brown *et al*., 2004). In *S. pneumoniae* ATCC 700669, this

~30-kb region, as defined by its low GC content (SPN23F09511-09860; Figure 3.4), contains the *pezAT* epsilon toxin-zeta antitoxin system, found on ICE*Sp*23FST81, as well as related elements in *S. suis* and *S. agalactiae* strains, and a cluster of three Tn*5252* conjugative machinery genes, including a relaxase and a MobC-domain protein. At the 3' end, coinciding with the edge of the low GC region, there is a further ~200-bp region of similarity (>90% identity at the nucleotide level) with ICE*Sp*23FST81, which is also shared with Tn*5253* and the ICE of the recently sequenced *S. pneumoniae* CGSP14 strain (Ding *et al*., 2009). A site-specific recombinase, similar to one found in *S. suis* ICEs, is found adjacent to this sequence in some pneumococcal strains, such as *S. pneumoniae* 14-BS69. Hence, it appears likely that this island originated as an ICE insertion that has subsequently degenerated, with the loss of genes required for the element's autonomous mobility. The gene clusters located between these ICE-like regions are very different in many of the strains for which genome data are available (Figure 3.5), suggesting this locus may be able to diversify through exchange of sequences with ICE via homologous recombination in the shared regions.

A further source of variation appears to be the extent to which the conjugative machinery at this locus has degenerated (Figure 3.5), with *S. pneumoniae* 18-BS74 missing all of the ICE-like regions of the island. In contrast, the 5' end of PPI-1, containing the *pit* iron transporter operon crucial for virulence of *S. pneumoniae* (Brown *et al*., 2001), is conserved among all strains. However, it seems likely the *pit* genes were acquired as part of the original ICE insertion, since they also lie within the low-GC-content region and, despite being ubiquitous among sequenced pneumococci, appear to be absent from other streptococci (Brown *et al*., 2001).

The ability to exchange sequences with conjugative elements may play a role in allowing this locus to rapidly evolve in response to changing selection pressures. In strains that have retained the variable region, CDSs similar to genes encoding daunorubicin efflux transporters, inhibitor-resistant methionyl tRNA synthetases, macrolide efflux pumps, and apparently truncated aminoglycoside phospho- and acetyltransferases are found. Furthermore, two of the sequenced Pittsburgh disease

isolates (9-BS69 and 14-BS69) seem to carry chloramphenicol acetyltransferase genes on the island. This is at least the fourth documented case in which Tn*5252*-type elements appear to have contributed to the transfer of chloramphenicol acetyltransferase genes to pneumococci. Tn*5253* of *S. pneumoniae* BM6001 and ICE*Sp*23FST81, both intact, potentially mobile, elements, carry chloramphenicol acetyltransferase genes at different positions. The PPI-1 locus and the IQ complex of *S. pneumoniae* 529, a genomic island containing Tn*5252*- and Tn*916*-like fragments along with chloramphenicol acetyltransferase and macrolide resistance genes (Mingoia *et al*., 2007), both appear to be fragments of conjugative transposons that have been integrated into the chromosome, suggesting that exchange of DNA between ICE and the pneumococcal genome is likely to be an important mechanism in the dissemination of antibiotic resistance throughout the *S. pneumoniae* population.

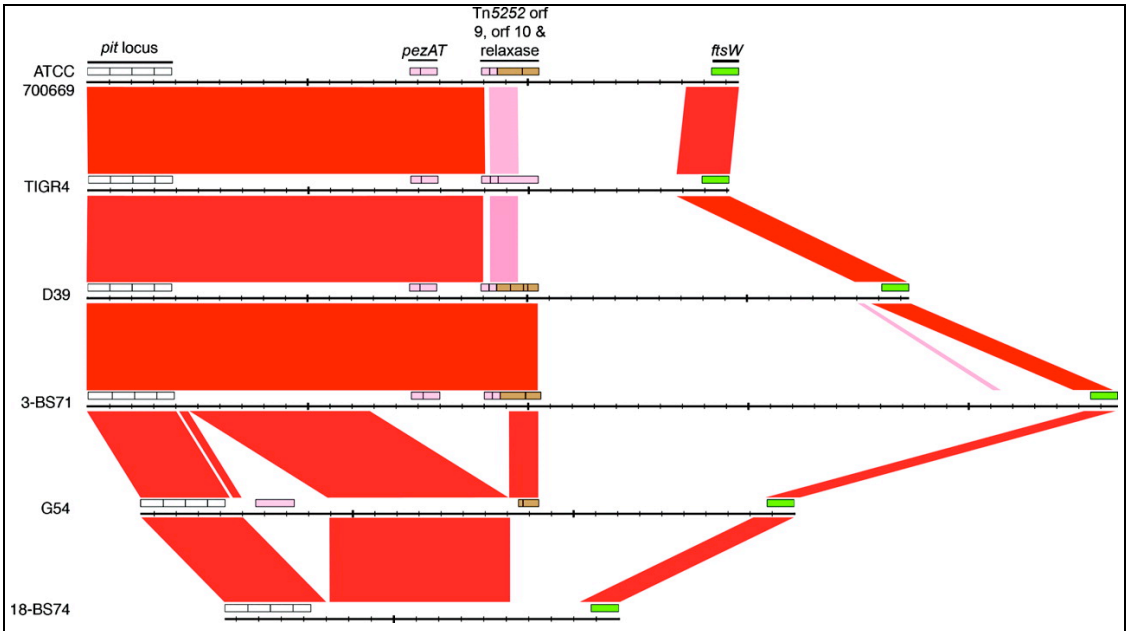


Figure 3.5 Alignment of PPI-1 from complete and draft pneumococcal genome data. Variation in the regions intervening between the relaxase and 3' end of PPI-1 (these boundaries both have sequence homology with Tn*5252*-type ICE) appears to be due to horizontal gene transfer, indicating that conjugative elements may contribute to the diversity found within this island through homologous recombination-mediated exchange. There is also variation putatively resulting from degeneration of the original conjugative element insertion: the PPI-1 of *S. pneumoniae* G54 has lost the *pezAT* toxin-antitoxin system and much of the cluster of conjugative transfer genes (although it has retained a CDS encoding a MobA-domain protein, which are typically associated with autonomously mobile elements, indicated in pink), while that of 18-BS74 appears to have lost all vestiges of the original element's conjugative machinery.

In *S. pneumoniae* ATCC 700669, PPI-1 contains an apparently incomplete lantibiotic synthesis operon. Lantibiotics (lanthionine-containing antibiotics) are small, secreted

cyclic peptides containing lanthionine rings formed by the stereospecific intramolecular addition of cysteine to dehydrated serine or threonine residues (Willey and van der Donk, 2007). They frequently function as bacteriocins, with different types hypothesized to operate through inhibiting peptidoglycan transglycoslyation or forming pores in cell membranes, but are also known to act as biosurfactants and phospholipase A2 inhibitors (Willey and van der Donk, 2007). The gene cluster present in the ATCC 700669 PPI-1 lacks a structural prepeptide gene but retains the CDS necessary for immunity. Comparison with the same locus in the serogroup 23 Pittsburgh isolate genome reveals a ~6-kb deletion in the ATCC 700669 gene cluster (Figure 3.6). The sequence absent in ATCC 700669 is flanked by thymidine dinucleotides and encodes two lantibiotic structural genes and fragments of two dehydratases. The putative product of the intact locus is a novel dimeric lantibiotic (Figure 3.6) likely to be similar to those produced by Bacillus and Lactococcus species (Willey and van der Donk, 2007).

3.2.6 Pneumocidins

In ATCC 700669, the bacteriocin-producing *blp* locus has undergone a rearrangement relative to that in TIGR4 (Figure 3.7), resulting in the deletion of *blpM* and *blpN*, both of which are required for *blp*-encoded bactericidal activity (Dawid *et al*., 2007; Lux *et al*., 2007). Given the high level of nasopharyngeal carriage of PMEN1 strains, it seems likely that loci elsewhere in the genome are able to compensate for the loss of this important pneumocidin. Consistent with this suggestion, bacteriocin production by PMEN1 strains has been observed to inhibit the growth of a larger number of indicator strains than a penicillin-sensitive serotype 23F isolate (Lux *et al*., 2007). In addition to the defunct operon in the ATCC 700669 PPI-1, ICE*Sp*23FST81 itself carries an intact lantibiotic synthesis gene cluster. The structural gene appears to be a novel group II lantibiotic, most closely related to mersacidin and lichenicidin (SPN23F12701; Figure 3.6), on the basis of its probable ring structure, diglycine cleavage motif, and net neutral charge. Adjacent to this operon is the probable transcriptional regulator, which is similar to *plcR* of Bacillus species. Such transcription factors can form a minimal extracellular signalling system in conjunction with peptide autoinducers (Slamti and Lereclus, 2002) and, in keeping with such a role, a small secreted peptide is found between the lantibiotic synthesis

genes and the *plcR* homologue.

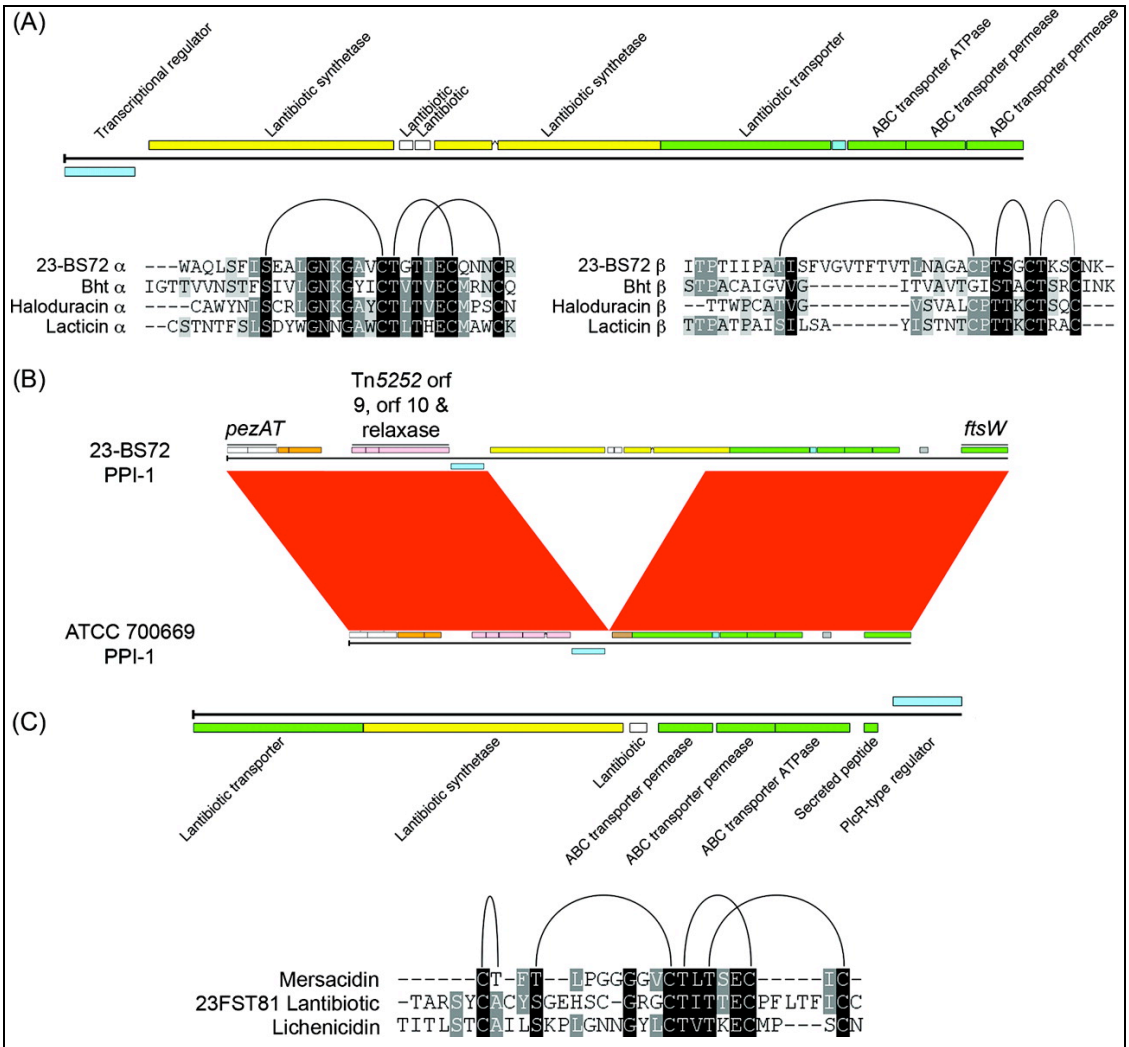


Figure 3.6 Lantibiotic synthesis gene clusters and predicted structures. The colour scheme is the same as that used in Figure 3.2. (A) Gene cluster encoded on the PPI-1 of 23-BS72. Each structural gene appears to be associated with its own synthetase. The ABC transporter is thought to be responsible for the self-immunity of the producer. The likely sequences of the mature peptides are shown aligned with other dimeric lantibiotics, with the lanthionine rings (formed by the dehydration of cysteine and serine or threonine residues) conserved between haloduracin and lacticin 3147 indicated by the curved lines above the alignment. The conservation of the functionally most important residues with these other bacteriocins indicates the 23-BS72 lantibiotic is likely to be bactericidal. (B) Alignment of the PPI-1 of *S. pneumoniae* 23-BS72 and ATCC 700669, revealing a ~6-kb deletion in the gene cluster of the latter. (C) Gene cluster encoded on ICE*Sp*23FST81. This is predicted to produce a monomeric lantibiotic, most similar to lichenicidin and mersacidin, produced by Bacillus species. The known ring structure of mersacidin is indicated, again showing that the functional, ring-forming residues are largely conserved in this new lantibiotic. It is the third ring (the most strongly conserved between the three sequences) that is thought to be most important for the bactericidal effects of mersacidin, resulting from the inhibition of peptidoglycan transglycosylation.

A similar *plcR* homologue-secreted peptide combination is present on the G54 ICE, and also adjacent to a different ~10-kb putative lantibiotic synthesis gene cluster

(SPN23F19690-19790) conserved in the chromosomes of ATCC 700669, TIGR4 and D39, next to the *rpoBC* genes. This locus also appears to be a recent addition to the genome on the basis of its atypical nucleotide composition (Vernikos and Parkhill, 2006) and absence from other pneumococcal sequences (*e.g.*, *S. pneumoniae* OXC141, EMBL accession code FQ312027 and *S. pneumoniae* INV104, EMBL accession code FQ312030). In addition, shortly upstream of the ICE*Sp*23FST81 insertion site lies another region (SPN23F12290-12330) in these three genomes that appears to have been recently horizontally acquired, on the basis of its nucleotide composition and flanking tandem repeats, which contains the remains of a set of lantibiotic processing machinery. Hence, the range of bacteriocins that individual pneumococcal strains are able to produce appears to change over time as new gene clusters are gained and old ones degenerate. If these antimicrobial peptides are associated with specific extracellular signalling systems, this could further increase the scope for competition between strains, since differences in signaling and regulation of the bacteriocins would also vary between different genotypes. This is likely to result in strains having a variable secretome that could strongly influence intra- and interspecific competition within the nasopharynx.

3.3 Discussion

Even in a species that exchanges genetic material as readily as *S. pneumoniae*, a small number of clones dominate the population of antibiotic-resistant pneumococci (Klugman, 2002), suggesting either that multiple resistances rarely successfully accumulate in a single strain, or that other factors in the chromosome are responsible for the apparently high fitness of particular 'pandemic clones', such as PMEN1. Two events were clearly important in the evolution of the lineage: recombinations affecting the *pbp* genes resulting in penicillin resistance, and the acquisition of ICE*Sp*23FST181, which carries the chloramphenicol and tetracycline resistances. As penicillin resistance has arisen on multiple occasions independently in *S. pneumoniae*, and ICE remnants seem to be quite widespread in the species, indicating such elements are common, it would appear likely that multiple resistances could accumulate in a single genome quite frequently. However, while ICE-carried resistance determinants are widely conserved, it should not be assumed that all penicillin-resistant *pbp* alleles are equally fit, and hence the prevalence of the lineage

may simply reflect a minimal cost of β lactam insensitivity.

Alternatively, if such arrangements are common and the associated fitness costs are not too variable, then the rest of the genotype would be likely to play a role in determining the success of the clone. In the case of PMEN1, the variation in the loci that encode bacteriocins, likely to be important in mediating competition within the nasopharynx, suggests one way in which the rest of the genome may contribute to differences in fitness between strains. The loss of the *blpMN* pneumocidin and PPI-1-encoded lantibiotic in ATCC 700669 suggests they may have become redundant during the evolution of the PMEN1 lineage. In both cases, the structural genes were lost, while those required for immunity were retained, ensuring the strain remained nonsusceptible to these compounds secreted by its competitors. A possible explanation is that the pneumocidin on the ICE may have superceded other bacteriocins produced by the strain and assisted nasopharyngeal colonization, and hence the spread, of the Spain^{23F} ST81 clone.

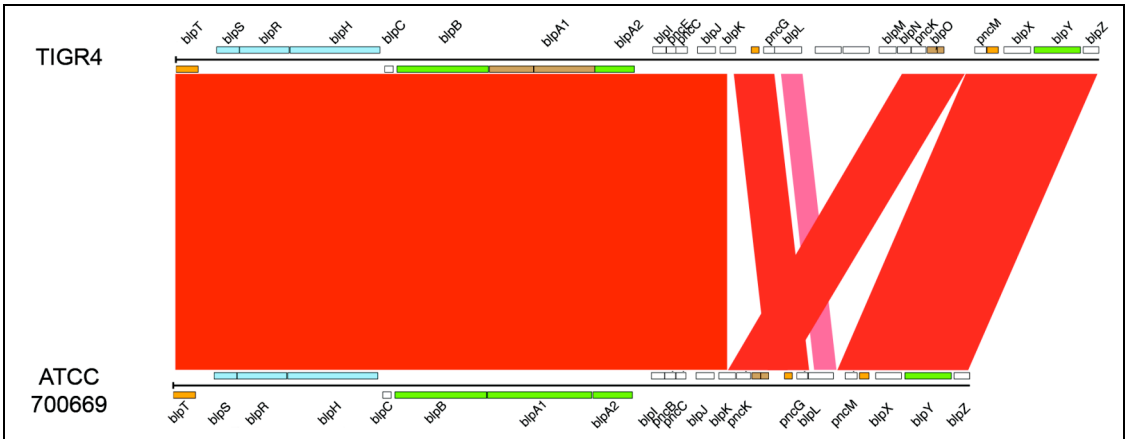


Figure 3.7 Comparison between the *blp* loci of *S. pneumoniae* TIGR4 and ATCC 700669. The *blpMN* genes, which encode a bacteriocin and have been demonstrated to be important in mediating intraspecific competition in a mouse model of pneumococcal colonisation, have been deleted from the locus in ATCC 700669.

Both this lantibiotic synthesis gene cluster and another elsewhere in the chromosome are found adjacent to genes encoding cell-density-dependent *plcR*-type transcriptional regulators and secreted peptides. Such regulation of the ICE-encoded lantibiotic synthesis gene cluster may be advantageous for the mobile element itself, since suppressing production of the antimicrobial compound until the ICE has saturated the

available population of potential hosts is likely to facilitate its horizontal transfer. Furthermore, if such quorum-sensing systems are involved in the regulation of bacteriocin production, this would add further layers of complexity onto the intercellular signalling already known to occur in *S. pneumoniae* (the two previously characterized pneumocidins, BlpMN and CibAB, are regulated by the pheromones BlpC and ComC, respectively) (de Saizieu *et al*., 2000; Claverys *et al*., 2007). Hence, just as multilocus epidemiological typing schemes are required to robustly identify lineages from an anthropic perspective, for bacteriocin-mediated competition between pneumococci to be effective several pheromone-controlled bacteriocin systems at different loci are likely to be required, as a system based on single locus with multiple alleles would be too easily confounded by a single recombination event.

4 The evolution of the PMEN1 lineage

4.1 Introduction

4.1.1 Global collection of PMEN1 strains

The complete genome of *S. pneumoniae* ATCC 700669 provides some insight into the evolution of features that are common to almost all members of PMEN1, such as penicillin, chloramphenicol and tetracycline resistance, but offers no information on the clinically important mechanisms of serotype switching and the acquisition of resistance to macrolides and fluoroquinolones. In order to study how this lineage has evolved as it has spread, Illumina sequencing of multiplexed genomic DNA libraries was used to characterise a global collection of 240 PMEN1 strains isolated between 1984 and 2008. These were identified either by using MLST or on the basis of serotype, drug-resistance profile, and targeted polymerase chain reaction (see Materials and Methods). Selected isolates were distributed among Europe (seven countries, 81 strains); South Africa (37 strains); America (six countries, 54 strains); and Asia (eight countries, 68 strains) and included a variety of drug-resistance profiles, as well as five serotypes distinct from the ancestral 23F: namely, 19F, 19A, 6A, 15B, and 3 (Appendix II: PMEN1 strains).

4.1.2 Detecting recombination in sequence data

Algorithms for constructing phylogenies assume that the entire alignment being studied has a single, common ancestry; this condition is violated in recombined sequences (Posada *et al*., 2002). Hence in order to reconstruct the history of a naturally transformable species, it is necessary to distinguish vertically inherited point mutations, which are informative about relationships between taxa, from horizontally acquired sequences, which may introduce many polymorphisms simultaneously but actually represent just a single mutational event. In the context of a phylogeny, recombinations can be defined as either ‘imports’, if the level of sequence divergence between the recombination donor and recipient is large compared to the diversity of taxa within the tree, or ‘exchanges’, if the divergence between donor and recipient is comparable to (or indeed, smaller than) the diversity within the tree. The major

impact of imports is on branch lengths, due to the increased divergence between the recipient and those taxa that retain the ancestral sequence. In addition, if multiple independent imports occur at the same locus in the studied population, then they can lead to distortions in the tree topology. This can either be the consequence of a genuine relationship between the donors in each case, such that independent imports lead to separate groups of taxa appear to look similar to one another through convergent evolution, or simply because two unrelated, but divergent, sequences can spuriously appear homologous due to long branch attraction effects. Exchanges have relatively little impact on branch lengths, unless they involve the transmission of a recent import through a population; instead, they primarily affect the topology of the tree, because distinct segments of the sequence will support different branching patterns, depending their recent history of transfer among the taxa. Hence phylogenies of populations in which recombinations, in particular exchanges, occur frequently will include a high level of homoplasy.

Early statistical tests for detecting recombination generally focussed on detecting exchanges through identifying patterns of polymorphisms in alignments that could not be explained by each mutation occurring only once in the maximally parsimonious reconstruction of the sequences' evolution. One of the first applications of this approach, using protein sequences, was an investigation of cytochrome evolution by Sneath *et al* (Sneath *et al*., 1975). Early examples using DNA sequences include tests for 'incongruent phylogenetic partititions' (Stephens, 1985) and distributions of discordant sites (Sawyer, 1989). Algorithms were also proposed for the detection of imports, such as the maximum χ^2 test (Maynard Smith, 1992), which essentially identified boundaries between regions of low, and high, SNP density. These tests all rely on detecting recombinations as uninterrupted runs of polymorphisms. However, when a population exists in complete linkage equilibrium, such contiguous segments of sequence with common ancestry no longer exist, as they are broken up by continual horizontal exchange (Maynard Smith, 1999). In such a situation, methods such as the 'homoplasy test' can be used to identify whether the population is recombining or not, but the lack of extended regions with common ancestry means identification of recombined loci themselves is no longer possible (Maynard Smith and Smith, 1998).

Another productive approach that was developed involved scanning the alignment using a moving window, identifying recombination boundaries as positions at which the surrounding upstream and downstream regions of the alignment implied different relationships between the taxa. At first, phylogenies were used to evaluate relationships: maximum parsimony (Fitch and Goodman, 1991; Hein, 1993), distance-based (McGuire *et al*., 1997) and maximum likelihood methods (Grassly and Holmes, 1997) were each implemented. An alternative involving comparing the distance matrices themselves along the alignment, thereby avoiding the need to construct trees, was also developed (Weiller, 1998). However, these methods often suffered from having a poor ability to detect imports, either because they did not consider branch lengths and therefore could only detect recombinations that changed the topology of the tree (Fitch and Goodman, 1991; Hein, 1993), or else struggled to distinguish imports from regions of the alignment where selection was relaxed, and hence diverged more quickly than the rest of the sequences (McGuire *et al*., 1997). Hence amendments to such algorithms were employed to improve the accuracy of import identification (McGuire and Wright, 2000).

Such approaches were subsequently adapted for Bayesian methods; rather than a moving window, a Hidden Markov Model (HMM), the states of which were phylogenies of different topologies, was used to scan the alignment for segments of sequence with different ancestries (McGuire *et al*., 2000). This was later modified to improve the distinction between heterogeneity in the rate of point mutation accumulation and imports, as for the scanning window approaches (Husmeier, 2005). These approaches were extended following the exposition of the concept of a 'clonal frame' in bacteria: that sufficiently closely related isolates would share a clonally descended fraction of their chromosomes, interrupted by a number of dispersed loci that had undergone recombination since their divergence (Milkman and Bridges, 1993). A Bayesian algorithm was developed that removed recombinant segments of sequences from the alignment, such that a final phylogeny was produced, based only on the clonal frame of each taxon (Didelot and Falush, 2007). The recombinations themselves were identified by an HMM on the basis of their elevated density of polymorphisms, rather than considering homoplasy; however, this analysis was not performed on the whole alignment, but instead independently on each branch of the

tree, using the patterns of reconstructed SNPs. This improves the sensitivity and resolution for predicting the occurrence of recombinations, and ameliorates, to an extent, the problem of distinguishing imports from loci accumulating mutations at a relatively high rate. This is because relaxed selection at a locus leads to mutations accumulating at a generally elevated rate throughout the tree, whereas in the event of a recombination, a large number of polymorphisms are introduced simultaneously. However, while this approach has been successfully applied to MLST datasets, its computational intensity means it is impractical for large, whole genome alignments.

4.2 Analysis of the PMEN1 population

4.2.1 Construction of the phylogeny

Sequence reads were mapped against the complete reference chromosome of *S. pneumoniae* ATCC 700669, identifying 39,107 polymorphic sites. Maximum likelihood analysis of these data produced a phylogeny with a high proportion of homoplasic sites (23%) and a weak correlation between the date of a strain's isolation and its distance from the root of the tree (Pearson correlation, N = 222, $R^2 = 0.05$, $p = 0.001$) (Figure 4.2). This suggested that variation was primarily arising through recombination and not through steady accumulation of base substitutions. In order to generate a phylogeny based on the clonal frame of the isolates, a maximum likelihood-based algorithm was designed to remove the recombinations from each taxon through analysis of the patterns of polymorphisms occurring on each branch, analogous to the Bayesian implementation of Didelot and Falush (Didelot and Falush, 2007). All recombinations were assumed to be imports; as PMEN1 constitutes a small fraction of the overall carried pneumococcal population, it was assumed that the rate of exchange between members of the lineage would be negligible.

Using the starting phylogeny constructed on the basis of all SNPs using RAxML (Stamatakis *et al*., 2005), the pattern of polymorphic events occurring on each branch of the tree was reconstructed using PAML (Yang, 2007). The positions of the SNPs occurring on each branch across the reference chromosome were analyzed using a one dimensional spatial scan statistic (Kulldorf, 1997) in order to detect clusters of polymorphisms that would indicate recombination events. The null hypothesis for

branch B, $H_{0,B}$, assumed the absence of any recombination events, therefore implying the SNPs occurring on the branch should be evenly distributed across the chromosome. This was considered a reasonable axiom considering the closely related nature of these isolates, as there should be minimal opportunity for selection to cause any significant spatial heterogeneity in the level of observed base substitutions. Hence $H_{0,B}$ was modelled as a binomial distribution, with SNPs uniformly distributed throughout the chromosome, of length g, occurring at a mean frequency of $d_{0,B}$, the mean number of SNPs per base, calculated separately for each branch B. This was tested using a moving window, which was altered in length, w, such that, given the number of polymorphisms occurring on the branch, the mean number of SNPs in a window, N, would be at least 10 according to $H_{0,B}$ (up to a maximum window length of 10 kb; Equation 4.1).

$$H_{0,B}: N \sim \text{Bin}(w, d_{0,B})$$

Equation 4.1

The test statistics were only calculated for the moving window at polymorphic sites, hence the threshold for significance was set as 0.05 divided by the number of SNPs occurring on the branch. Each region of the chromosome, r, where $H_{0,B}$ could be rejected at this threshold was treated as containing a recombination, and hence conforming to an alternative hypothesis, $H_{1,B,r}$, that it contained a higher density of SNPs, $d_{1,B,r}$, calculated as the mean number of SNPs within the region (Equation 4.2).

$$H_{1,B,r}: N \sim \text{Bin}(w, d_{1,B,r})$$

Equation 4.2

However, the size of these identified regions exceeded the length of the recombination they contained. In order to more precisely delineate the borders of the recombination event within the regions identified by the moving window, the block was first reduced in size such that its boundaries were the outermost SNPs within the region. Each end of the block was then progressively moved inwards until the density of SNPs within the block was more likely under $H_{1,B,r}$ than $H_{0,B}$. Once the boundaries of the putative recombination had been identified, the inequality Equation 4.3 had to

be satisfied as a final test for rejection of $H_{0,B}$ on the basis of the length of the block, b, and the number of SNPs it contained, N.

$$\frac{0.05}{g/b} > 1 - \sum_{i=0}^{i=N-1} \binom{b}{i} d_{0,B}{}^{i} (1 - d_{0,B})^{b-i}$$

Equation 4.3

This condition was required to eliminate false positive events generated as artifacts of the window length spanning the edges of neighboring, but separate, clusters of SNPs. The block identified in this manner as having the smallest likelihood ratio, calculated as the probability of the block under $H_{0,B}$ divided by its probability under $H_{1,B,r}$, was then removed from the dataset, and $d_{0,B}$ was recalculated as the mean density of SNPs across the remainder of the chromosome outside of this recombination block. The identification of recombinations was then repeated, with the process iterating until either no more loci deviated from $H_{0,B}$ or the minimum number of SNPs within a window required to identify a recombination fell below three. This approach was taken to avoid SNP-dense regions reducing the power to detect other recombinations occurring on the same branch.

The loci corresponding to these putative recombination events were then treated as missing data in all taxa downstream of the branch on which the recombination was estimated to occur when redrawing the phylogeny with RAxML. Subsequently all mutations, including those occurring within putative recombination events, were reconstructed on the new tree and recombinations re-identified as described above. This process was repeated for five iterations to produce the final dataset. The algorithm rapidly converges on a topology, as assessed by comparing the phylogenies produced by each iteration using ftreedist (Rice *et al*., 2000) (**Table 4.1**). Additionally, extending the analysis for a further four iterations resulted in few changes in the tree (**Table 4.1**), suggesting that the output of the algorithm in the case of this study is robust. The only alterations between iterations involve rearrangements concerning very short branches near the base of the tree that are difficult to resolve, with the annotated clades identified in Figure 4.1 consistently identified in all phylogenies from iteration 2 onwards.

From this analysis (Figure 4.1, Figure 4.3), a total of 57,736 single-nucleotide polymorphisms (SNPs) were reconstructed as occurring during the history of the lineage, 50,720 (88%) of which were introduced by 702 recombination events. This gives a per site *r/m* ratio (the relative likelihood that a polymorphism was introduced through recombination rather than point mutation) of 7.2, less than the previously calculated value of ~66 from MLST data (Feil *et al*., 2000). By removing recombination events from the phylogeny, the number of homoplasic sites is reduced by 97%, and the tree has significantly shortened branches, such that root-to-tip distance more strongly correlates with date of isolation ($R^2 = 0.46$, $p = <2.2 \times 10^{-16}$; Figure 4.2). The rate at which base substitutions occur outside of recombinations suggests a mutation rate of 1.57×10^{-6} substitutions per site per year (95% confidence interval 1.34 to 1.79×10^{-6}), close to the estimate of 3.3×10^{-6} substitutions per site per year from *Staph. aureus* ST239 (Harris *et al*., 2010) and much higher than that of $\sim 5 \times 10^{-9}$ substitutions per site per year found between more distantly related isolates (Ochman *et al*., 1999). Furthermore, by excluding SNPs introduced through recombinations, the date of origin of the lineage implied by the tree moved from about 1930—which predates the introduction of penicillin, chloramphenicol, and tetracycline—to about 1970 (Figure 4.2).

This method inevitably underestimates the level of recombination occurring in the population, as it only allows for the detection of imports that generate a sufficient level of sequence diversity. Hence the estimate of the *r/m* ratio is effectively a lower bound for the value. In order to quantify a probable upper bound for this parameter, it is necessary to consider how many of the SNPs identified as substitutions may actually have resulted from recombinations. It is possible that the 348 substitutions that are homoplasic with SNPs found in recombinations in the dataset may have originated through short recombinations importing a single polymorphism; this would raise the estimate of *r/m* to 7.7. In addition, if the substitution homoplasies were considered to have arisen once through point mutation, whilst the remaining instances represented horizontal transfer of this SNP, then the value of *r/m* would rise to 8.1. Finally, if all homoplasies were considered to have arisen through recombination, then the value of *r/m* would be 8.2.

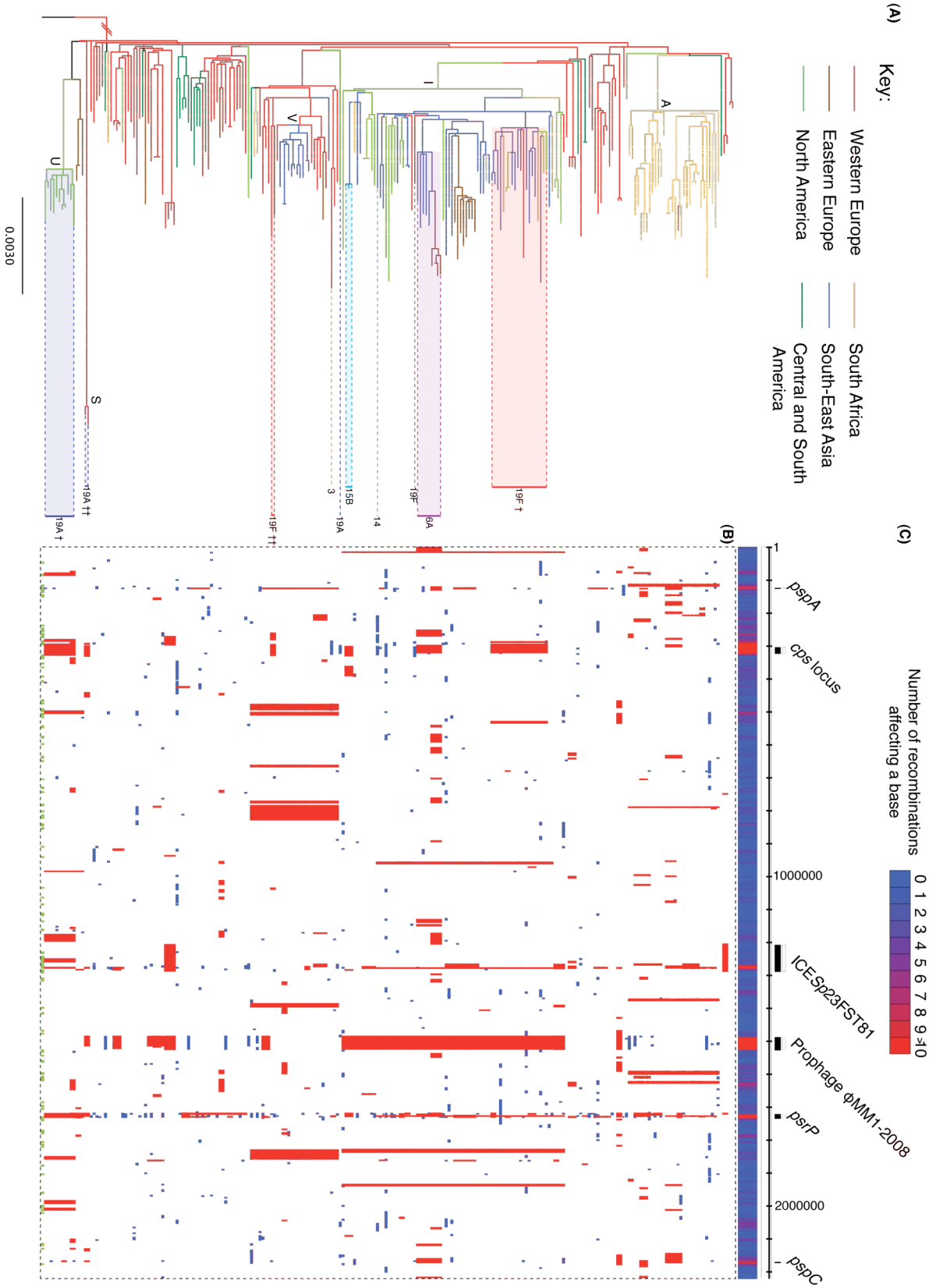

(A)
Key:
Western Europe
Eastern Europe
North America
South Africa
South-East Asia
Central and South America
0.0030
U
S
V
I
A
19A †
19A ††
19F ††
3
19A
15B
14
19F
6A
19F †
(B)
(C)
Number of recombinations affecting a base
0 1 2 3 4 5 6 7 8 9 ≥10
1
pspA
cps locus
1000000
ICESp23FST81
Prophage ϕMM1-2008
psrP
2000000
pspC

Figure 4.1 Phylogeography and sequence variation of PMEN1. (A) Global phylogeny of PMEN1. The maximum likelihood tree, constructed using substitutions outside of recombination events, is coloured according to location, as reconstructed through the phylogeny by using parsimony. Shaded boxes and dashed lines indicate isolates that have switched capsule type from the ancestral 23F serotype. Independent switches to the same serotype are distinguished by annotation with daggers (†). Specific clades referred to in the text are marked on the tree: A (South Africa), I (International), V (Vietnam), S (Spain 19A), and U (USA 19A). (B) Recombinations detected in PMEN1. The panel shows the chromosomal locations of the putative recombination events detected in each terminal taxon. Red blocks are recombinations predicted to have occurred on an internal branch and, therefore, are shared by multiple isolates through common descent. Blue blocks are recombinations predicted to occur on terminal branches and hence are present in only one strain. The green blocks indicate recombinations predicted to have occurred along the branch to the outgroup (*S. pneumoniae* BM4200), used to root the tree. (C) Biological relevance of recombination. The heat map shows the density of independent recombination events within PMEN1 in relation to the annotation of the reference genome. All regions that have undergone 10 or more recombination events are marked and annotated (Tn*916* is encompassed within ICE*Sp*23FST81).

All of these estimates remain considerably lower than the equivalent values estimated from MLST data (Feil *et al*., 2000). It should be noted that the genome-wide data demonstrate that these averages do not apply uniformly across the chromosome, but instead will vary considerably between different loci. Furthermore, it remains possible that an even higher proportion of the substitution SNPs could be the result of recombinations. Despite this underestimation of the rate at which recombination imports variation, the net rate of point mutation remains quite similar to that of *Staph. aureus* ST239, which has a much lower *r/m* value. This emphasises that recombination will overwrite base substitutions, as well as introducing variation.

4.2.2 Recombination and antigenic variation

Even in this sample of a single lineage, 74% of the reference genome length has undergone recombination in at least one isolate, with a mean of 74,097 bp of sequence affected by recombination in each strain. This encompasses both site-specific integrations of prophage and conjugative elements and homologous recombinations mediated by the competence system. The 615 recombinations outside of the prophage and ICE vary in size from 3 bp to 72,038 bp, with a mean of 6.3 kb (Figure 4.4). Within these homologous recombinations, there is a distinct heterogeneity in the density of polymorphisms, although it is unclear whether this represents a consequence of the mechanism by which horizontally acquired DNA is incorporated or a property of the donor sequence.

Recombination hotspots are evident in the genome where horizontal sequence transfers are detected abnormally frequently (Figure 4.1). One of the most noticeable is within Tn*916*, concentrated around the *tetM* gene. Excepting the prophage, the other loci—*pspA*, *pspC*, *psrP*, and the *cps* locus—are all major surface structures. Hence, it seems likely that these loci are under diversifying selection driven by the human immune system, and consequently, the apparent increase in the frequency of recombination in these regions is due to the selective advantage that is offered by the divergent sequence introduced by such recombination events.

In addition to base substitutions, 1,032 small (<6-bp) insertion and deletion events can be reconstructed onto the phylogeny, of which 61% are concentrated in the 13% of the genome that does not encode for CDSs, probably because of selection against the introduction of frameshift mutations. Throughout the phylogeny, 331 CDSs are predicted to be affected by either frameshift or premature stop codon mutations. Modeling these disruptive events as a Poisson distributed process occurring at a rate proportional to the length of the CDS, 11 CDSs were significantly enriched for disruptive mutations after correction for multiple testing (Table 4.2). These included *pspA* and a glycosyltransferase posited to act on *psrP* (SPN23F17730). This again suggests there may be a selective pressure acting either to remove (*pspA*) or alter (*psrP*) two major surface antigens. Furthermore, the longest recombination in the data set spans, and deletes, the *psrP*-encoding island, which shows that such non-essential antigens can be quickly removed from the chromosome. These data imply that the pneumococcal population is likely to be able to respond very rapidly to the introduction of some of the protein antigen–based pneumococcal vaccines currently under development.

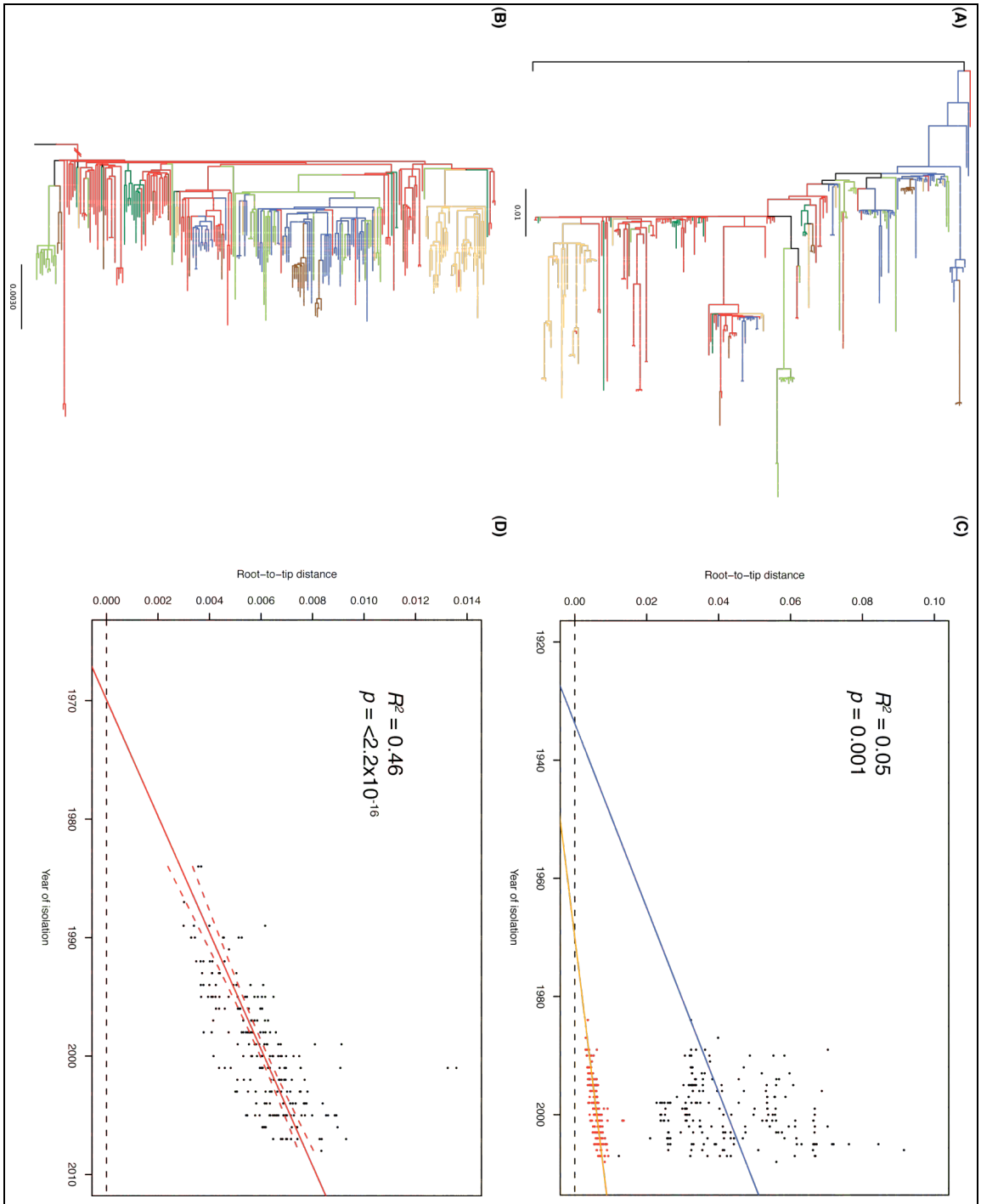


Figure 4.2 Construction of the PMEN1 phylogeny. The maximum likelihood phylogeny constructed on the basis of all SNPs, coloured according to geographical distribution as in Figure 4.1, is shown in (A), and that derived by excluding those SNPs falling within putative recombination events is shown in (B) for comparison. A plot of root-to-tip distance against date of isolation, for taxa for which dates were available, is shown in (C): points in black correspond to tree (A), and those in red correspond to tree (B), with the regression lines coloured blue and orange, respectively. In (D), the points corresponding to tree (B) are shown in greater detail, with the 95% confidence interval indicated by the dashed red lines. The two outlying points correspond to the Spanish 19A isolates (clade 'S' in Fig. 1), which lie on a long branch that indicates they may have accumulated mutations at an unusually high rate at some point in their recent history. This graph suggests the PMEN1 lineage originated around 1970.

4.2.3 Population and serotype dynamics

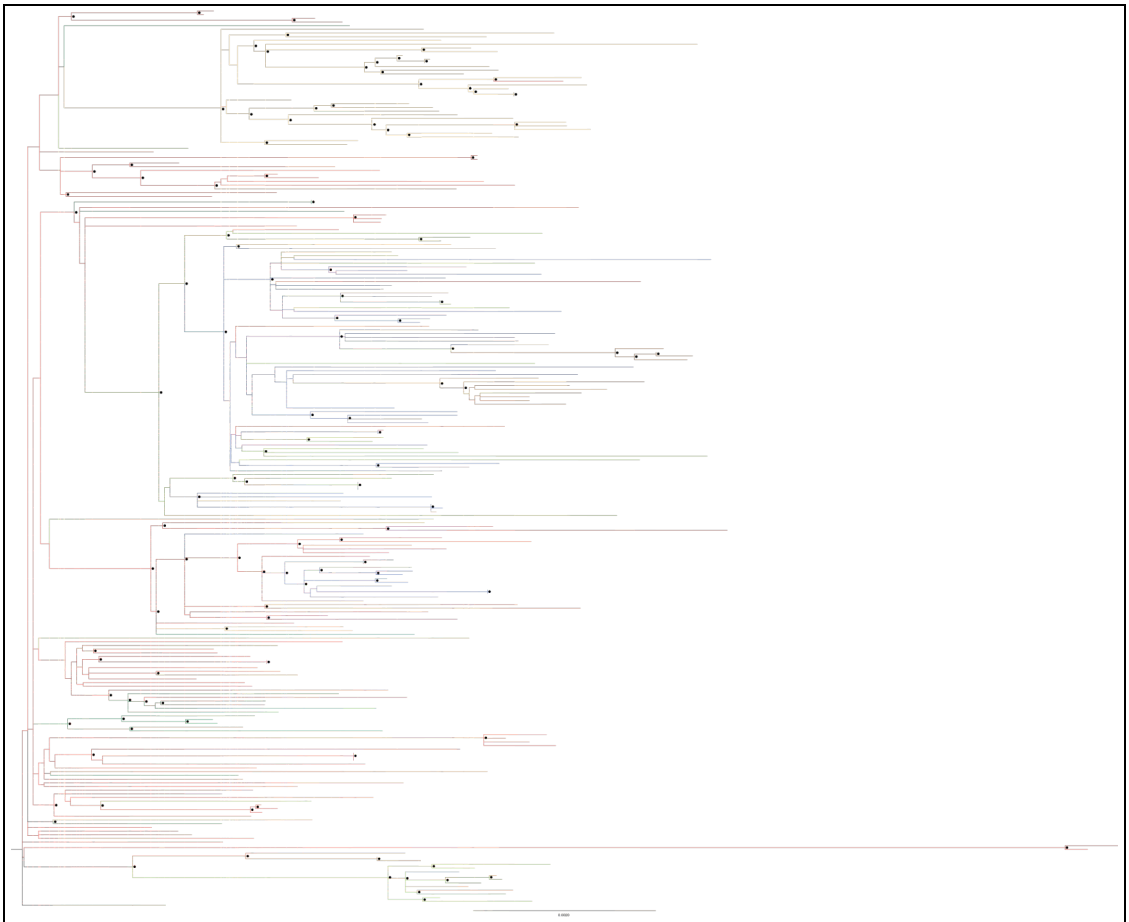


Figure 4.3 Robustness of the PMEN1 phylogeny. All nodes in the tree with bootstrap values greater than, or equal to, 75 are marked with a black dot (•).

The spread of PMEN1 can be tracked using the phylogeography indicated by the tree (Figure 4.1). There are several European clades with their base near the root of the tree, and a parsimony-based reconstruction of location supports a European origin for the lineage. Interspersed among the European isolates are samples from Central and South America, which may represent an early transmission from Spain, where the clone was first isolated, to Latin America, a route previously suggested to occur by data from *Staph. aureus* (Harris *et al*., 2010). One clade (labelled A in Figure 4.1), containing South African isolates from 1989 to 2006, appears to have originated from a single highly successful intercontinental transmission event. There is also a cluster of isolates from Ho Chi Minh City (labeled V), representing a transmission to Southeast (SE) Asia. However, the predominant clade found outside of Europe (labelled I) appears to have spread quite freely throughout North America, SE Asia,

and Eastern Europe, which implies that there are few barriers to intercontinental transmission of *S. pneumoniae* between these regions.

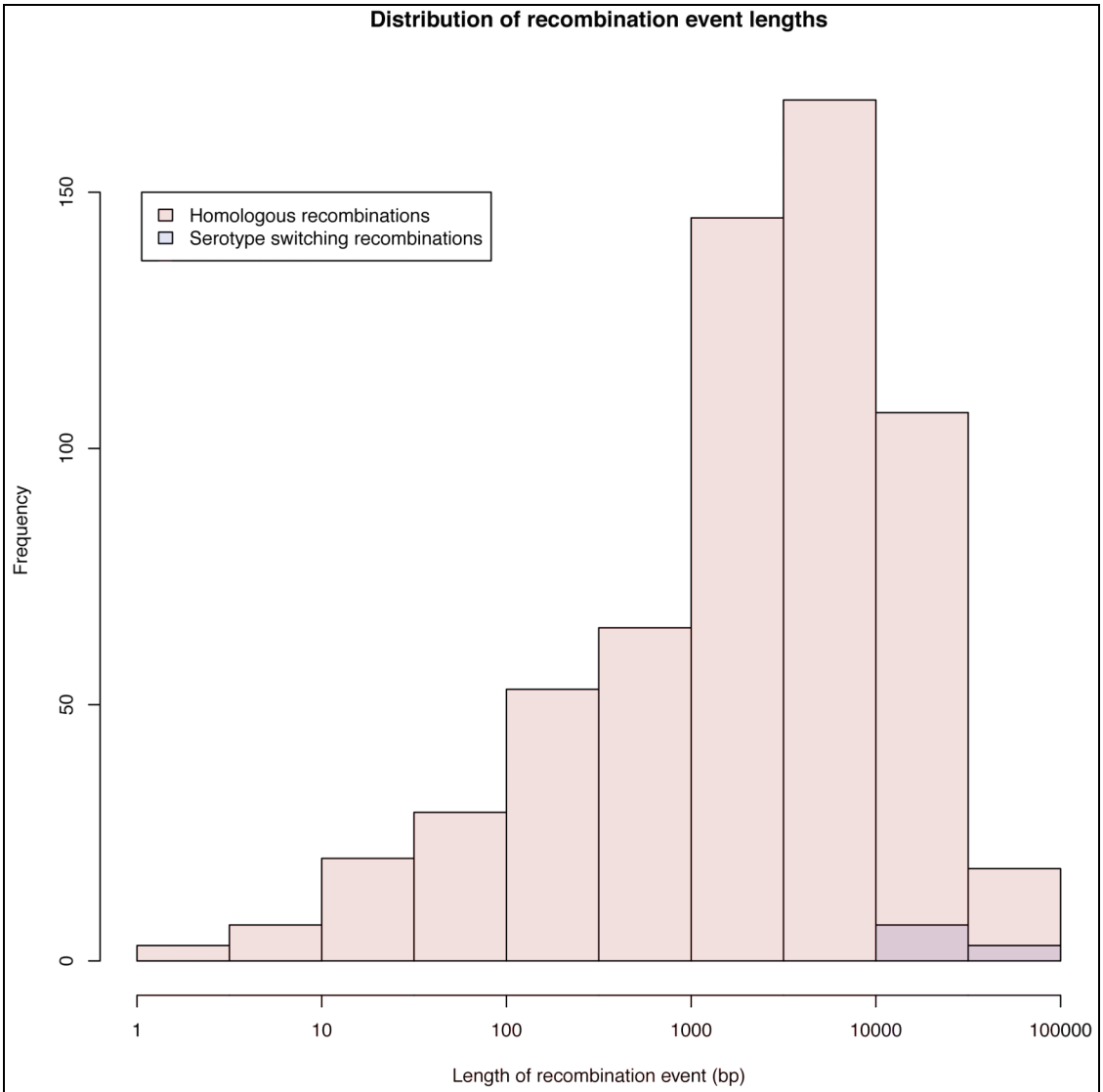


Figure 4.4 Histogram showing the variation of recombination event lengths. The bars in red show the distribution of lengths for the 615 homologous recombinations occurring outside of ICE*Sp*23FST81 and the ϕMM1-2008 prophage on a $\log_{10}$ scale. The bars in blue represent the ten recombinations causing serotype switching. This clearly shows that, although such events are within the range of homologous recombination lengths found elsewhere in the chromosome, they are atypically large. Hence it is unusual that serotype switches are observed so frequently within the PMEN1 population.

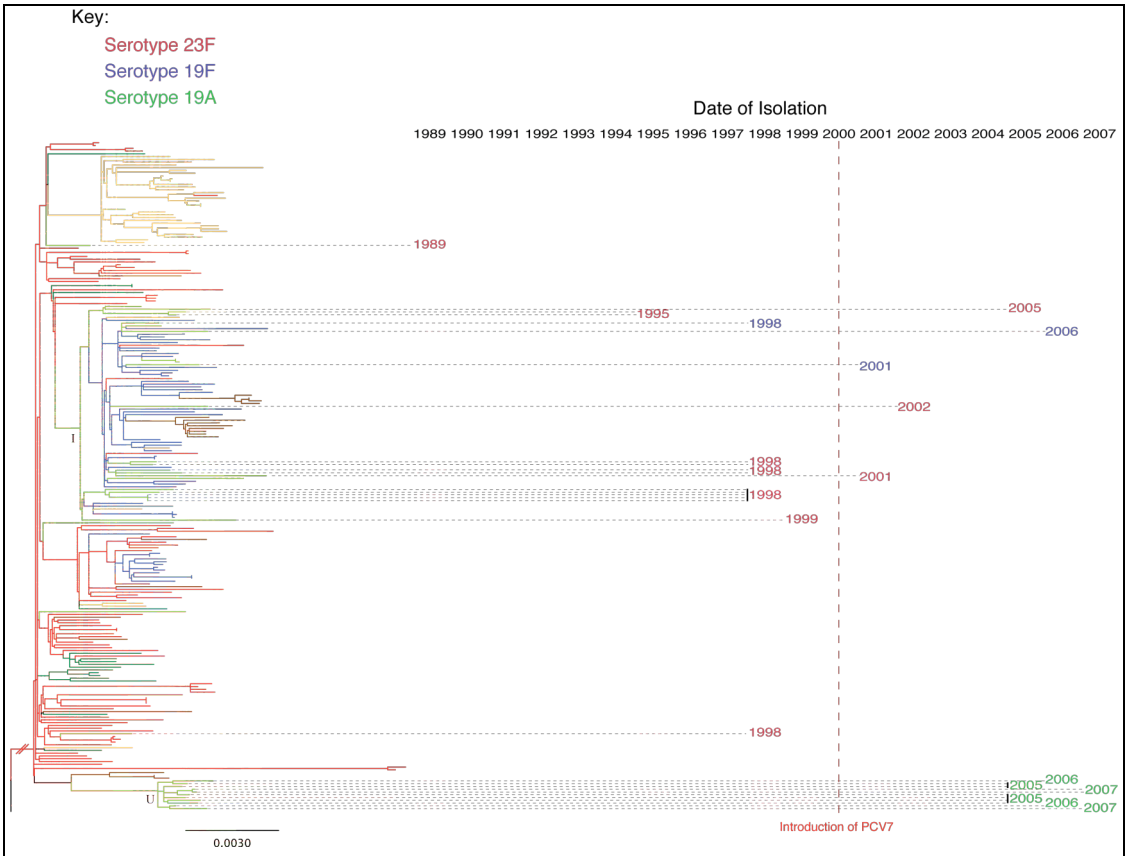


Figure 4.5 Population dynamics of PMEN1 in the USA. All PMEN1 isolates in the phylogeny from the United States, where the PCV7 conjugate polysaccharide vaccine was introduced in 2000, are labelled with their year of isolation and coloured according to their serotype. The earliest isolate, from Cleveland in 1989, appears to represent an initial, limited transmission event. The more successful colonization seems to have come from clade I, which contains all but one of the isolates from the 1990s. This group includes US isolates of serotypes 23F and 19F, both of which correspond to antigens present in PCV7. Hence there is a shift in the population to the serotype 19A isolates (clade U), not targeted by the vaccine, in the later years of this study.

The final non-European group consists of serotype 19A U.S. isolates (labelled U). These all date from between 2005 and 2007 and are distinct from all other U.S. PMEN1 isolates, which have capsular types included in PCV7 (Figure 4.5). This is evidence of a shift in the PMEN1 population in the USA: rather than a change in capsule type occurring among the resident population, it has been eliminated by the vaccine and replaced by a different subpopulation within the lineage that has expanded to fill the vacated niche. Similarly, a pair of Spanish isolates from 2001 (labeled S in Figure 4.1), the year in which PCV7 was introduced in Spain, that have independently acquired a 19A capsule are not closely associated with any other European isolates. The estimated times of origin for clades U (1996; 95% credible interval 1992–1999) and S (1998; 95% credible interval 1996–1999) both predate the introduction of PCV7, and accordingly a third 19A switch, from Canada, was isolated

in 1994. Hence, it appears that these changes in serotype after vaccine introduction result from an expansion of pre-existing capsular variants, which were relatively uncommon and not part of the predominant population, and would have therefore been difficult to detect before the existence of the selection pressure exerted by the vaccine.

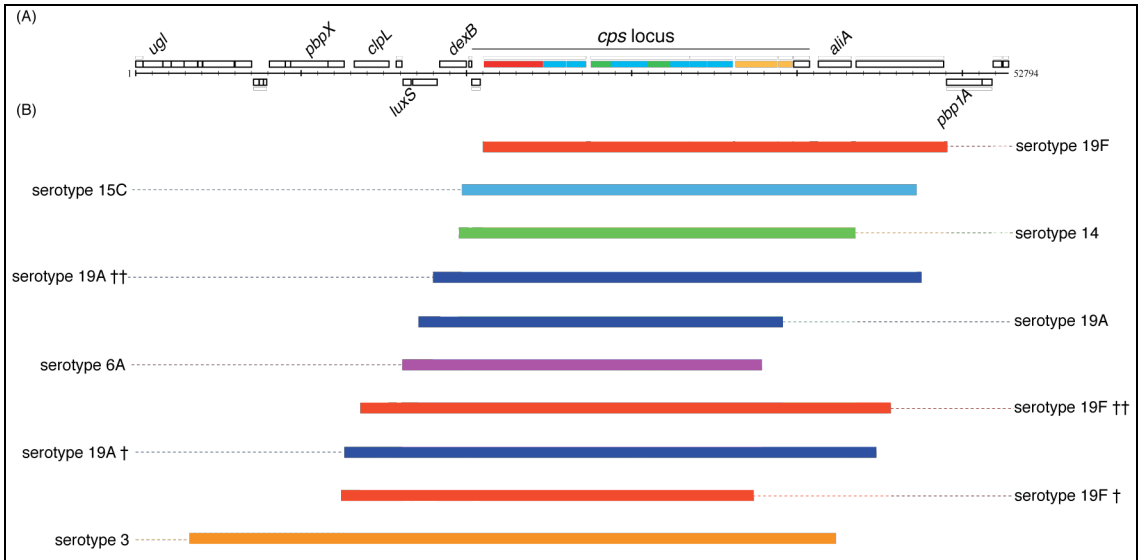


Figure 4.6 Recombinations causing serotype-switching events. (A) The annotated *cps* locus of the reference strain. CDSs involved in capsule biosynthesis are coloured according to their role. Genes in red are regulatory, those in blue synthesize and modify the oligosaccharide subunits, those in green are involved in polymerization and transport, and those in orange are required for the synthesis of rhamnose. (B) Below are delineated the recombinations leading to changes in serotype, coloured according to the serotype of the sequence donor. The different events are labelled to correspond with Figure 4.1.

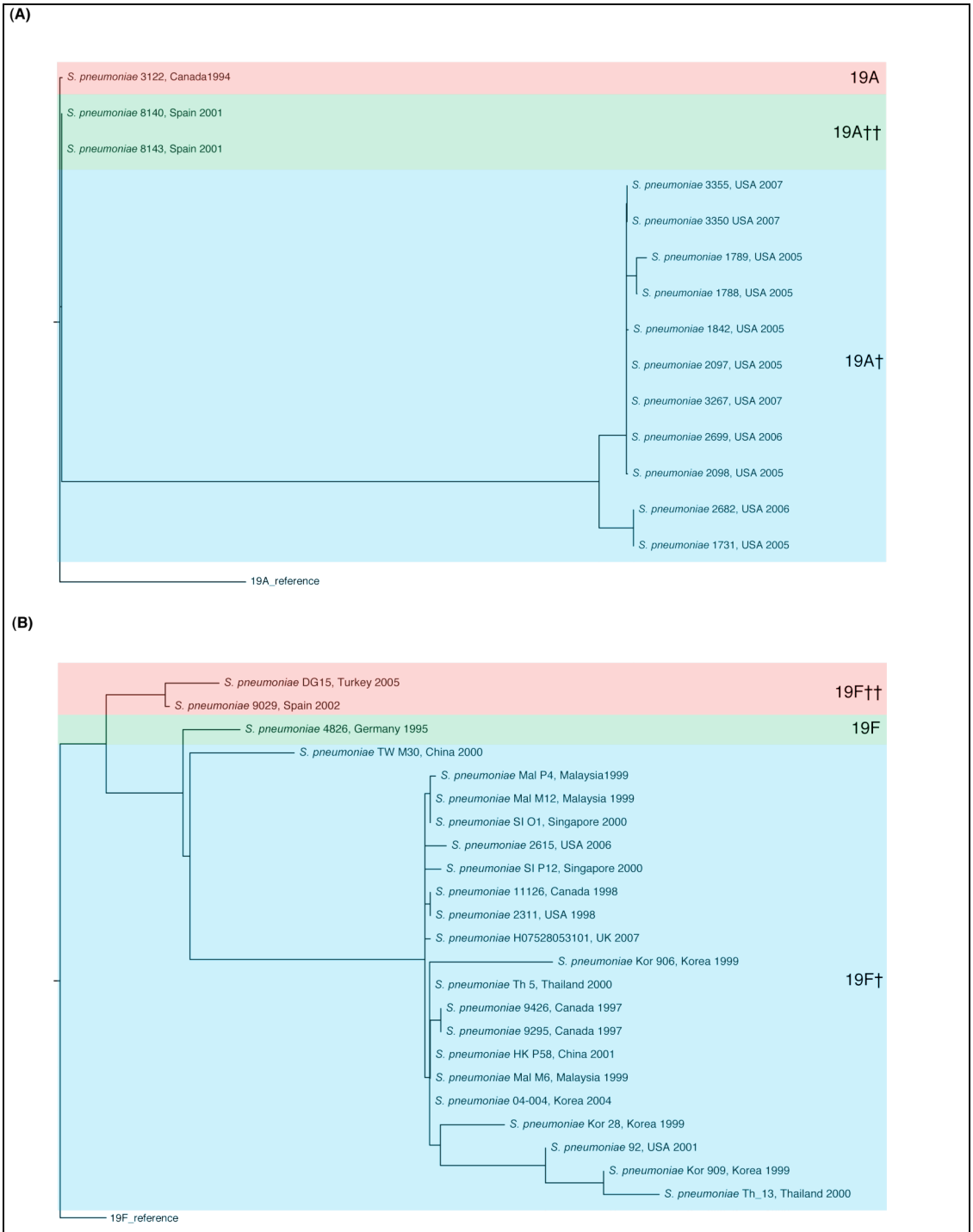


Figure 4.7 Variation in the serogroup 19 *cps* loci. (A) The *cps* loci of the serotype 19A isolates were assembled *de novo*, aligned and a maximum likelihood tree generated. The three acquisition events are coloured differently and marked according to the scheme in Figure 4.1. This shows the *cps* locus acquired by the US isolates is quite distinct from that received by the Spanish or Canadian isolates. There is variation in this region even between the US isolates: *S. pneumoniae* 1731 and 2682 appear to have a second recombination affecting the *wzg* and *wzh* genes at the 5' end of the *cps* locus. (B) The *cps* loci of the serotype 19F isolates were analyzed in the same manner. Here, *S. pneumoniae* TW M30 differs from the other isolates with which it shares its 19A *cps* locus through common descent due to a secondary recombination affecting the *rmlABCD* gene cluster.

Seven further serotype-switching events can be detected in the data (Figure 4.6), including three switches to serotype 19F. The polyphyletic nature of these 19F isolates is supported by the variation observed between the acquired *cps* loci, as is also the case for the 19A isolates (Figure 4.7). The previously known switches to serotypes 3, 6A, and 15B are only found to occur once each in the phylogeny, and in addition, a single Korean sample that had not been typed was identified as a serotype 14 variant by mapping reads to known *cps* loci (Bentley *et al*., 2006). The recombination events leading to these switches ranged from 21,780 bp to 39,182 bp in size, with a mean of 28.2 kb. Only 35 homologous recombinations of an equivalent size or larger occur elsewhere in the genome; most such events are much smaller (Figure 4.4), which makes it surprising that serotype switching occurs with such frequency, indicating a role for balancing selection at this locus. Additionally, the span of these events appears to be limited by the flanking penicillin-binding protein genes, the sequences of which are crucial in determining β-lactam resistance in pneumococci (Trzcinski *et al*., 2004). Only the recombination causing the switch to serotype 3 affects one of these, and it introduces just a single SNP into the *pbpX* CDS, which does not appear to compromise the strain's penicillin resistance (Appendix II: PMEN1 strains). Hence, the positioning of these two genes may hinder the transfer of capsule biosynthesis operons from penicillin-sensitive to penicillin-resistant pneumococci via larger recombinations, although size constraints alone could also cause such a distribution.

4.2.4 Resistance to non–β-lactam antibiotics

The strong selection pressures exerted by antibiotics on the PMEN1 lineage are manifest as multiple examples of geographically disparate isolates converging on common resistance mechanisms. Single base substitutions causing reduced susceptibility to some classes of antibiotics have occurred multiple times throughout the phylogeny, as observed in *Staph. aureus* (Harris *et al*., 2010) and *Salmonella* Typhi (Holt *et al*., 2008) populations, including mutations in *parC*, *parE*, and *gyrA*, which cause increased resistance to fluoroquinolone antibiotics (Pletz *et al*., 2004), and changes in *rpoB* causing resistance to rifampicin (Ferrandiz *et al*., 2005). The S79F, S79Y, and D83N mutations in *parC* are estimated to occur nine, three, and five times, respectively, in PMEN1; additionally, D435N in the adjacent *parE* gene is

found to happen three times. The S81F and S81Y substitutions, in the same position of *gyrA*, are found four and two times, respectively. None of these mutations are predicted to have been introduced by recombination, whereas changes at position H499 of *rpoB* causing rifampicin resistance are introduced twice by horizontal transfer and three times by means of base substitution.

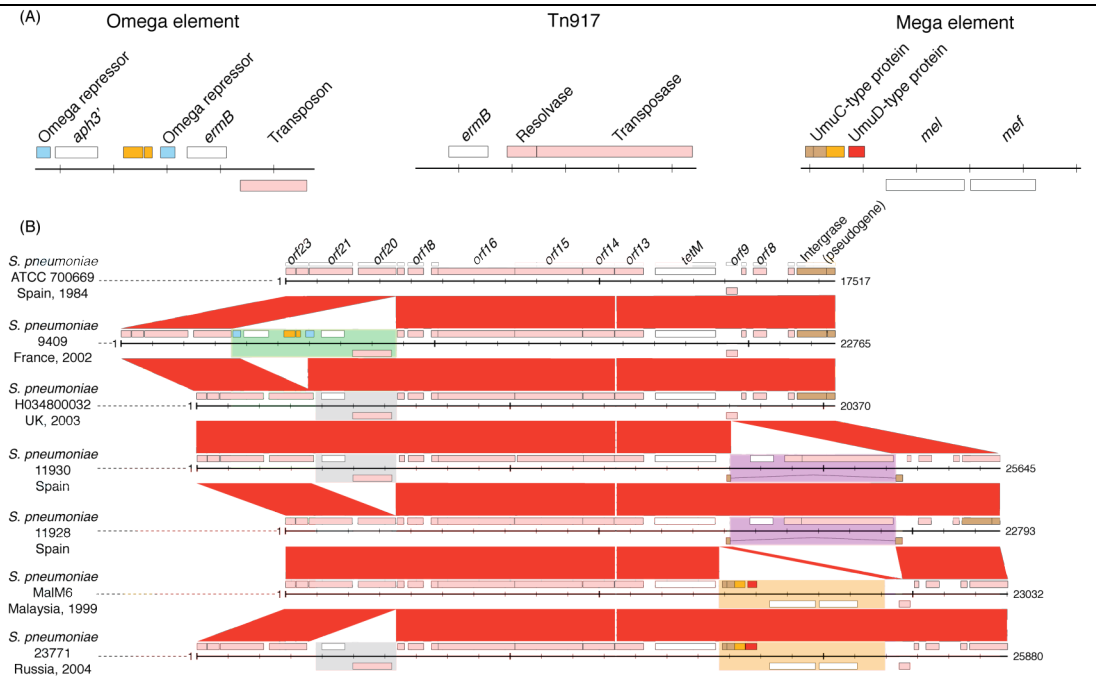


Figure 4.8 Acquisition of macrolide-resistance cassettes. The three full-length resistance cassettes are shown in (A): the omega element, which carries an *aph3′* aminoglycoside-resistance gene and an *ermB* macrolide-resistance gene; Tn*917*, which carries just the *ermB* methylase; and the mega element, which carries the *mel*/*mef* macrolide efflux system. (B) A comparison of the different Tn*916* variants in the PMEN1 lineage. Red bands between the sequences indicate BLASTN matches. The omega element is shaded green when present at full length and shaded grey where present as a remnant resulting from a recombination between the omega repressor–encoding genes, which concomitantly leads to the fusion of an omega transcriptional repressor domain to the 3′ of the *orf20* CDS. Tn*917* is boxed in purple, with the two parts of *orf9*, into which it inserts, indicated on either side. The mega element is boxed in orange.

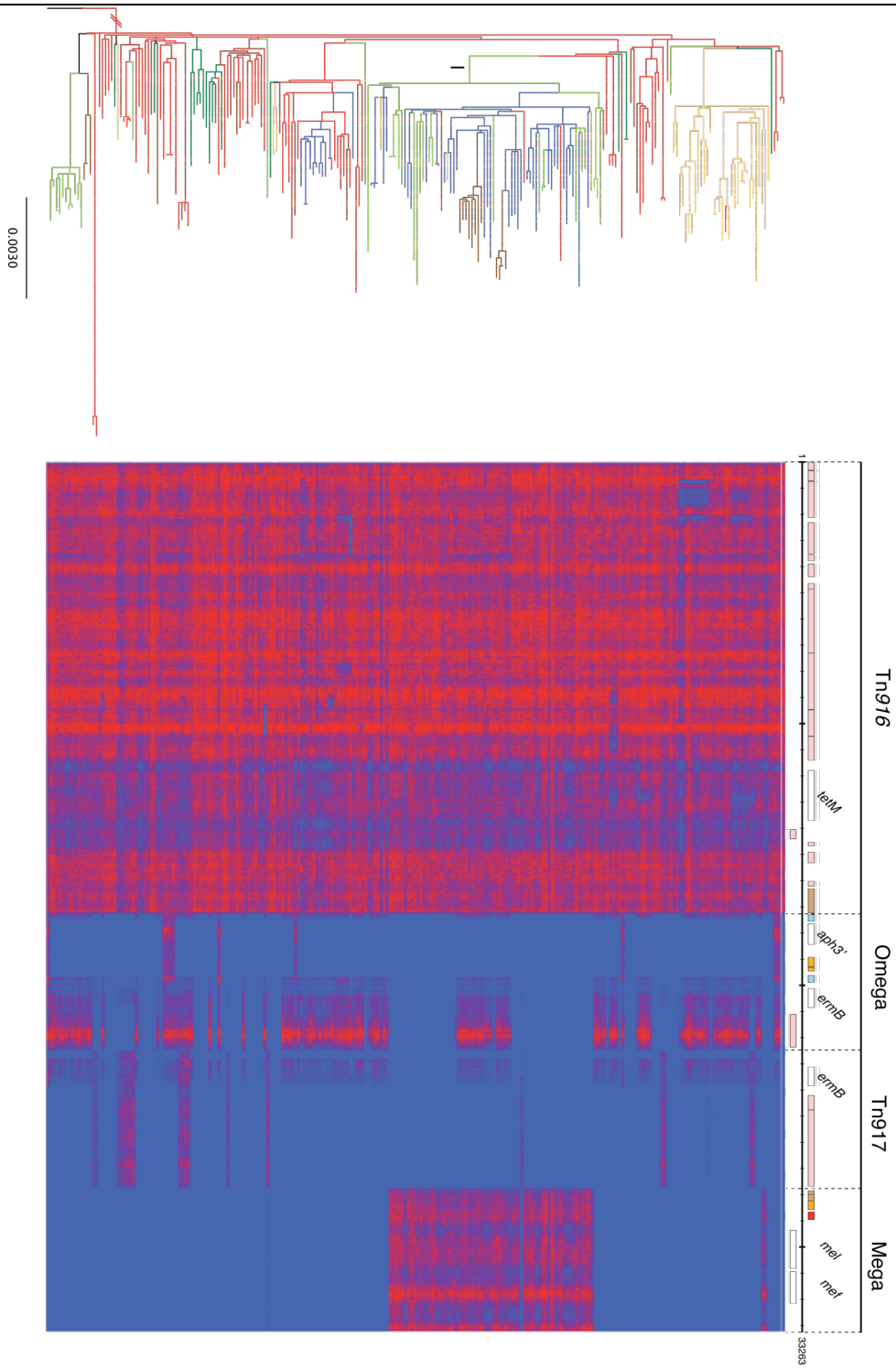


Figure 4.9 Distribution of macrolide resistance cassettes amongst the PMEN1 isolates in relation to the phylogeny, on which the 'International' clade is marked with an 'I'. The reference Tn*916* annotation is shown at the top, along with the Omega, Tn*917* and Mega elements. The level of sequence read mapping from each isolate in the phylogeny is displayed as a heatmap, with blue indicating low mapping and red indicating high coverage. Many isolates only have coverage of the 3' half of the Omega element, because the 5' end has been deleted through a recombination between the Omega repressor encoding genes (Figure 4.8). There is sufficient similarity between the *ermB* nucleotide sequences in the Omega and Tn*917* elements to allow for cross mapping of Illumina reads between them.

Resistance to macrolide antibiotics tends not to derive from SNPs, but from acquisition of CDSs facilitating one of the two common resistance mechanisms: methylation of the target ribosomal RNA by *erm* genes and removal of the drug from the cell by the macrolide efflux (*mef*)-type efflux pumps. Both can be found in the PMEN1 population, and in all cases, the genes appear to be integrated into the Tn*916* transposon (Figure 4.8). They are carried by three different elements. Tn*917*, consisting of an *ermB* gene with an associated transposon and resolvase, inserts into open reading frame *orf9* of Tn*916* (Shaw and Clewell, 1985). A second has been characterized as the macrolide efflux genetic assembly (mega) element (Del Grosso *et al*., 2006), which carries a *mef*/*mel* efflux pump system and, in PMEN1, inserts upstream of *orf9*. A third element (henceforth referred to as an omega element, for omega and multidrug-resistance encoding genetic assembly) carries both an *ermB* gene and an aminoglycoside phosphotransferase, with the latter flanked by direct repeats of omega transcriptional repressor genes, and is found just downstream of *orf20*.

Rather than a single acquisition of these elements occurring, and the resulting clones spreading and replacing macrolide-sensitive isolates, all three elements appear to have been acquired multiple times across the phylogeny (Figure 4.9). The mega element is predominantly shared by isolates in clade I, although the *ermB*-encoding omega element appears to have been subsequently acquired on two occasions, and Tn*917* has entirely superseded the mega element in one isolate. This is congruent with the known advantages of target methylation over drug efflux as a broader-spectrum resistance mechanism (Del Grosso *et al*., 2007). In most instances of the omega element, only the *ermB*-encoding part remains; the aminoglycoside phosphotransferase appears to have been deleted through a recombination between the omega-encoding genes, which leaves only an omega domain–encoding open reading frame fused to *orf20* as a scar. This implies that the benefit of the aminoglycoside-resistance element may have not been sufficient to maintain it on the ICE.

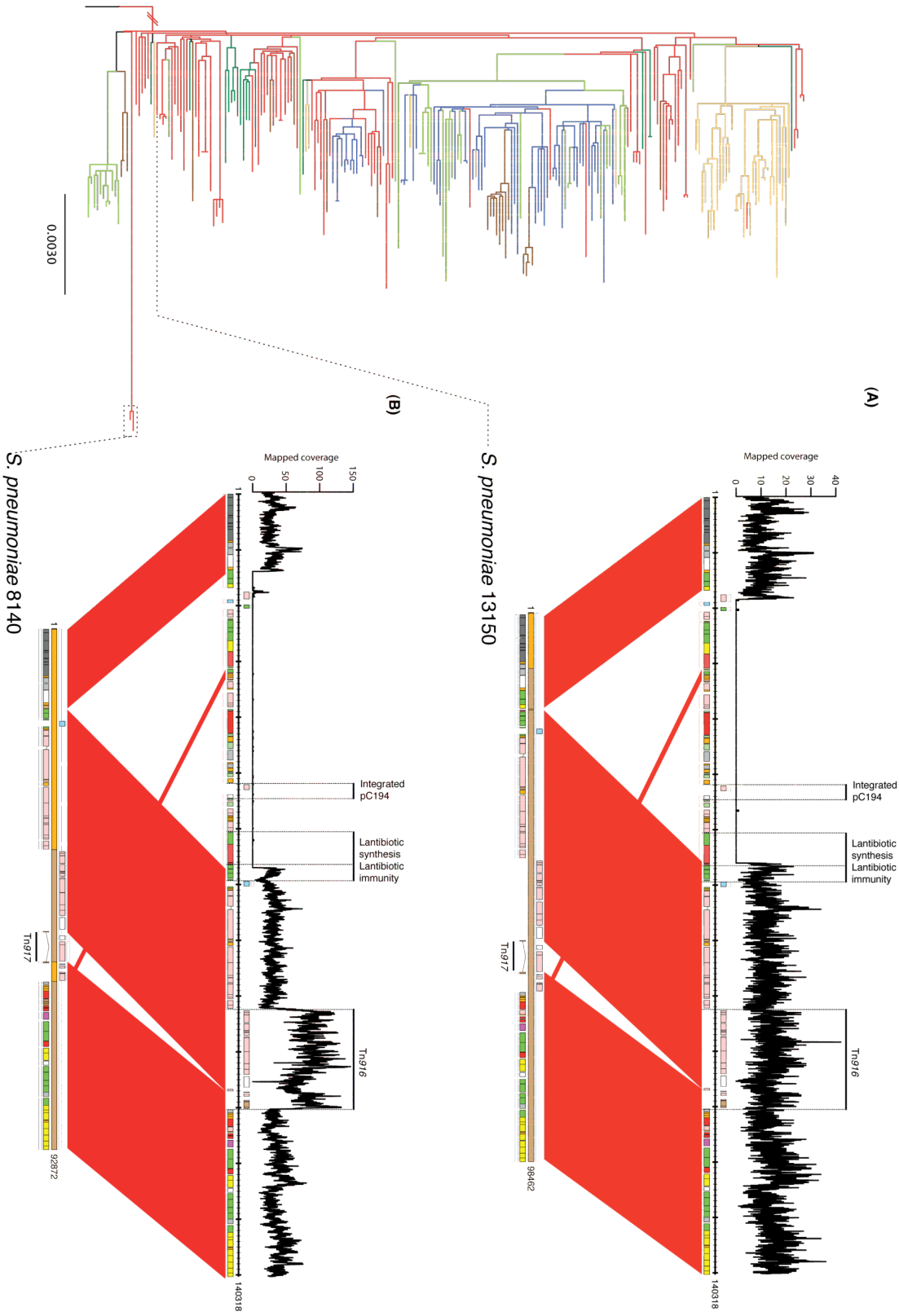

0.0030
(A)
S. pneumoniae 13150
Mapped coverage
0
10
20
30
40
Integrated pC194
Lantibiotic synthesis
Lantibiotic immunity
Tn916
Tn917
1
98462
140318
(B)
S. pneumoniae 8140
Mapped coverage
0
50
100
150
Integrated pC194
Lantibiotic synthesis
Lantibiotic immunity
Tn916
Tn917
1
92872
140318

Figure 4.10 Deletions affecting ICE*Sp*23FST81. Two deletions are displayed; in both cases, *de novo* assemblies are shown relative to the reference annotation, with red bands between the sequences indicating BLASTN matches. Alternate brown and orange boxes covering the scale line indicate scaffolds produced by assembly of Illumina reads, which have been ordered against the reference sequence. A graph illustrating the depth of Illumina read mapping to the reference sequence is shown above the reference annotation to demonstrate that the deletions do not represent deficiencies in the short read assemblies. (A) *S. pneumoniae* 13150, isolated in Germany in 1998, lacks the 5' end of the ICE, with the boundary of the deletion meaning the genes for the biosynthesis of the lantibiotic carried by the ICE are lost, but the genes encoding the ABC transporter presumed to be required for self-immunity are retained. The integrated pC194 plasmid that carries the chloramphenicol acetyltransferase gene of ICE*Sp*23FST81 is also lost. (B) *S. pneumoniae* 8140, isolated in Spain in 2001, has a similar deletion, starting slightly upstream but again ending in a position that retains most of the self-immunity transporter genes but removes the lantibiotic synthesis CDSs and the chloramphenicol acetyltransferase on pC194. The Tn*916* transposon of this strain has a depth of sequence read mapping approximately twice that of the rest of the locus, suggesting that a second copy of this element has been acquired somewhere in the chromosome.

4.2.5 Components of the accessory genome

Other than the insertion of these cassettes, the ICE itself is otherwise relatively unchanged throughout the population. In two cases, the 5′ region of the element up to, and including, the lantibiotic synthesis machinery is deleted, whereas the self-immunity genes are retained (Figure 4.10). This deletion, which also removes the integrated chloramphenicol-resistance plasmid, is analogous to that observed in the PPI–1 of the PMEN1 lineage, in which all that remains are the immunity genes from a once-intact lantibiotic synthesis machinery. In two other cases, the ICE has been supplanted by alternative transposons, both of which are similar composites of Tn*5252*- and Tn*916*-type elements: In *S. pneumoniae* 11876, a wholesale replacement at the same locus entails the gain of an omega element at the expense of losing resistance to chloramphenicol (Figure 4.11), whereas, in isolate 11930, the new ICE inserts elsewhere in the chromosome and carries two *ermB* genes, as well as a chloramphenicol acetyltransferase (Figure 4.12). The only other identified conjugative element was an ICE*St*1-type transposon shared by isolates 8140 and 8143 (Figure 4.13), and the only extrachromosomal element present in the data set was the plasmid pSpnP1 (Romero *et al*., 2007), found in isolate SA8.

The accessory genome is primarily composed of prophage sequence (Figure 4.14), with little evidence of much variation in the complement of metabolic genes. Viral sequences appear to be a transient feature of the pneumococcal chromosome (Figure 4.15), with few persisting long enough to be detected in related isolates. Four of the new prophage that could be assembled were found to insert into the competence pilus structural gene *comYC*, which lies within an operon shown to be essential for competence in *S. pneumoniae* (Pestova and Morrison, 1998). In two cases where such phage appear to be shared through common descent by pairs of isolates, no recombination events can be detected that are unique to either member of the pair, consistent with a nonfunctional competence system in these isolates. Furthermore, assaying the competence of available lysogenic strains in vitro also suggested that these phage insertions abrogate the ability of their host to take up exogenous DNA (Figure 4.16).

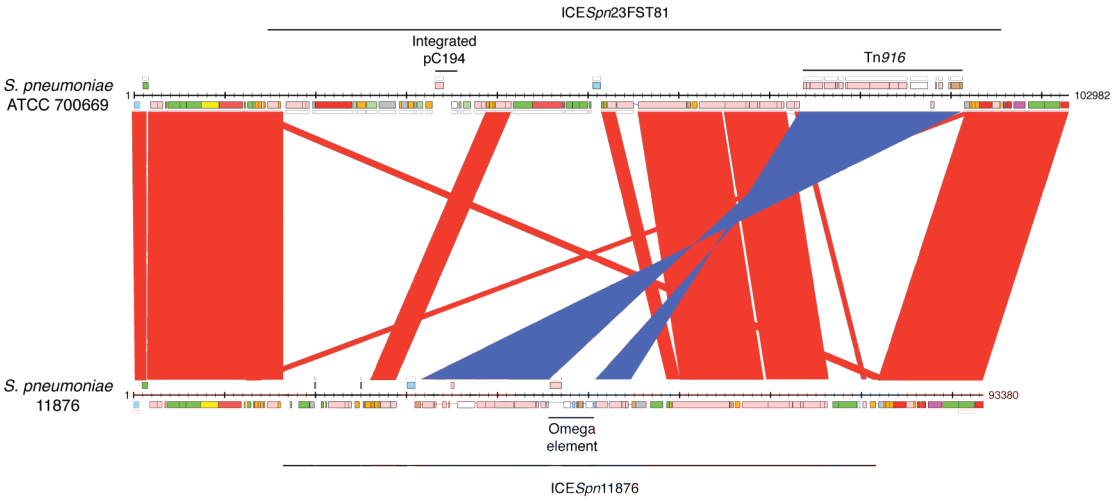


Figure 4.11 The putative ICE ICE*Sp*11876 displayed relative to ICE*Sp*23FST81. Red bands indicate BLASTN matches between the sequences in the same direction, blue bands indicate BLASTN matches between the sequences in the inverse orientation. ICE*Sp*11876 is inserted into the same locus as ICE*Sp*23FST81, shares the majority of the conjugative machinery and also carries a Tn*916* element, albeit at a different position and in the opposite orientation. This Tn*916* element carries an Omega element (Figure 4.8), and has inserted in a metalloprotease gene, which itself is part of a four gene island similar to one found in *S. equi*. The majority of the other ICE*Sp*11876 cargo genes are regulators or transporters, and it lacks the linearised pC194 plasmid that carries the chloramphenicol resistance gene found in ICE*Sp*23FST81.

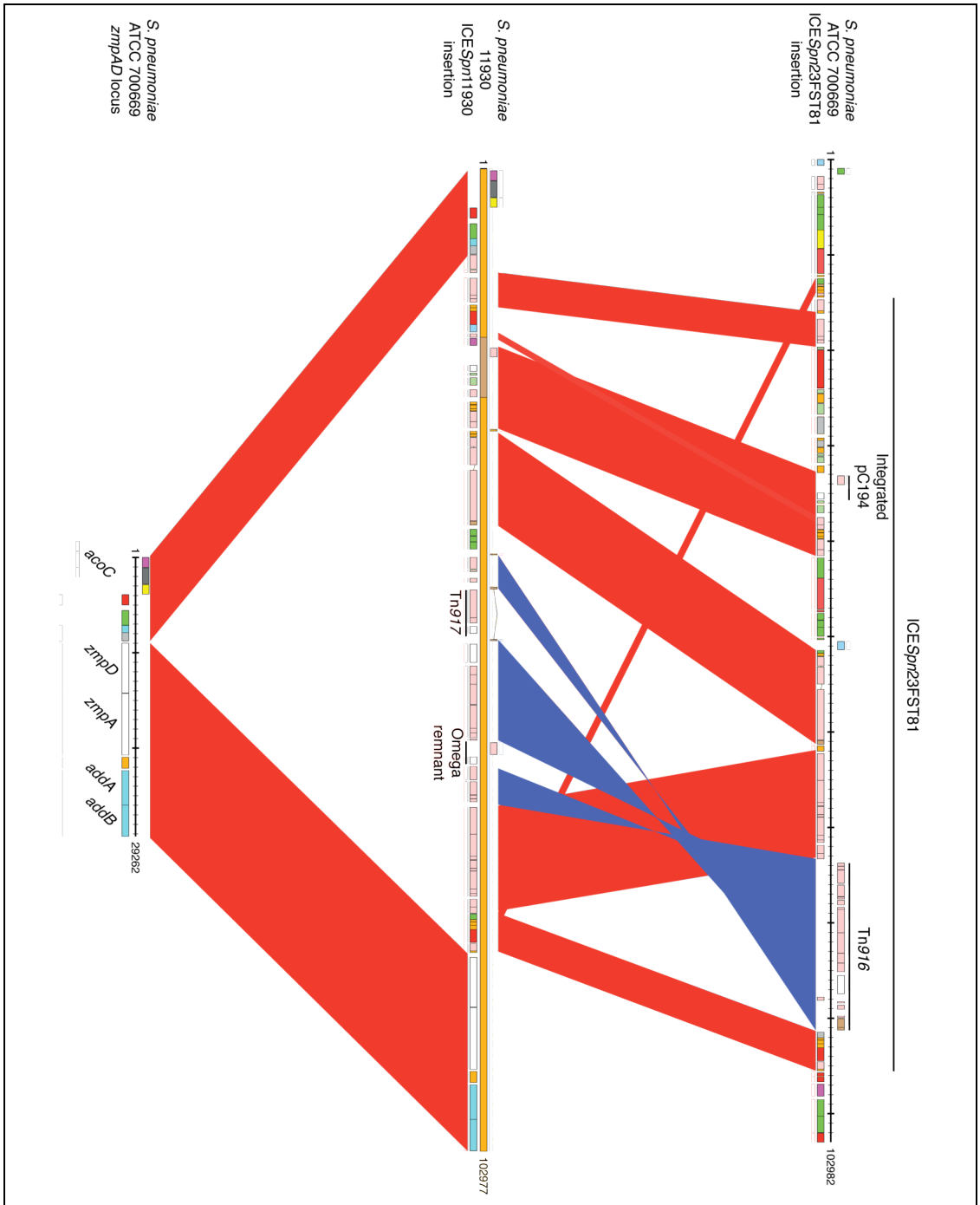


Figure 4.12 The putative ICE ICE*Sp*11930 displayed relative to ICE*Sp*23FST81, as described in Figure 4.11. Alternate brown and orange boxes covering the scale line indicate scaffolds produced by assembly of pyrosequencing and Illumina reads, which have been ordered against the reference sequence. This ICE has inserted downstream of the *zmpAD* genes, in a different position to ICE*Sp*23FST81, which is deleted in this strain. The majority of the conjugative machinery is shared between the ICE and ICE*Sp*11930, and both carry Tn*916*, albeit in different positions and in opposite orientations. The Tn*916* component of ICE*Sp*11930 carries two *ermB* genes, as the Tn*917* and Omega cassettes are both inserted into this transposon. Unlike ICE*Sp*11876 (Figure 4.11), ICE*Sp*11930 appears to carry the linearised pC194 plasmid, which contains the chloramphenicol acetyltransferase gene.

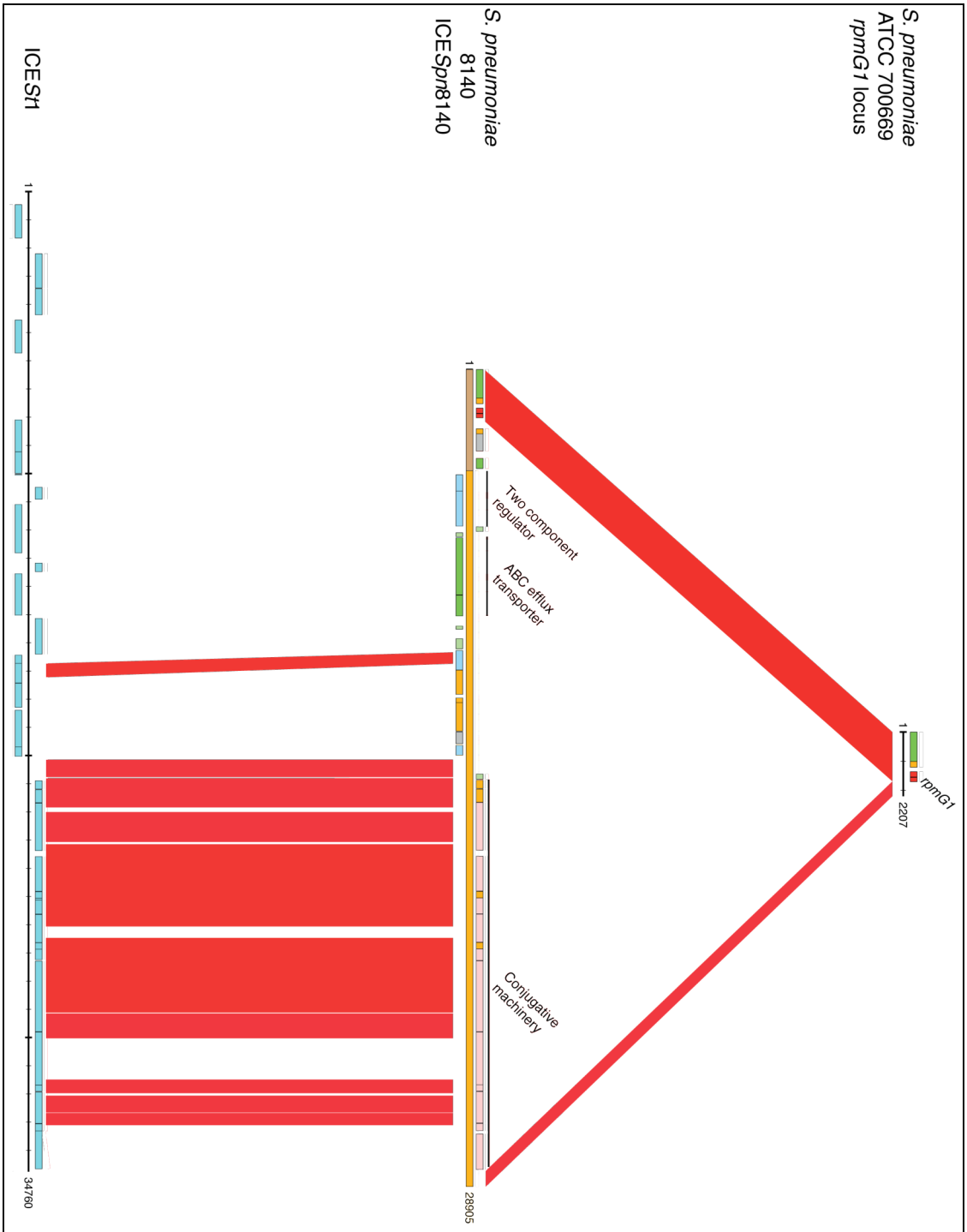


Figure 4.13 The putative ICE*St*1-like element ICE*Sp*8140. The two ends of this small conjugative element were assembled *de novo* from Illumina reads. The brown and orange boxes covering the scale line indicate scaffolds produced by assembly of Illumina reads, which have been ordered against the reference sequence. Similarity between the sequences is indicated as described in Figure 4.11. The element inserts shortly downstream of *rpmG1* and shares the conjugative machinery of ICE*St*1. However, the two elements have very dissimilar cargo complements, with ICE*Sp*8140 carrying a putative two component regulator and ABC transporter.

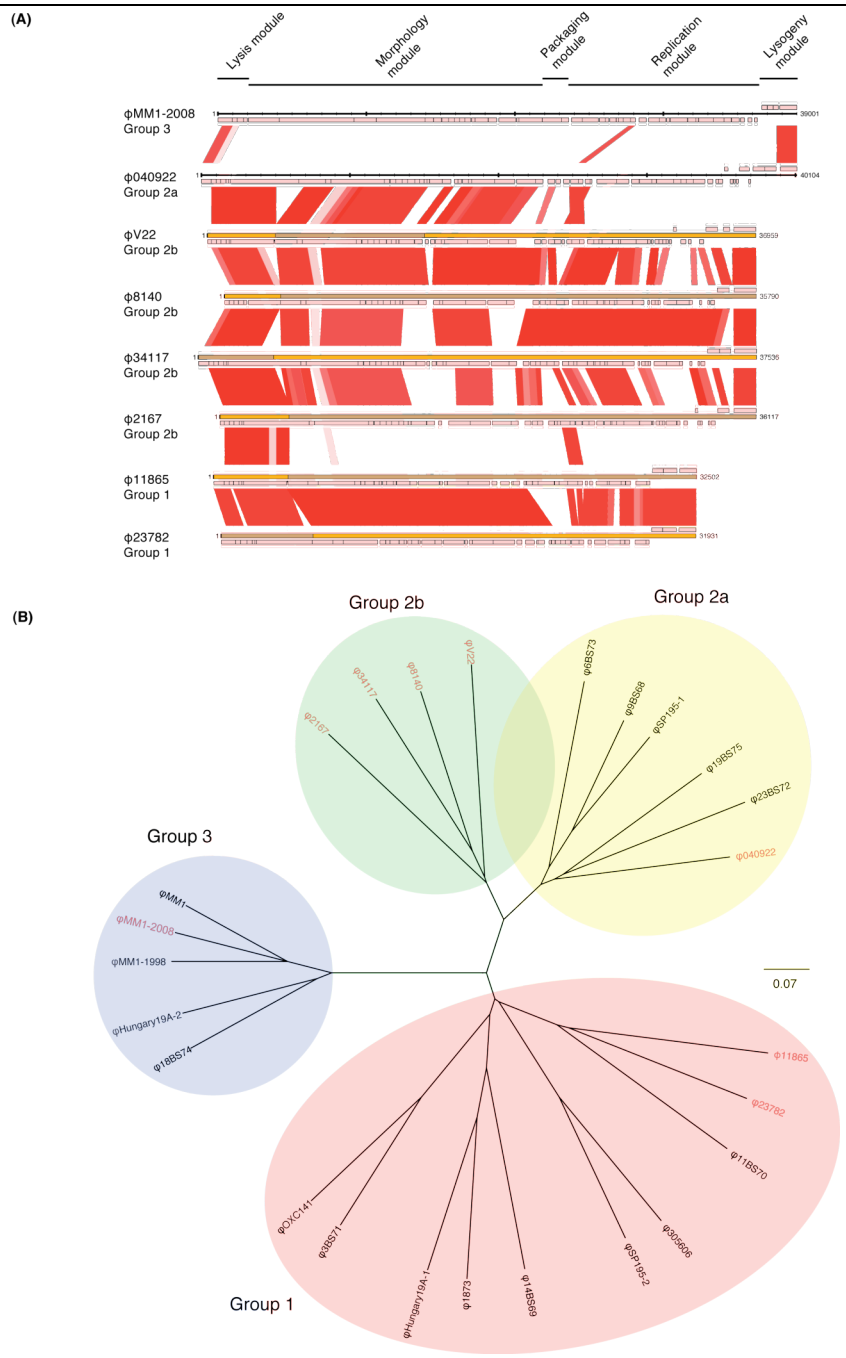


Figure 4.14 Prophage sequences in PMEN1. (A) Comparison of seven prophage assembled from PMEN1 isolates to prophage ϕMM1-2008 from the reference genome, *S. pneumoniae* ATCC 700669. The modular structure of such pneumococcal prophage genomes are indicated along the top. Alternate orange and brown boxes covering the scale line indicate scaffolds produced by assembly of Illumina reads. Red bands between sequences indicate TBLASTX matches between prophage in the same orientation, with the colour intensity reflecting the strength of the sequence similarity. (B) Clustering of pneumococcal prophage. All available identified prophage from *S. pneumoniae* genomes were included in this analysis. The names of prophage assembled from PMEN1 isolates are coloured red, showing that representatives from every group of temperate pneumococcal phage can be identified in the PMEN1 lineage. In addition, the previously unidentified group 2b prophage that insert into the *comYC* gene can be found in the isolates samples in this study.

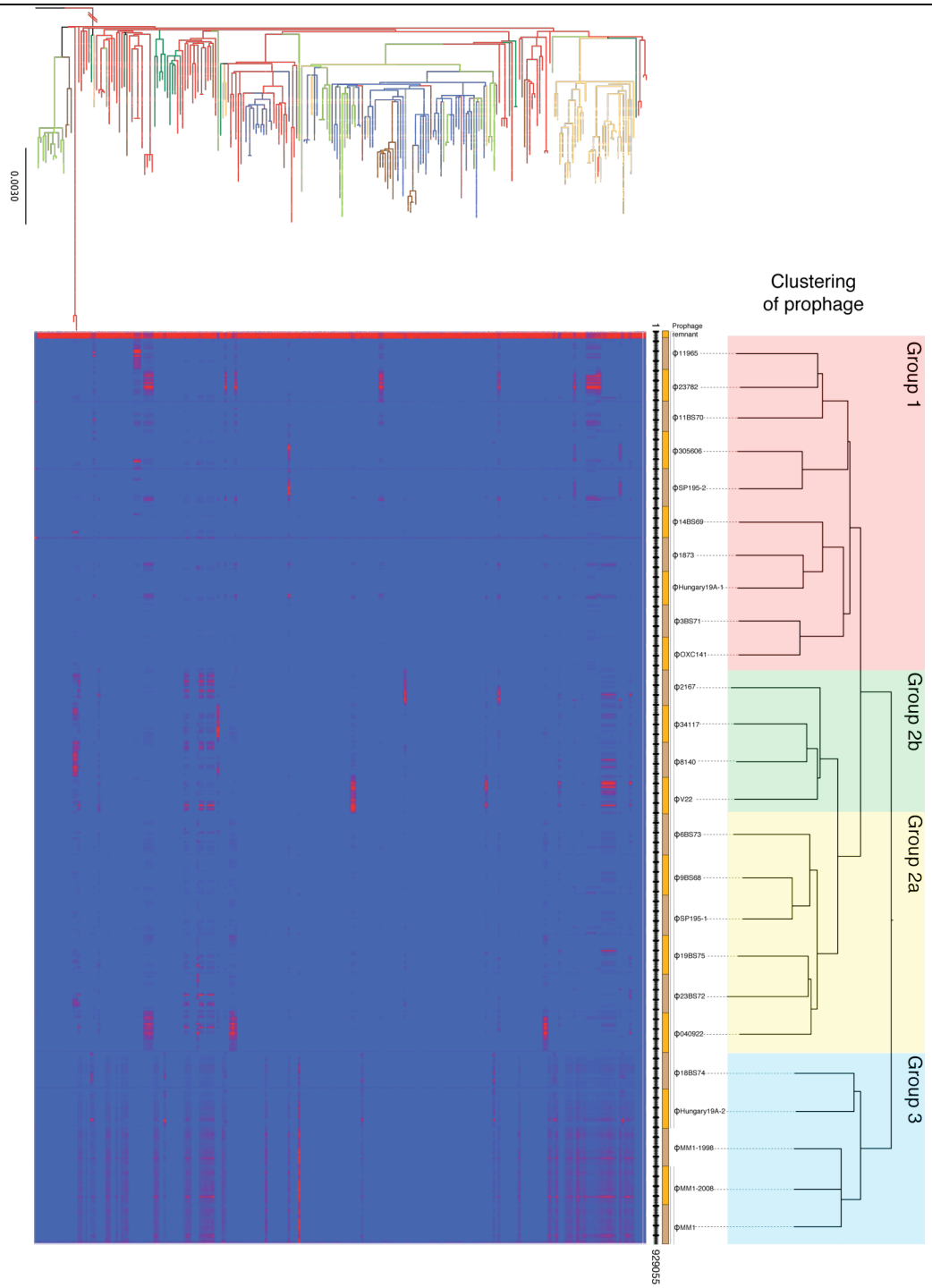


Figure 4.15 Distribution of prophage sequences between PMEN1 isolates. Across the top, the clustering of the prophage sequences is shown (Figure 4.14). The first sequence, not including in the clustering, is the prophage remnant found in the *S. pneumoniae* ATCC 700669 genome. This is conserved among all sequenced members of the lineage, and therefore used to standardize each heatmap, displayed as described in Figure 4.9, to the level of coverage obtained for each sample. The level of nucleotide sequence similarity between the closely related group 3 phage leads to a large amount of cross mapping of reads between these five prophage, hence the mean level of signal across these elements is typically lower than that for prophage in groups 1 or 2.

4.3 Discussion

The ability to distinguish vertically acquired substitutions from horizontally acquired sequences is crucial to successfully reconstructing phylogenies for recombinogenic organisms such as *S. pneumoniae*. Phylogenies are in turn essential for detailed studies of events such as intercontinental transmission, capsule type switching, and antibiotic-resistance acquisition. Although current epidemiological typing methods have indicated that recombination is frequent among the pneumococcal population, they cannot sufficiently account for its impact on relations between strains at such high resolution. Only the availability of such a sample of whole-genome sequences makes it possible to adequately reconstruct the natural history of a lineage. The base substitutions used to construct the phylogeny have accumulated over about 40 years and occur, on average, once every 15 weeks. Recombinations happen at a rate about ten fold more slowly but introduce a mean of 72 SNPs each. The responses to the different anthropogenic selection pressures acting on this variation are distinct. The apparently weak selection by aminoglycosides and chloramphenicol has led to the occasional deletion of loci encoding resistance to these antibiotics. By contrast, resistance to macrolide antibiotics has been acquired frequently throughout the phylogeny, with selection strong enough to drive supplementation or replacement of the resistance afforded by the *mef* efflux pump with the broader-range resistance provided by *ermB*-mediated target modification. The response to vaccine selection is different and involves the depletion of the resident population before it can respond to the selection pressure and thereby opens the niche to isolates that already expressed nonvaccine serotypes. This is likely to reflect the high host population coverage of PCV7 in the USA, as opposed to macrolides or other antibiotics, and the relative likelihood of the recombination events that underlie these responses.

Over a few decades, this single pneumococcal lineage has acquired drug resistance and the ability to evade vaccine pressure multiple times, demonstrating the remarkable adaptability of recombinogenic bacteria such as the pneumococcus. PMEN1 is, nevertheless, only one lineage of this pathogen. Our relative ignorance of the forces that affect bacterial evolution over the long term is illustrated by BM4200 (Buu-Hoi and Horodniceanu, 1980), a multidrug-resistant serotype 23F isolate of ST1010 sequenced as the outgroup for this analysis (Figure 4.1). This isolate dates to 1978 but, despite its apparent similarity to PMEN1 strains, has been found very rarely

since then. Hence, the multidrug-resistant phenotype is not sufficient to guarantee success, suggesting that the nature of the resistances themselves, or other factors in the genotype, may be important for the relative prevalences of these two clones.

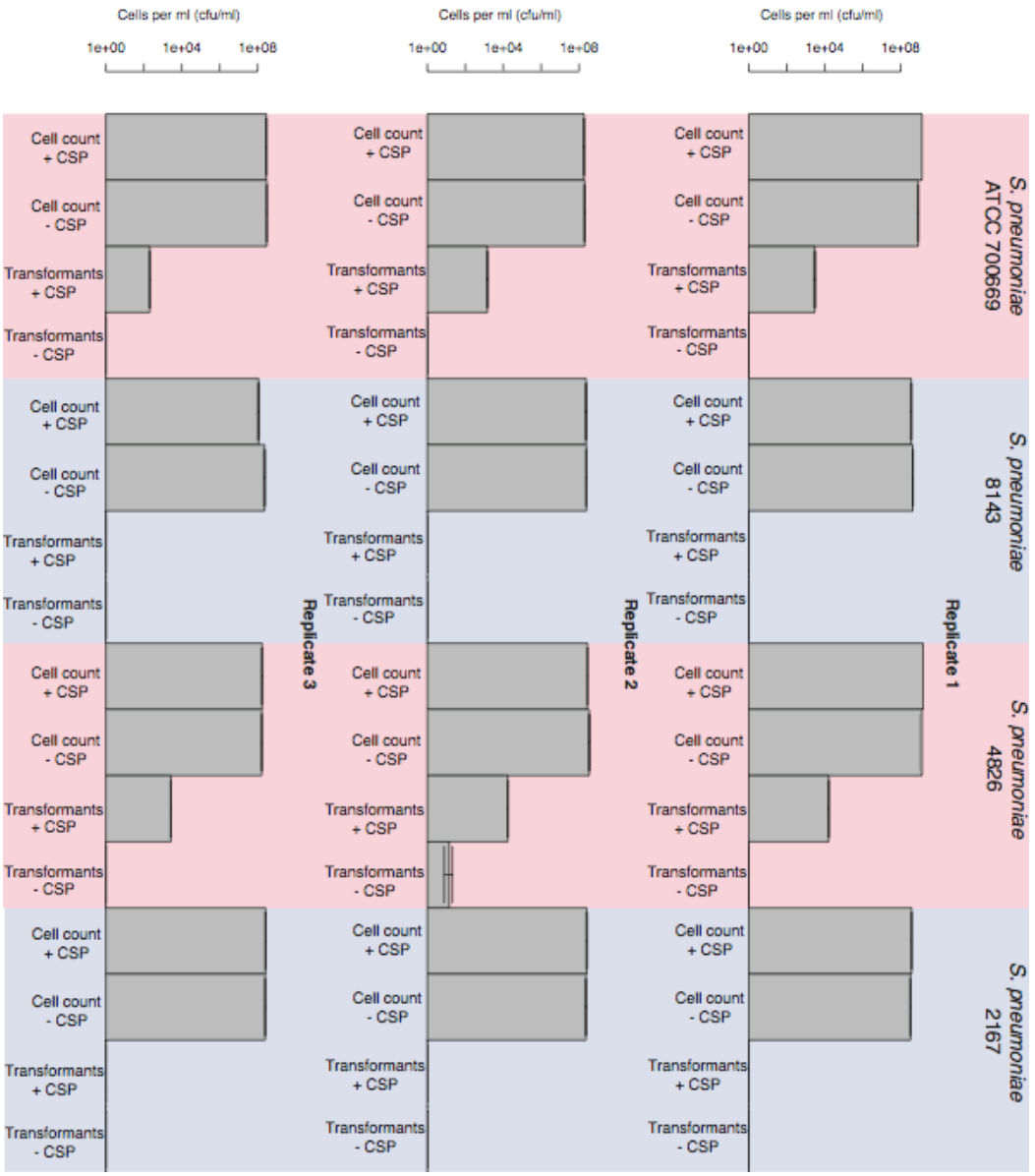


Figure 4.16 Testing the competence of isolates carrying prophage inserted in *comYC*. Four strains (two lysogens carrying group 2b prophage, *S. pneumoniae* 2167 and 8143, shaded blue, and related strains with intact *comYC* genes, *S. pneumoniae* ATCC 700669 and 4826, shaded red) were assayed for their ability to incorporate a kanamycin resistance marker into the *cps* locus in a competence stimulating peptide (CSP)-dependent manner (see Methods). The bars show either the total cell count or the number of transformants, as indicated, in colony forming units per milliliter, with error bars showing the standard error of the mean ($N = 3$ for each measurement). No CSP-dependence in the total population of cells was observed, suggesting that the prophage, if they are active, are not induced by CSP. Kanamycin-resistant colonies were only observed in experiments using the strains with intact *comYC* genes, suggesting that prophage that disrupt the *comYC* gene abrogate the competence of the pneumococcus and prevent it taking up DNA from the environment.

Table 4.1 Convergence of the PMEN1 phylogeny. Quantification of the similarity between phylogenies produced by subsequent iterations of the algorithm used to detect recombination and construct the tree. The branch score difference takes the length of branches into account, while the Robinson-Foulds metric is based only on the topology of the tree.

Comparison	Branch score distance	Robinson-Foulds metric
Iteration 1 vs Iteration 2	1.00	181
Iteration 2 vs Iteration 3	6.49×10^{-3}	41
Iteration 3 vs Iteration 4	1.44×10^{-3}	19
Iteration 4 vs Iteration 5	9.86×10^{-4}	7
Iteration 5 vs Iteration 6	3.87×10^{-4}	5
Iteration 6 vs Iteration 7	5.39×10^{-4}	14
Iteration 7 vs Iteration 8	7.34×10^{-4}	11
Iteration 8 vs Iteration 9	4.83×10^{-4}	14

Table 4.2 CDSs frequently disrupted by mutations in the PMEN1 phylogeny. CDSs affected by a significantly high number of disruptive mutations in the PMEN1 phylogeny.

CDS	Gene Name	Product	Length (bp)	Disruptions	*p* value
SPN23F17730	-	Putative *psrP* glycosyltransferase	905	24	0
SPN23F01290	*pspA*	Pneumococcal surface protein A (pseudogene)	2176	15	3.33×10^{-16}
SPN23F15600	-	Putative phage protein	317	5	4.08×10^{-8}
SPN23F19760	-	Lantibiotic processing protease	1739	8	4.76×10^{-8}
SPN23F15300	*hol*	Antiholin	332	5	5.13×10^{-8}
SPN23F06290	-	Membrane protein	752	6	9.60×10^{-8}
SPN23F05270	-	IS1239 transposase (pseudogene)	449	5	2.26×10^{-7}
SPN23F12860	-	Uncharacterised ICE protein	362	4	3.92×10^{-6}
SPN23F21150	-	Putative DNA binding protein	440	4	8.41×10^{-6}
SPN23F14790	-	IS1239 transposase	1007	5	1.13×10^{-5}
SPN23F17840	-	Transposase	479	4	1.17×10^{-5}

5 *In vitro* transformation of *S. pneumoniae* ATCC 700669

5.1 Introduction

During the 1960s, independent studies were conducted in which mutations causing resistance to erythromycin (Iyer and Ravin, 1962), aminopterin (Ephrussi-Taylor *et al*., 1965), streptomycin (Chen and Ravin, 1966) or amethopterin (Sirotnak *et al*., 1969), or markers that overcame auxotrophic mutations preventing catabolism of maltose (Lacks, 1966) or biosynthesis of uracil (Morse and Lerman, 1969), were transferred between pneumococci. All of these studies concluded that independent mutations giving the same phenotype were transferred at reproducibly different rates, allowing markers to be categorized according to their transformation efficiency (Ephrussi-Taylor *et al*., 1965; Lacks, 1966; Sirotnak *et al*., 1969). The observation that markers induced by specific mutagens tended to transfer with similar efficiencies suggested that the disparity in transformation rates reflected the ease with which different types of mutation were transferred.

Following the advent of DNA sequencing, contemporaneous work studying the aminopterin resistance (*amiA*) and amylomaltase loci identified transversion mutations A•T↔C•G and C•G↔G•C as markers transferred with a high efficiency, the transversion A•T↔T•A having an intermediate efficiency, while transitions acted as low efficiency markers (Claverys *et al*., 1981; Lacks *et al*., 1982; Claverys *et al*., 1983), although some inconsistencies, ascribed to neighbouring sequence context, were identified. Deletions 3 bp or shorter also acted as low efficiency markers (Lacks *et al*., 1982; Gasc and Sicard, 1986; Gasc *et al*., 1987), while those 5 bp or longer were transferred as ‘very high efficiency markers’ (Claverys *et al*., 1981), although there is some evidence that this property is somewhat tempered as the deletion increases in length (Lacks, 1966; Claverys *et al*., 1980; Claverys *et al*., 1981; Lacks *et al*., 1982; Claverys *et al*., 1983; Gasc *et al*., 1987). Somewhat contradictory data indicating both insertions and deletions increase the rate with which flanking markers are transferred through transformation have also been reported (Lefevre *et al*., 1989; Pasta and Sicard, 1996).

Early in the study of transformation, a 'high efficiency integration' (*hex*) phenotype was recognized in recipient cells that no longer discriminated between low and high efficiency markers (Lacks, 1970), although deletions of the appropriate length remained very high efficiency markers in both backgrounds (Claverys *et al*., 1980). Strains exhibiting the *hex* phenotype were found to have an elevated spontaneous mutation rate, suggesting they had lost a DNA repair system (Tiraby and Fox, 1973). This function was found to be encoded by two genes, *hexA* (Balganesh and Lacks, 1985) and *hexB* (Prats *et al*., 1985), which were subsequently identified as being homologous with the mismatch repair (MMR) system in *Escherchia coli* and eukaryotes (Haber *et al*., 1988; Priebe *et al*., 1988; Mankovich *et al*., 1989; Prudhomme *et al*., 1989). Therefore the difference in transfer efficiency between base substitution markers reflects the rate at which the non-canonical base pairings formed by the invasion of genomic DNA by the acquired ssDNA are corrected by the mismatch repair system.

However, problems arise when trying to extrapolate from these relatively simple scenarios, the transfer of small numbers of selectable polymorphisms at specific loci, to understanding the frequent exchanges of sequence among the natural pneumococcal population. The interactions between small numbers of markers with different transformation efficiencies was found to be very complex (Gasc *et al*., 1989), and if correction of each mismatch within a recombination were as efficient as observed for individual polymorphisms many recombinations observed to occur *in vitro* and *in vivo* would be impossible (Lacks, 1966; Majewski *et al*., 2000). This appears to be a consequence of the saturation of the mismatch repair system by a relatively small number of polymorphisms in imported DNA (Humbert *et al*., 1995).

Nevertheless, a linear relationship between the mean level of sequence divergence and the logarithm of the frequency of recombination events is observed, implying polymorphisms do constitute a significant barrier to the exchange of sequence between bacteria (Roberts and Cohan, 1993; Vulic *et al*., 1997; Majewski *et al*., 2000). This has been suggested to be the consequence of the requirement for a minimum threshold length of perfect sequence identity (a 'Minimal Efficiently Processed Segment', or MEPS) at each end of a recombination to allow a strand

exchange to occur (Majewski and Cohan, 1998). Based on the changing frequency of transfer with donor sequences of different levels of divergence from the recipient, the minimum summed length of the two MEPS flanking *S. pneumoniae* recombinations was estimated to be 27 bp (Majewski *et al*., 2000). In order to test how recombination events could occur with such constraints, yet allow the diversification observed in the PMEN1 isolates, an *in vitro* transformation of *S. pneumoniae* ATCC 700669 was performed.

5.2 Analysis of *in vitro* transformants

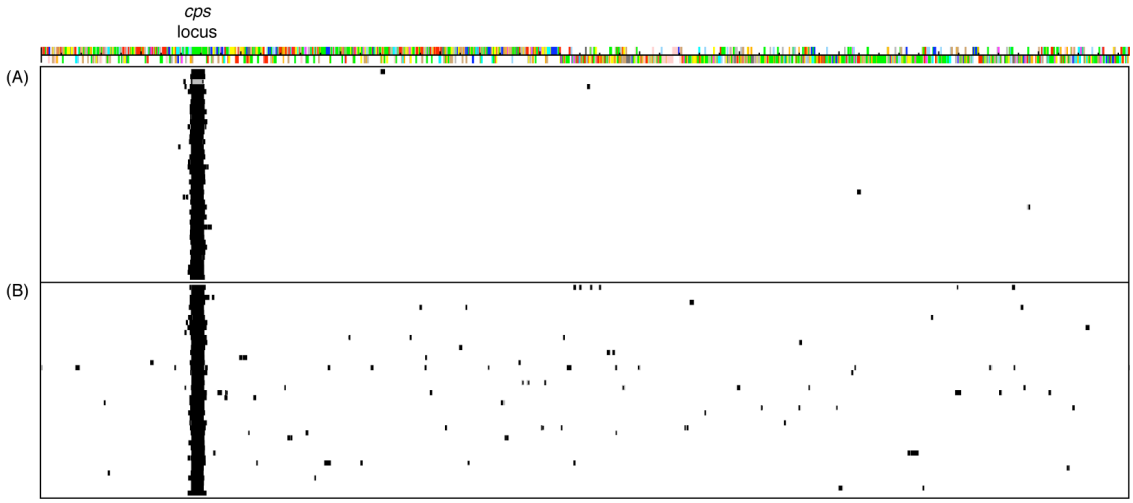


Figure 5.1 Distribution of recombination events across the genome. The annotation of the 2,182,008 bp *S. pneumoniae* ATCC 700669^{lab} genome is displayed across the top, coloured as described in Chapter 3. The *cps* locus, spanned by the selected recombination in transformants, is labelled. Underneath, each row represents an individual sequenced transformant. RSSs are indicated on the rows by black blocks, representing the donor sequence, flanked by grey FRs, which show the ambiguous boundaries between RSSs and the flanking polymorphisms with recipient alleles. The transformants are grouped as (A) *S. pneumoniae* ATCC 700669^{lab} transformed at 2 ng mL^{-1} donor DNA and (B) *S. pneumoniae* ATCC 700669^{lab} transformed at 200 ng mL^{-1} donor DNA.

5.2.1 Genome-wide exchange of sequence between pneumococci

The precise isolate of *S. pneumoniae* ATCC 700669 used in the experiment is hereafter designated *S. pneumoniae* ATCC 700669^{lab} (see Material and Methods). This was transformed with genomic DNA from a rough derivative of *S. pneumoniae* TIGR4 that carries a kanamycin resistance marker at the capsule biosynthesis (*cps*) locus (Pearce *et al*., 2002). Multiple transformations were performed using a concentration of either 2 ng mL^{-1} or 200 ng mL^{-1} of donor genomic DNA. Recombinations affecting the *cps* locus were detected either through selection with

kanamycin alone, or kanamycin supplemented with penicillin. This latter condition was used to test the hypothesis that the transfer of *cps* loci from penicillin-sensitive pneumococci, such as TIGR4Δ*cps*, to penicillin resistant strains, such as ATCC 700669, may be inhibited by selection against any co-transfer of the flanking antibiotic-sensitive penicillin-binding protein gene alleles.

With selection on just kanamycin, the transformation with the lower concentration of DNA produced 38 fold fewer transformants (Wilcoxon rank sum test, $p = 1.5\text{x}10^{-4}$), suggesting that the availability of the marker was limiting the rate of transformation. Dual selection with penicillin as well caused a small, but non-significant, decrease in transformation rates: with 2 ng mL^{-1} DNA, a 12.6% decrease was observed (Wilcoxon rank sum test, $p = 0.53$), while with 200 ng mL^{-1}, there was a fall of 9.6% (Wilcoxon rank sum test, $p = 0.21$). Therefore, selection for β-lactam resistance does not appear to significantly inhibit exchange with penicillin sensitive lineages at the *cps* locus, instead suggesting a strong limitation on the size of recombination events reducing the impact of linkage between genes. To test this hypothesis, 21 isolates from each of the four examined conditions (low and high DNA concentration, and with and without penicillin selection) were sequenced using the Illumina platform.

Alignment of the complete genome sequences of the donor and recipient strains identified 21,512 base substitutions, 476 insertions in TIGR4Δ*cps* relative to ATCC 700669^{lab} (1 bp – 14,153 bp in size) and 578 insertions in ATCC 700669^{lab} (1 bp – 76,827 bp in size). By mapping Illumina reads simulated from the donor sequence to that of the recipient, it was possible to identify 16,067 base substitutions, all but 15 of which were also found through whole genome alignment. Sequence data from the 84 transformants identified 2,347 polymorphic sites, of which just 67 did not correspond to polymorphisms transferred from the donor. Eleven of these sites represent difficulties with mapping; a further nine appear to have arisen through intragenomic recombinations affecting an IS element and the repetitive surface protein gene *pclA*. The remaining 47 sites appear to be spontaneous mutations occurring *in vitro*; 25 of these are C•G→T•A substitutions likely to represent the consequences of cytosine oxidation and deamination (Kreutzer and Essigmann, 1998), which may be caused by

the high levels of hydrogen peroxide produced by *S. pneumoniae* during aerobic growth (Pericone *et al*., 2000).

Recombinant sequence segments (RSSs) were detectable in transformant sequence data as loci containing donor alleles at polymorphic sites, defining the minimum size of the recombination, bounded by recipient alleles at the flanking polymorphic sites, demarcating the maximum size of the exchange (Figure 5.1). The actual length may be estimated as being the median (L50) between these two limits, positioning the boundary half way through each flanking region (FR), expressed as a distance relative to the donor ($L50_D$) or recipient ($L50_R$) genome. The selected recombination at the *cps* locus was detected in all transformants, with at least one of the 84 isolates having recombinant sequence between the genomic loci 294,349 bp and 340,522 bp; this region of the chromosome is henceforth referred to as the 'primary locus', as these recombinations have been driven by selection. Furthermore, 112 unselected, 'secondary' recombinations were observed outside the *cps* locus, with one strain having a total of 18 RSSs. The mean proportion of the recipient genome found to have undergone recombination was 1.4%, ranging up to a maximum of 2.3%. Secondary recombinations were significantly more common in the strains transformed at a high concentration of DNA (mean of 2.48 secondary events per strain) than at a low concentration (mean of 0.26 secondary events per strain; Wilcoxon rank sum test, $p = 2.0 \times 10^{-8}$). Hence the effective concentration of DNA available for recombination inside the cell can vary. This implies that recombination events involving separate DNA strands can occur within the same cell concurrently and independently, rather than all arising from the import of a single large molecule of DNA, as has been observed in *S. agalactiae* (Brochet *et al*., 2008) and inferred from *C. difficile* genome sequences (He *et al*., 2010).

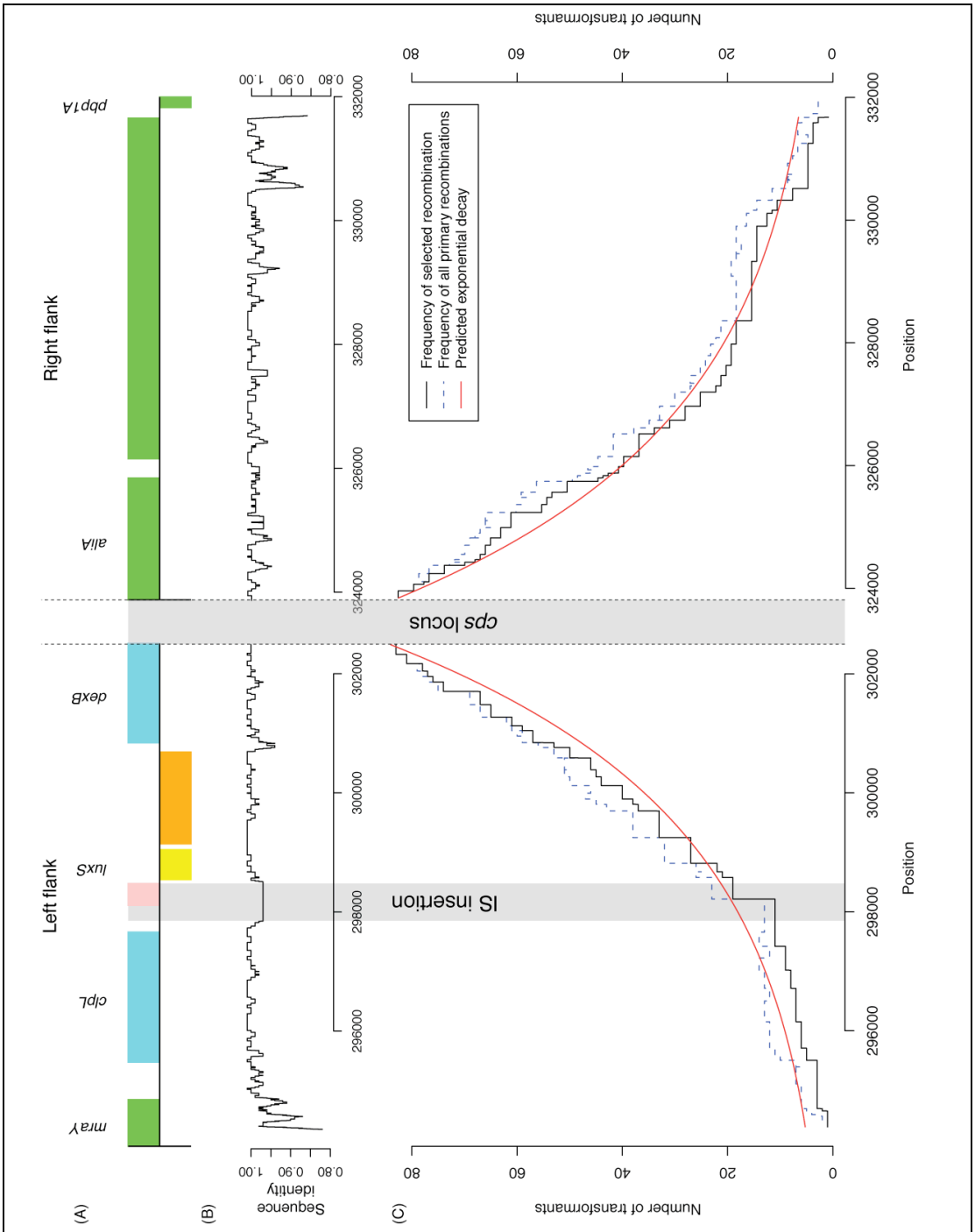


Figure 5.2 Structure of recombinations within the primary locus. (A) Annotation of two sections immediately flanking the *cps* locus, within the primary locus. The position of the *cps* locus, lying between the two flanks, is marked by the grey bar. Similarly, the extent of an IS element insertion in *S. pneumoniae* ATCC 700669[lab], not present in *S. pneumoniae* TIGR4**Δ***cps*, is marked by a grey bar within the left flank. (B) A plot of sequence identity showing the level of similarity between the donor and recipient throughout the locus. The value on the graph represents the proportion of the aligned flanking 100 bp that is identical in both strains. (C) Distribution of recombination sizes. At the *cps* locus itself, all 84 isolates have the donor allele. The black line traces the rate at which this selected recombination ends relative to the extent of the locus. This process can be modelled as an exponential decay, which is represented by the red line for each flank. Other recombinations occur within the primary locus, and the overall prevalence of these, along with the selected recombination, is indicated by the blue dashed line. This shows the total number of isolates in the recombinant state throughout the displayed locus.

5.2.2 Characterisation of 'capsule switching' recombinations

The high density of SNPs within the primary locus allows a high-resolution view of the boundaries of the selected RSSs that span the *cps* locus. For each distance from the edge of the selected *cps* locus, on the two sides independently, the number of isolates with a selected RSS extending to, or beyond, that point can be plotted (Figure 5.2C). This outlines the rate at which the primary RSSs end relative to their distance from the selected marker, best modelled as an exponential decay, with similar estimates of the decay constant on both sides: 3.42×10^{-4} (95% confidence interval, $3.41\text{-}3.43 \times 10^{-4}$) on the left flank, and 3.40×10^{-4} (95% confidence interval, $3.39\text{-}3.41 \times 10^{-4}$). The symmetrical nature of the decays on the two flanks is not disrupted by the presence of an IS element insertion in the recipient distinguishing it from the donor. Correspondingly, the exponential decay on the left hand side fits the position relative to the recipient (residual sum-of-squares, 172,966) better than that relative to the donor (residual sum-of-squares, 264,836). Such observations imply that the size of the RSSs is dictated by a Poisson process that involves the recipient's DNA, hence may represent a process such as resolution of the heteroduplex, rather than events during the pre-processing of the donor strand, such as endonucleolytic cleavage during DNA import.

The symmetry is also in spite of the low level of correlation in terms of sequence identity between the donor and recipient between these two regions, either side of the *cps* locus (Pearson correlation, $R^2 = 0.0011$), suggesting the density of SNPs observed in this region is not enough to significantly affect the distribution of recombination events (Figure 5.2B). In the absence of sequence identity affecting the exponential declines, the number of isolates in the recombinant state should halve over each 2 kb stretch of sequence. On the basis of this rate, of the recombinations that directly affect the *cps* locus, 7.4% will affect *pbp1A* and 5.8% will affect *pbpX*. Less than 0.1% of recombinations encompassing the *cps* locus would be expected to replace both *pbpX* and *pbp1A* in their entirety, explaining the lack of a significant inhibition of *cps* transfer by ampicillin selection.

5.2.3 Mosaic recombinations in the *cps* locus

Thirty-six further RSSs occur in the primary locus but do not span the *cps* gene cluster (Figure 5.2C). This high density of unselected primary recombinations suggests that they are associated with that spanning the *cps* locus. Supporting this hypothesis, strains transformed with the lower concentration of DNA actually have a larger mean number of primary RSSs (1.52 per strain) than those exposed to the higher concentration (1.33 per strain), although this difference is not significant (Wilcoxon rank sum test, p = 0.33). This suggests that the frequency of recombinations in close proximity to the selected event is independent of the external DNA concentration, indicating that these mosaic recombinations (MRs) are not a consequence of the locus acting as a hotspot for integrations by several imported strands, but rather reflects multiple RSSs originating from the same piece of donor DNA.

A histogram of the $L50_R$ of the recombination events within the primary locus reveals a bimodal distribution (Figure 5.3A). While the selected recombinations spanning the *cps* locus, 21,373 bp long in *S. pneumoniae* ATCC 700669^{lab}, have a modal $L50_R$ of around 27 kb, the nearby flanking events are mainly 5 kb or less in size. The shapes of the two distributions are also distinct. The smaller events form an approximate exponential distribution, supporting the suggestion that RSSs are generated through a Poisson process with a per base probability of strand exchange, λ_R, of 5.0×10^{-4} bp^{-1}. The theoretical mean length, λ_R^{-1}, is therefore 2 kb. By contrast, the lengths of the events that span the *cps* locus are not exponentially distributed. This is an artefact of selection; although the longer events are less frequent, they are more likely to span the selectable marker, and hence are observed at an unusually high frequency relative to shorter events when compared with unselected recombinations.

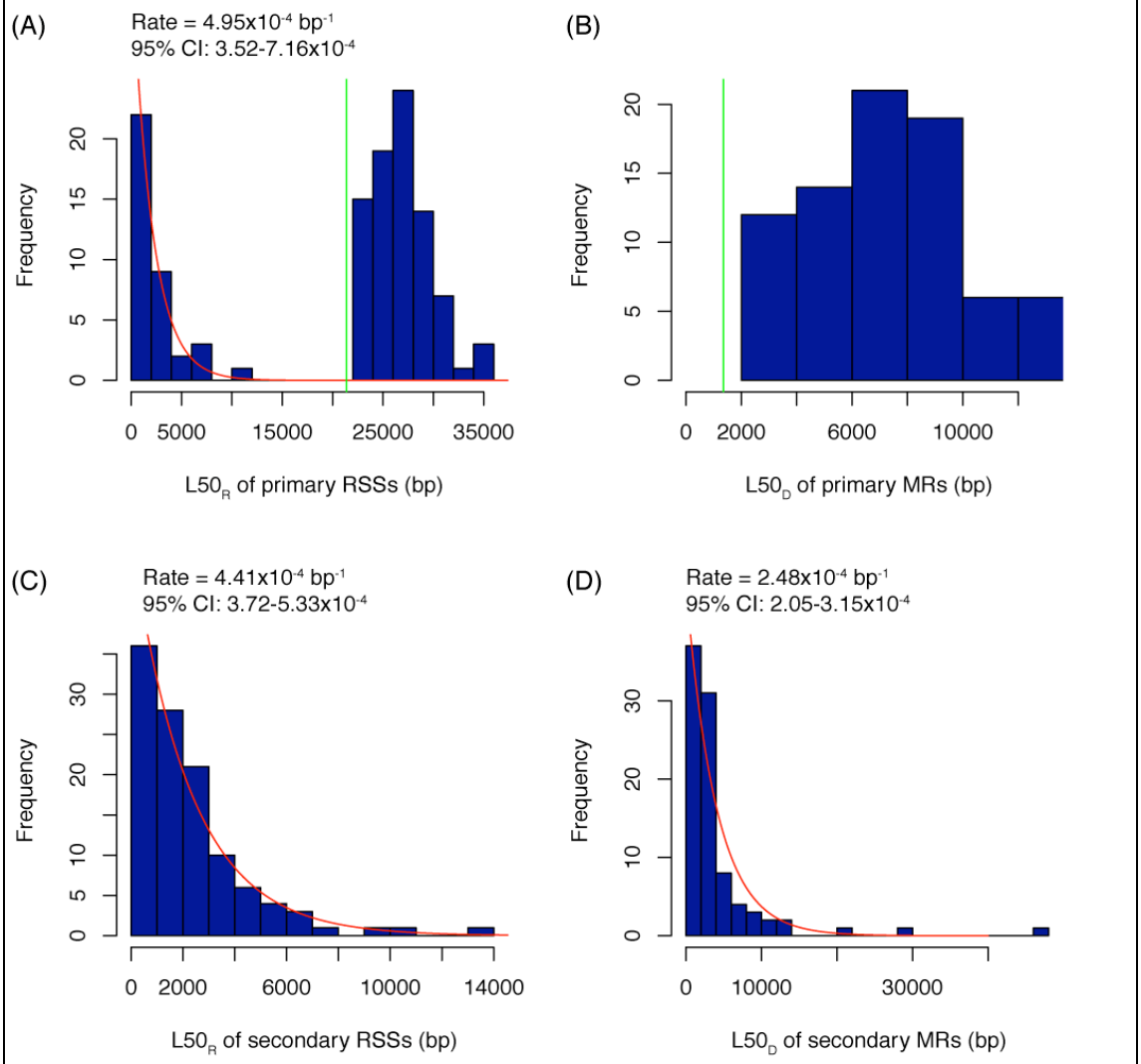


Figure 5.3 Distribution of recombination sizes. (A) Histogram showing the distribution of RSS sizes as $L50_R$ lengths. The vertical green line shows the length of the *cps* locus relative to the recipient (21,372 bp). The smaller, unselected recombinations are modelled as an exponential distribution (red line), the rate parameter of which is displayed. (B) Histogram showing the distribution of donor strand lengths participating in primary recombinations, as estimated using the $L50_D$ lengths of the MR boundaries. The vertical green line shows the length of the *cps* locus relative to the donor (1,354 bp). (C) Histogram showing the distribution of secondary RSS sizes as $L50_R$ lengths. These are modelled as an exponential distribution, indicated by the red line. (D) Histogram estimating the sizes of donor strands participating in secondary recombinations, showing the distribution of MR $L50_R$ lengths. These are modelled as an exponential distribution, indicated by the red line.

As all the primary RSSs in each strain, forming a single MR, originate from the same molecule of DNA, the length of the donor strand participating in the recombinations around the primary locus can be estimated (Figure 5.3B). The median $L50_D$ of these measurements is 7.3 kb, with no significant differences between strains transformed with high or low levels of DNA were detected (Wilcoxon rank sum test, $p = 0.93$). This reflects the consequences of selection at this locus: donor strands smaller than

the kanamycin resistance gene evidently will not be observed, and furthermore the selection process will bias the range towards larger MRs. Despite these constraints, the $L50_D$ is still comparable in length to the median size of the DNA strands imported inside the cell (~6.6 kb) (Morrison and Guild, 1972). Within the primary locus boundaries, the mosaic pattern of transfer could arise either through multiple exchanges involving the same donor strand, which may be enforced by cleavage of the original molecule into several pieces on entry, or through localised action of repair processes acting on a single, larger transfer.

5.2.4 Analysis of secondary recombinations

The lengths of the secondary RSSs are exponentially distributed with a λ_R of $4.4x10^{-4}$ bp and a median of 2.3 kb (Figure 5.3C), very similar to the unselected RSSs within the primary locus but contrasting with the size and distributions of the selected recombination events. They also exhibited a similar pattern of mosaicism to those at the primary locus. A bootstrapping algorithm (see Materials and Methods) was used to organise the 112 secondary RRSs into 90 MRs, each likely to have been derived from a single donor strand. All RSSs less than 8 kp apart were linked into MRs; although the majority of MRs consisted of only one RSS, up to four could be found in significantly close proximity. The stretches of unmodified recipient sequence between linked RSSs were sometimes only identifiable by a single SNP, although they usually contained multiple polymorphic sites; their median length was 509 bp (mean length of 2.5 kb). The most distant sequences joined were 43.6 kb apart; this occurred in a strain where the only three secondary RSSs all fell within a 45 kb region of the genome. Unlike those at the *cps* locus, MRs themselves were exponentially distributed like their RSS components (Figure 5.3D).

The mean sequence divergence across the $L50_R$ of detected RSSs was 0.9%, which falls to 0.7% when considering the $L50_R$ of the entire MRs. Hence the heterogenous pattern of SNP density observed occurring *in vivo* appears to represent an intrinsic property of the mechanism of pneumococcal transformation. However, such estimates are based only on RSSs that can be detected through transfer of SNPs. Using a sliding window analysis (see Materials and Methods), it can be estimated that a quarter of the total number of recombinations outside the primary locus are not detectable. These

are typically short events, which would decrease the overall median length estimate and SNP density values. However, it may be that an even greater proportion is undetectable, if the presence of a single polymorphism is enough to significantly inhibit transfer. Although no such effect was observable within the primary locus, this problem can also be investigated using the detected secondary recombinations.

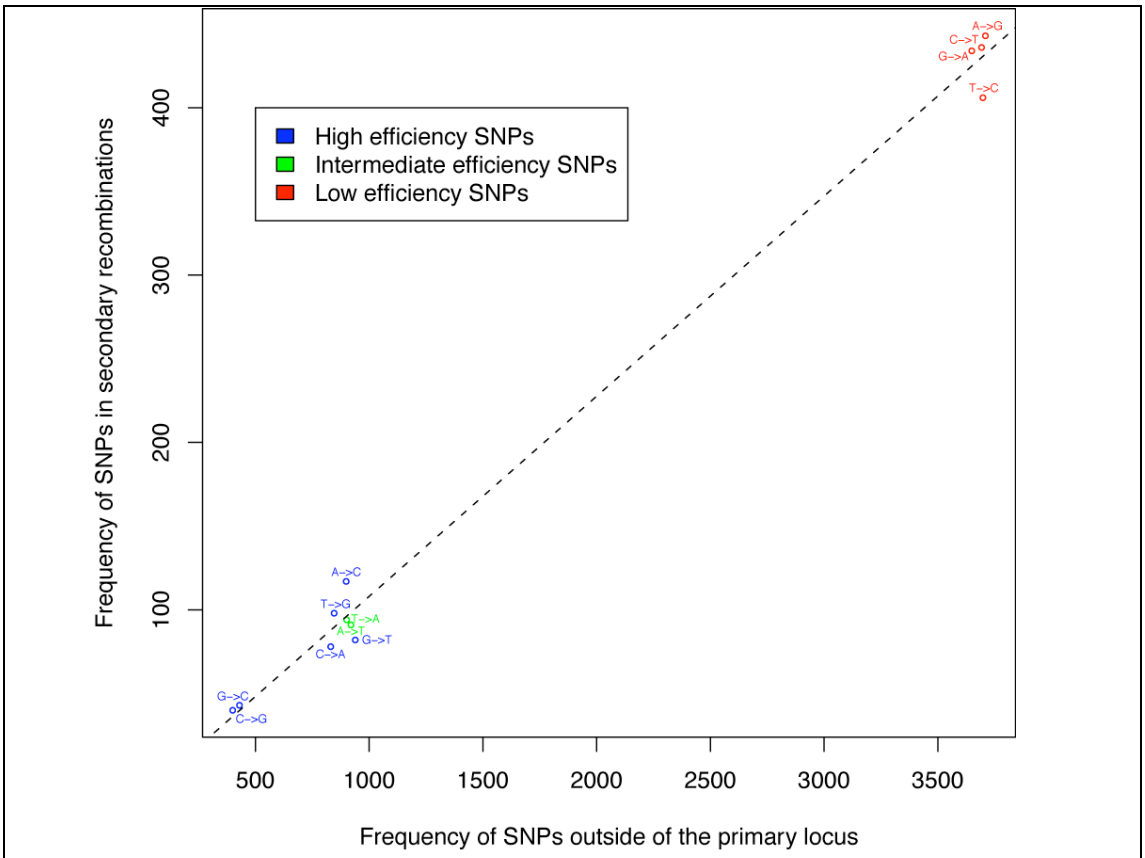


Figure 5.4 Correlation of SNP type frequencies outside the primary locus with those detected within secondary recombinations. Points are coloured according to the efficiency with which they are transferred through transformation when mismatch repair is active. This shows that SNPs repaired effectively by mismatch repair are the most common, and are represented in secondary recombinations at the same level as those SNPs repaired less effectively.

The sliding window analysis found that 62-70% of observed recombinations had a lower mean level of sequence diversity than expected from the distribution of SNPs between the donor and recipient, with the result varying as the length of surrounding sequence considered in calculating the sequence identity was changed between 100 bp and 2 kb. However, this deviation was not significant at any of the tested surrounding sequence lengths (Fisher's exact test, p = 0.080-0.11). Hence no enrichment of RSSs in regions of high sequence similarity can be observed. This may reflect the only constraints, in terms of sequence similarity, being the length of the MEPS. In this

dataset, only one RSS contravenes the suggested 27 bp threshold (Majewski *et al*., 2000); this event transfers a single SNP and has a total flanking identity length of 26 bp. Despite largely conforming to this condition, the population of FRs were not found to be significantly longer than the population of distances between SNPs: the sliding window analysis found half of the 5' and 3' MEPS were larger than the expected length given the $L50_R$ of the RSS (Fisher's exact test, $p = 1.0$). Hence it appears that sequence diversity causes few limitations on the exchange of sequences between *S. pneumoniae* genotypes.

5.2.5 Efficiency of polymorphism transfer

As just a few hundred SNPs appear to be sufficient to overwhelm the pneumococcal MMR system, it is not expected that this form of repair is likely to have impacted on recombinations to a great extent. Correspondingly, the number of each type of SNP outside the primary locus correlates tightly with the frequencies of these mutations in the secondary recombinations ($R^2 = 0.99$, $p = 1.5 \times 10^{-12}$; Figure 5.4). Furthermore, the frequency of each SNP on the outermost position of each RSS is proportional to its prevalence in the nearest flanking unchanged position ($R^2 = 1.0$, $p < 2.2 \times 10^{-16}$). Hence there is no evidence that the low efficiency markers lead to entire recombinations being lost at a higher frequency, or that they trigger localised repair, which might have been a mechanism for the formation of MRs. The alignment of the donor and recipient does show, however, that the most frequent mutations distinguishing them are the transversions, supporting the hypothesis that MMR has evolved to repair the most common mutations most efficiently (Figure 5.4).

Past work has shown that the efficiency with which indels are transferred appears to depend upon their length. The density of deletions in the donor, relative to the recipient, within the secondary recombinations is not significantly different to that outside such regions (Table 5.1), suggesting they have no effect on recombination rates. However, insertions in the donor relative to the recipient are significantly excluded from secondary recombinations (Table 5.1). The mechanism behind this inhibition may be inferred by comparing the sizes of the insertions found in the recombinations; while there is no significant difference in length between the insertions in the recipient within and outside secondary recombinations (Wilcoxon

rank sum test, p = 0.97), the transferred insertions in the donor are significantly smaller than expected (Wilcoxon rank sum test, p = 0.018). This is despite there being no statistically significant difference between the distributions of indels distinguishing the two strains (Wilcoxon rank sum test, p = 0.94). Hence it appears that constraints on the size of the donor strand, likely resulting from the cleavage of the DNA molecule as it is imported, inhibits the acquisition of large insertions in the donor sequence.

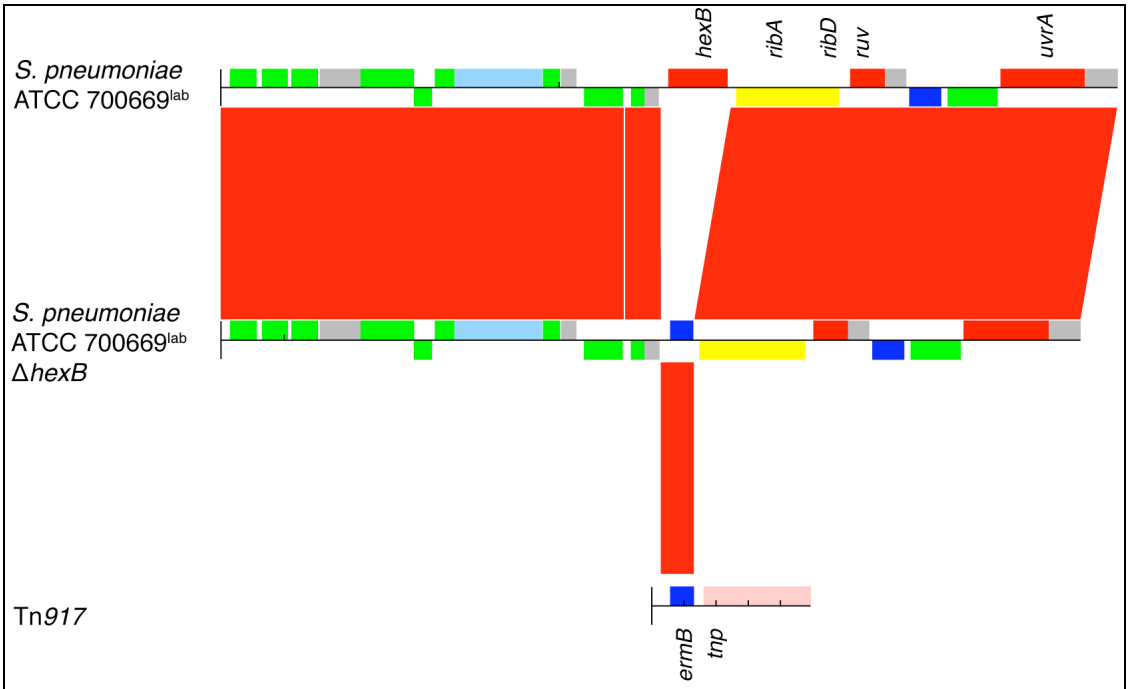


Figure 5.5 Knock out of the *hexB* gene in *S. pneumoniae* ATCC 700669^{lab}Δ*hexB*. This shows an ACT comparison (as described in Figure 3.2) of the *hexB* locus of the *S. pneumoniae* ATCC 700669^{lab} reference strain compared against the orthologous region of an MMR-defective transformant, which has an *ermB* erythromycin resistance in place of *hexB*. The selectable marker was derived from *Tn*917, shown below.

The density of indels in FRs was examined to test the hypothesis that they are capable of stimulating the transfer of adjacent sequence, even when they themselves are not transferred. Both insertions and deletions in the donor were found to be excluded from the FRs, although the result was only significant for deletions (Table 5.1). However, for both types of indels, those in the FRs were significantly larger than expected (Wilcoxon rank sum test, p = 0.00029 for deletions and 0.030 for insertions), although manual inspection of the short read alignments suggested that very few of these indels were actually transferred as part of the associated RSS. Overall, this implies that small indels, of just a few bases, are particularly strongly

excluded from FRs, likely because they interfere with strand exchange in MEPS, though it remains possible that larger indels may trigger the transfer of neighbouring loci.

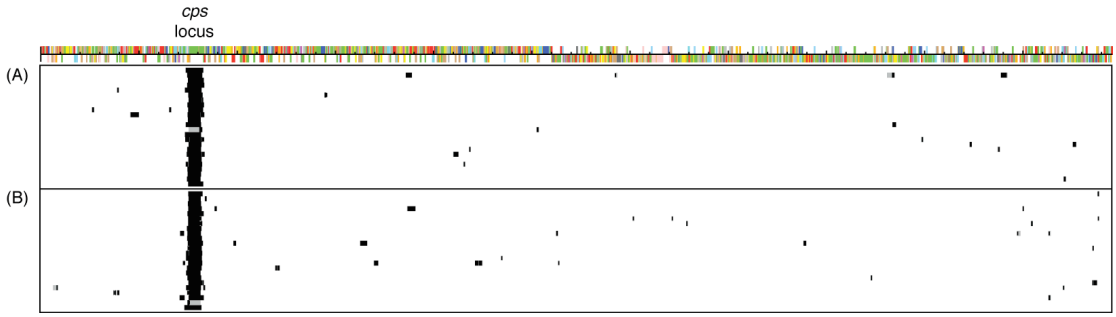


Figure 5.6 Distribution of recombination events across the genome when testing the effect of MMR, displayed as described in Figure 5.1. The transformants are grouped as (A) *S. pneumoniae* ATCC 700669^{lab} transformed at 200 ng mL^{-1} donor DNA and (B) *S. pneumoniae* ATCC 700669^{lab}Δ*hexB* transformed at 200 ng mL^{-1} donor DNA in the second experiment.

5.2.6 The role of mismatch repair

In order to confirm whether mismatch repair played any role in the transformation of *S. pneumoniae* with divergent donor DNA, both *S. pneumoniae* ATCC 700669^{lab} and a mutant with the *hexB* MMR gene knocked out (*S. pneumoniae* ATCC 700669^{lab}Δ*hexB*) were transformed with 200 ng mL^{-1} DNA from *S. pneumoniae* TIGR4Δ*cps*. Twenty-four transformants of each background were then sequenced, confirming the knockout (Figure 5.5) and allowing RSSs to be identified (Figure 5.6). This revealed that the mosaicism within the primary locus, redefined in this experiment as lying between coordinates 293,436 and 333,345 using the same criteria as described previously, was observed in both backgrounds: the wild type isolates had a mean of 1.83 primary RSSs per strain, while the Δ*hexB* isolates had a mean of 1.92 (Wilcoxon rank sum test, $p = 0.57$). Hence the observed mosaicism within MRs does not result from localised repair, but instead appears to represent multiple independent invasions of the recipient genome by the same donor strand.

Correspondingly, the number of secondary RSSs also did not differ significantly, with the Δ*hexB* isolates actually having a slightly higher mean (1.96 per genome, as opposed to 1.04; Wilcoxon rank sum test, $p = 0.12$). Although the median secondary $L50_R$ length was smaller in the MMR defective background (386 bp) than in the wild

type strain (1,453 bp), this difference was not significant (Wilcoxon rank sum test, p = 0.084). Therefore, under the tested conditions, MMR does not appear to be responsible for fully, or partially, repairing any recombination events.

5.3 Discussion

Previous studies of transformation have necessarily focused on rates of transfer of particular selectable mutations. The advantage of high-throughput genome sequencing, coupled with a controlled *in vitro* system, is the ability to characterise multiple recombinations occurring across the chromosome in great detail. Using these data, it can be observed that multiple RSSs, with an exponential size distribution, can be generated from a single donor strand of DNA, thereby forming MRs. Furthermore, multiple MRs, each from a different donor strand, can be generated during a single period of competence.

The heterogenous density of SNPs observed within the MRs corresponds with that observed within the recombination events identified in the PMEN1 isolates. While the median length of homologous recombinations in the PMEN1 population was 2.9 kb, the equivalent measurement from this *in vitro* transformation (the minimum size of MRs relative to the recipient's sequence) is 1.7 kb. However, the overall distribution of sizes from the two studies are similar: the recombination events identified in the clinical isolates follow an approximately exponential distribution with λ_R of about 1.6×10^{-4} bp^{-1} (95% confidence interval of 1.5×10^{-4}-1.7×10^{-4} bp^{-1}). Therefore, rather than the discrepancy representing an overestimation of the lengths of events in the PMEN1 population, this difference results from a lack of sensitivity in identifying small events, due to the conservative nature of the algorithm employed. Comparing the events that span the *cps* locus reveals a more accurate correspondence between the datasets, with the median lengths of those from the *in vitro* data being 27.2 kb, while those from the PMEN1 population have a median of 27.9 kb (Wilcoxon rank sum test, p = 1.0). Hence the method used to reconstruct the history of that lineage appears to have been successful in defining transformation events mediated by the competence system.

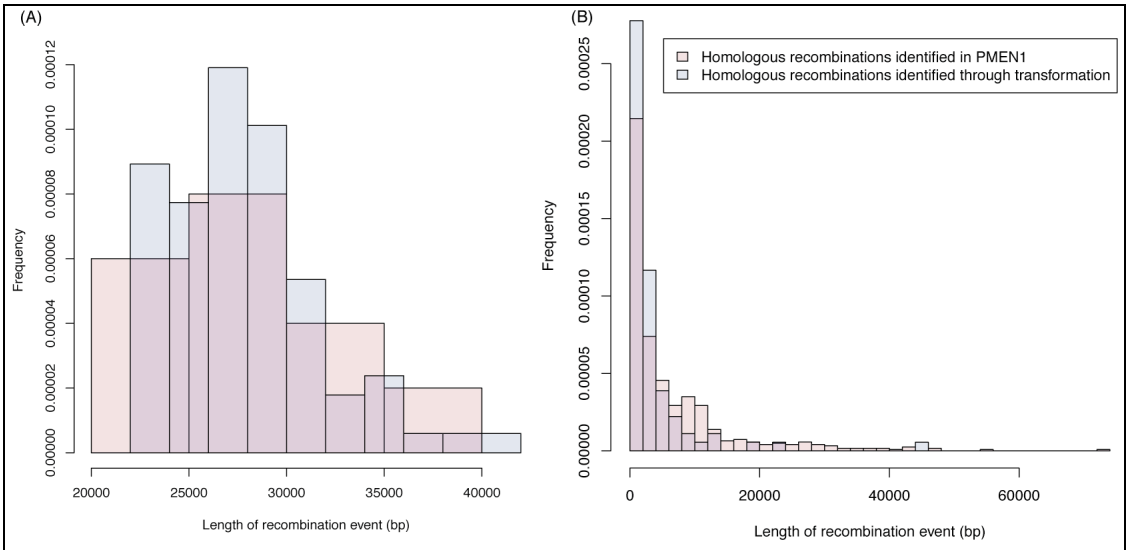


Figure 5.7 Histogram comparing frequencies of detected recombination sizes in PMEN1 isolates and *in vitro* transformants. In both graphs, the blue bars show the data from this experiment, whereas the red bars represent that from the PMEN1 strains. (A) Comparison of the minimum primary MR lengths (*i.e.*, the distance between the outermost polymorphic sites of the donor allele) with the ten serotype switching recombinations characterised in PMEN1. (B) The distribution of secondary MR minimum sizes identified in this experiment with the equivalent size distribution for 615 homologous recombinations (*i.e.* excluding events involving the prophage or conjugative elements) identified as occurring in the PMEN1 isolates.

The distribution of secondary RSSs observed in this study clearly shows that there are few constraints on the exchange of DNA between pneumococci, congruent with the widespread locations and polymorphism densities observed in the PMEN1 population. As discussed, this is likely to be a consequence of the level of divergence between these strains (a mean of one SNP per 101 bp in this experiment). Assuming 27 bp, or more, of identical sequence is required for a strand exchange to occur (Majewski *et al*., 2000), 95% of the aligned length of the recipient genome is capable of participating in strand exchanges. Such a level of divergence is typical between *S. pneumoniae* chromosomes, as comparing all complete pneumococcal genomes to the recipient reveals a minimum and maximum SNP density of one SNP per 150 bp (*S. pneumoniae* JJA) and one SNP per 81 bp (*S. pneumoniae* Hungary 19A-6), respectively. By contrast, using the same approach to compare *S. mitis* B6 (Denapaite *et al*., 2010) with the recipient sequence, just 54% of the sequence would have a sufficiently low SNP density to permit strand exchanges to occur. Hence, the main determinant on the distribution of sequence exchanges between pneumococci across the chromosome is likely to be the random uptake of sequence by the competence system. This implies the observed patterns of recombination frequencies across the *S.*

pneumoniae PMEN1 genome are determined by selection and the level of detectable sequence divergence in the extant population, not by sequence-based constraints limiting the transfer of sequence.

One factor that does appear to constrain the positioning of transformation events is the incidence of indels. The exponential distribution of sizes implies that, following the formation of a boundary at one end, the extent of the RSS is determined by an event that occurs with a fixed per base probability of λ_R as it becomes further removed from the initial site. The impact of indels is at least partly governed by whether λ_R applies to recipient bases, donor bases or aligned bases only. The evidence from the exponential decay of selected recombination boundary positions suggests that λ_R applies to each recipient base. However, this would predict that deletions, but not insertions, would be excluded from RSSs, whereas the reverse is actually observed; furthermore, larger deletions would be anticipated to be excluded to a greater extent, which is also not the case. That the observations indicate only large insertions being excluded from RSSs suggests that the actual situation *in vivo* may be more complicated.

Instead, I propose a model in which λ_R applies only to aligned bases, which interact within the heteroduplex resulting from invasion of the ssDNA strand. It seems likely that non-aligned bases in indels loop out of the heteroduplex, and thereby destabilise it to some extent; hence the drop in selected recombinations spanning the IS element insertion in the recipient, despite there being no aligned bases, and the exclusion of indels from the FRs, where the heteroduplex must be more stable for the process of strand exchange to occur. This destabilisation is unlikely to be simply related to the length of the indel, given the range of deletion sizes found within RSSs in this experiment. The exclusion of large insertions from RSSs would not, therefore, be a consequence of the process governed by λ_R, but instead relate to the cleavage of ssDNA as it is imported into the cell. Further investigation will be necessary to test this model.

The exclusion of insertions from RSSs, and the exponential distribution of RSS sizes, has the consequence that, at all loci, if there is a difference is size between alleles then

the smaller allele will transfer between cells more quickly. Hence, in the absence of selection, the smallest allele at any given locus will drift to fixation in a population, meaning homologous recombination has a reductive effect on genome size. In this characteristic, it opposes site-specific recombination, which leads to the integration of mobile elements into the chromosome. These two opposing forces are likely to shape the evolution of pneumococcal lineages.

Table 5.1 Association of indels with *in vitro* recombination events. Comparing the density of donor indels, outside of the primary locus, observed within RSSs and FRs relative to that in sequence that does not undergo recombination. For both insertions and deletions, Fisher exact tests were performed comparing the number of events within RSSs or FRs with those outside of both, and the number of aligned bases between the donor and recipient with the same category with those outside of both. The resultant *p* values are displayed in the columns on the right.

	Outside recombinations	**Within RSSs**	**Within FRs**	***p* value, within RSS**	***p* values, within FR**
Sequence length (bp)	1,646,830	200,193	101,784	-	-
Donor insertions	509	40	20	0.023	0.050
Donor deletions	409	40	11	0.22	0.0033

6 A simple method for directional RNA-seq

6.1 Introduction

One of the drawbacks of the initial RNA-seq studies, relative to microarray work, was the lack of information on the direction of transcription. These protocols sequenced ds-cDNA, thereby masking directionality by showing equal signal on both strands (Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008). This has resulted in a number of techniques being developed in order to retain information on the direction of transcription in the output of such studies. Such data are crucial for resolving overlapping genetic features, detecting antisense transcription and assigning the sense strand for non-coding RNA (ncRNA). These published methods for directional RNA-seq either modify the RNA molecules prior to reverse transcription, through attaching RNA linkers (Lister *et al*., 2008) or by bisulfite-induced cytosine deamination (He *et al*., 2008), or by modifying the first cDNA strand prior to second strand synthesis by adding cytosine residues to the 3' end in a template switching PCR (Cloonan *et al*., 2008). These techniques, all developed for the study of eukaryotic cells, modify the nucleic acid sample prior to the second strand of the duplex being synthesised, thereby allowing reads to be assigned to a specific strand of the genome. However, adding extra steps to any sample preparation protocol increases the risks of sample biases being introduced or exacerbated. Furthermore, the high ribonuclease activity within bacterial cells makes mRNA highly unstable: prokaryotic mRNA typically has a half-life of minutes, whereas in eukaryotic cells such transcripts usually have a half-life on the order of an hour (Rauhut and Klug, 1999). Hence a protocol that minimizes sample manipulation, whilst retaining information on the template strand of transcription, is ideal for studying bacterial gene expression.

6.2 Description and validation of the RNA-seq methodology

6.2.1 Illumina sequencing libraries can be generated from ss-DNA

Sequencing using the Illumina platform requires the ligation of adapters, necessary for PCR amplification, flow cell attachment and sequencing reaction priming, onto

either end of a DNA molecule (Bentley *et al*., 2008). The standard Illumina library preparation protocol requires that samples are prepared in a double-stranded form and subjected to an end repair reaction, using either Klenow to resect 3' overhangs or T4 polymerase to extend from recessed 3' ends to give 'polished' blunt-ended products. These are subsequently 3' monoadenylated and the Illumina adapters, in the form of dimers with a 3' monothymidine overhang, are ligated.

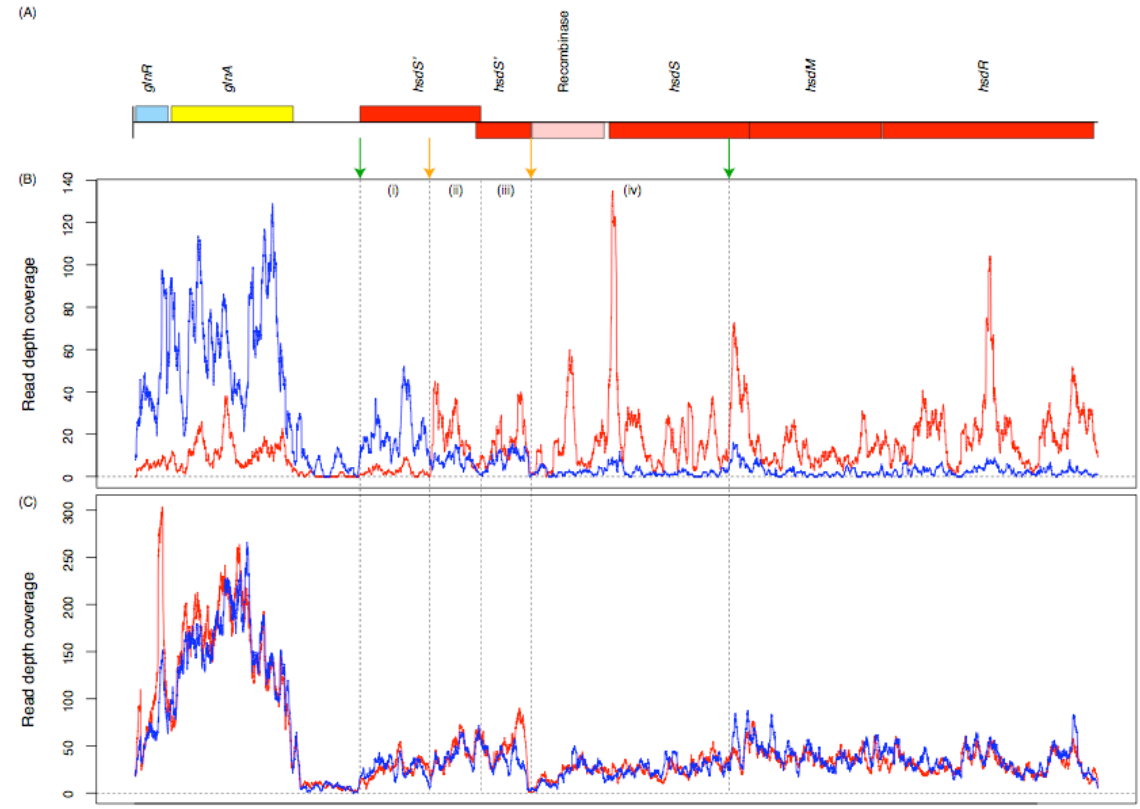


Figure 6.1 Example of the data produced by sequencing ss-cDNA and ds-cDNA. (A) The displayed locus contains the *hsdRMS* operon, which encodes a functional type I restriction system, along with two gene fragments (both labelled *hsdS'*) corresponding to alternative C termini of the specificity subunit, *hsdS*. Also present is a site-specific recombinase, able to invert sequence between two 15 bp repeats, marked with orange arrows, and two 85 bp repeats, marked with green arrows. (B) The data produced by the sequencing of ss-cDNA is shown as Illumina read coverage relative to the annotation. The blue line shows the coverage of reads corresponding to transcription of the top strand of the annotation; the red line shows the data corresponding to transcription of the bottom strand. There appears to be rapid inversion between the 15 bp repeats, hence regions (ii) and (iii) are bidirectionally transcribed. However, inversion between the 85 bp repeats seems to be slower, with no significant antisense transcription observable in regions (i) and (iv). (C) The data produced by ds-cDNA sequencing is shown as read coverage relative to the annotation in the same manner as in (B). In this experiment, the directional information is lost, due to the symmetrical nature of the cDNA being sequenced, but the relative levels of expression of each gene are similar to those in (B).

Unexpectedly, it was found to be possible to produce Illumina libraries from ss-cDNA, generated from *S. pneumoniae* ATCC 700669 RNA. When sequenced, it was

found that such samples retained information on the direction of transcription that generated the template RNA molecule (*e.g.*, Figure 6.1B). Four mechanisms by which ss-cDNA might undergo correct processing to generate Illumina libraries were proposed (summarised in Figure 6.2). The first required the ligation of adapters to the ss-cDNA molecules (Figure 6.2A). This is possible because T4 DNA ligase can ligate ss-DNA molecules, albeit at low efficiency (Kuhn and Frank-Kamenetskii, 2005) (Figure 6.2A, iii), and directionality would be maintained because the second strand is never synthesized (Figure 6.2A, iv). The alternate possibilities involved the formation of duplexes during the end repair reaction (Figure 6.2B–D). Either annealed RNA fragments (the remains of transcripts that served as templates in the reverse transcription reaction; Figure 6.2B, ii) or inter or intramolecular hybridization of cDNA (Figure 6.2C, ii) and Figure 6.2D, ii), were suggested to prime complementary strand synthesis, leading the formation of blunt-ended, double-stranded constructs that could then function as the substrate for the efficient ligation of adapters. If complementary strand synthesis were primed by annealed RNA fragments, this strand would be composed of both RNA and DNA (Figure 6.2B, iii), which cannot be amplified and sequenced by DNA-dependent DNA polymerases. Consequently, only the original ss-cDNA strand would be sequenced (Figure 6.2B, iv). If complementary strand synthesis were primed by intra or intermolecular cDNA annealing, then 3' end processing would produce a reverse complement of the annealed cDNA's 5' end (Figure 6.2C, ii and Figure 6.2D, ii). Hence, sequences with different orientations relative to the original transcript would be segregated into the 3' and 5' regions of the cDNA strands, so by sequencing only the 5' end, all sequence reads maintain the same orientation relative to the original RNA molecule (Figure 6.2C, iv and Figure 6.2D, iv).

In order to determine which of these mechanisms described above occurs during library preparation, a 48 nt DNA oligonucleotide composed of a defined 5' sequence tag and RNA oligonucleotide binding site separated by two stretches of random sequence (Figure 6.3A) was designed. Solutions containing either this DNA oligonucleotide alone, or in the presence of a 12 nt RNA oligonucleotide complementary to the binding site, were subjected to standard Illumina sample preparation and sequencing reactions (see Materials and Methods).

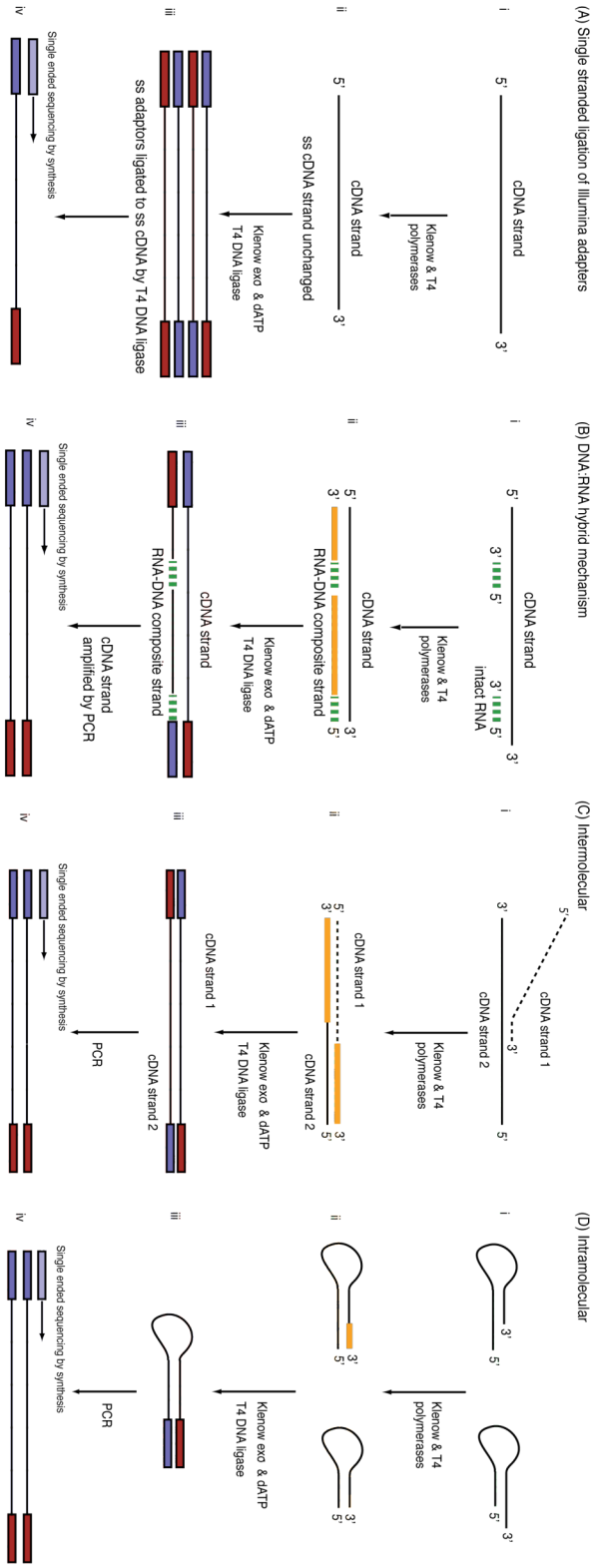


Figure 6.2 The hypotheses proposed to account for the attachment of Illumina adapted dimers to ss-cDNA: (A) attachment of adapters to ss-cDNA, and priming of second strand synthesis by (B) RNA fragments, (C) intermolecular cDNA annealing and (D) intramolecular cDNA annealing.

Libraries were successfully generated both in the presence and absence of the RNA oligonucleotide, demonstrating that adapter ligation did not require RNA-primed complementary strand synthesis. Analysis of 2,162,655 paired 36 nt sequence reads generated from libraries produced in the absence of RNA revealed that in 88% of the DNA molecules, the RNA binding site had been partially replaced by sequence representing the reverse and complement of the known 5' end tag of the 48-mer DNA oligonucleotide (as shown in Figure 6.3C). This indicated that duplexes had been formed through intra or intermolecular annealing followed by processing of the 3' end. The most common species (29% of the sequenced population) had 9 nt of reverse complement of the 5' tag at the 3' end (equivalent to a 9 bp 'duplex length'), which is likely to have arisen from the scenarios outlined in Figure 6.3C.

In cases where more than 12 nt of sequence is generated at the 3' end, the calculated duplex length depends on whether annealing occurs intra or intermolecularly. If annealing is intramolecular, then the reverse complement of the 5' end of the random sequence region is found near the 3' end, resulting in a duplex length greater than 12 nt. This is observed in around a third of cases. However, if intermolecular hybridization occurs, then the reverse complement of the annealed molecule's 5' region is synthesized at the 3' end of the sequenced molecule. In such a case, a duplex length of 12 nt will usually be observed. This is because only the 12 nt 5' tag, common to all molecules, can be identified as having its reverse complement at the 3' end; 3' end processing otherwise replaces random sequence with the reverse complement of another molecule's random sequence. Such a scenario is likely to account for much of the 12% of the sequenced population with a 12 bp duplex length. Similar results are observed when libraries are constructed from the ssDNA in the presence of the RNA oligonucleotide (data not shown). Hence, this shows that Illumina libraries can be constructed from ss-cDNA using standard protocols, with both intra and intermolecular annealing occurring to a comparable extent and contributing to the formation of duplexes during the end repair reaction.

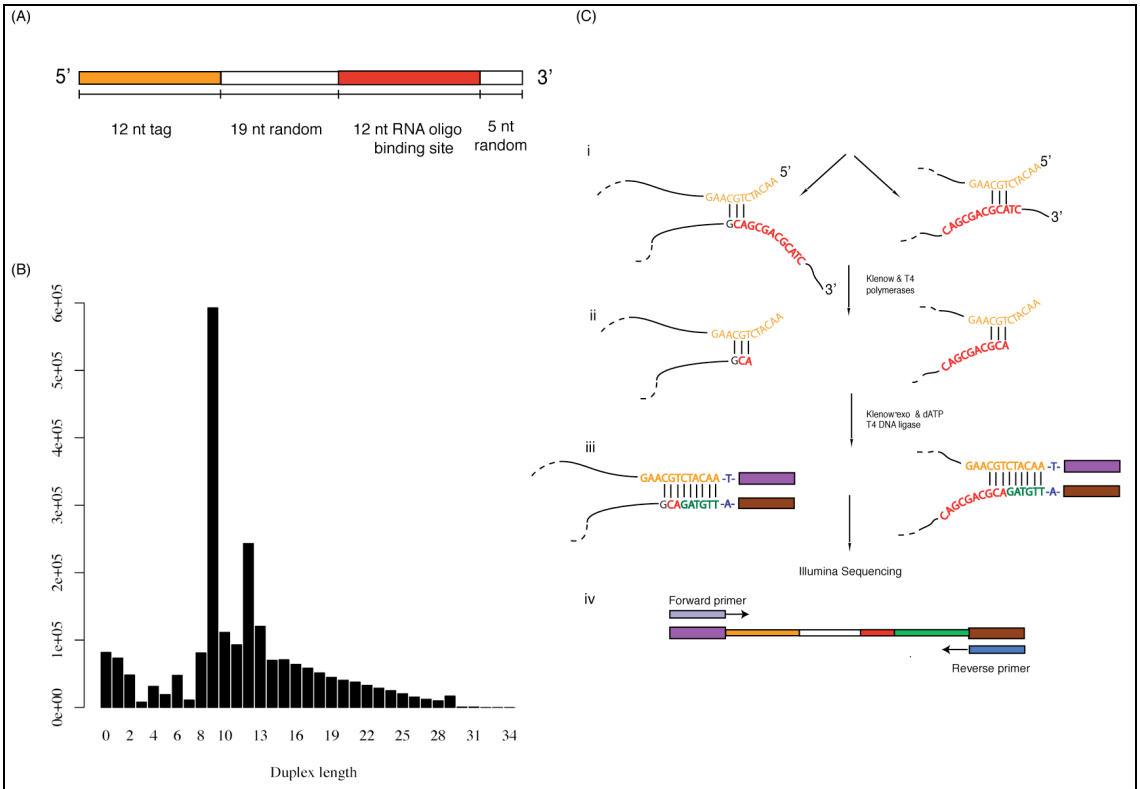


Figure 6.3 (A) Schematic representation of the DNA oligonucleotides from which Illumina libraries were generated. (B) Distribution of duplex lengths amongst a sequenced sample of single-stranded DNA oligonucleotides. Reads corresponding to the oligonucleotide were extracted by searching the output data for the 12 nt known sequence tag. Duplex lengths were then calculated by counting the number of bases at the 3' end found to be the reverse complement of those at the 5' end. This revealed a smooth distribution of values over a range of sizes, with large peaks at 12 bp (likely resulting from intermolecular annealing) and 9 bp (probably the consequence of a 3 bp duplex that can form between the known sequence tag and RNA binding site). (C) Two proposed mechanisms for the formation of species with 9 nt of reverse complementarity between the 5' and 3' ends, the most common duplex length observed. The vast majority of these were found to have 3 bp 'seed duplexes' formed by base pairing between –CGT– in the sequence tag and either the –GCA– in the 3' half of the RNA oligonucleotide binding site, or the CA dinucleotide at the start of the binding site when the preceding nucleotide (the last of the 19 nt random sequence) was G.

6.2.2 ss-cDNA sequencing retains data directionality

An RNA sample from an exponentially growing pneumococcal culture was processed to produce an ss-cDNA sample, half of which underwent second strand synthesis to produce an equivalent ds-cDNA sample. Both of these were then sequenced on the Illumina platform to give 54 nt paired end reads, revealing only the ss-cDNA sample retained information on the direction of transcription (Figure 6.1). The proportion of reads mapping to CDSs in the sense direction varied with the level of putative antisense transcription, the degree to which it overlapped with convergently

transcribed CDSs and the amount of experimental noise, with a median percentage of 87%. This compares to a value of 50% for the ds-cDNA sample.

A high proportion of the cDNA aligns to rDNA sequences (68% in the case of ss-cDNA, 48% in the case of ds-cDNA), which map redundantly to the four almost identical pneumococcal rRNA operons. Hence for the ss-cDNA and ds-cDNA samples, about 24% and 46%, respectively, of the mapping reads could be aligned as pairs in which both read mates could be assigned to unique locations on the chromosome. Of these, 97% of the pairs in the ds-cDNA sample correspond to 'proper pairs', which map as such data are expected to, with the two reads aligning to complementary strands of the genome at sites separated by an insert size congruent with that expected from library construction. The comparable figure for the ss-cDNA sample is 62%; these represent cases where the level of 3' end processing of the cDNA is not sufficient to interfere with the mapping of the reverse read of the pair. The majority of the remainder (36% of the uniquely aligned pairs) mapped to the genome with an insert size greater than 1 kb, with the two reads either mapping to the same strand or complementary strands. This is indicative of intermolecular annealing: the sequence of the reverse read originates from a different cDNA molecule to that of the forward read, resulting in a chimeric molecule that maps to widely separated regions of the genome. The relative orientations of the two reads within the pair is determined by whether the two annealing cDNA molecules were produced from RNA transcribed from the same strand of the genome, or not; correspondingly, half (51%) align to the complementary strand, and the remainder to the same strand. Hence, in accordance with the results of the model oligonucleotide system, intermolecular annealing is observed to occur between cDNA strands during library generation.

Similar fractions of both the ss-cDNA and ds-cDNA samples (1.9%) map to the same strand of the genome to sites within 1 kb of each other. Such an arrangement would be expected as a consequence of intramolecular annealing, with the reverse read sequence originating through reverse complementation of the forward read. That there is no detectable excess of such read pairs when processing ss-cDNA rather than ds-cDNA indicates that intramolecular annealing is not a significant contributory process to library formation in complex transcriptome samples.

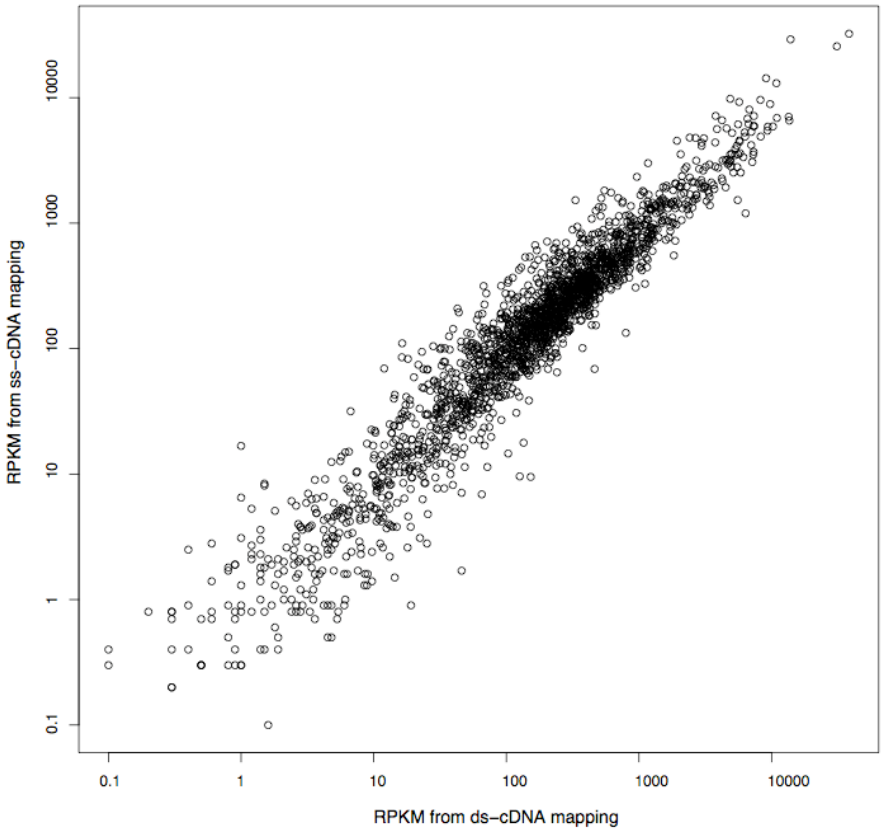


Figure 6.4 Correlation between RPKM values calculated from mapping of ss-cDNA and ds-cDNA to the genome of *S. pneumoniae* ATCC 700669. The RPKM values for each CDS in the genome. with an RPKM greater than zero in both samples, are shown plotted against one another, revealing a tight correlation ($R^2 = 0.84$, $p < 2.2\text{x}10^{-16}$).

6.2.3 ss-cDNA sequencing is quantitative

In order for this technique to be used for quantitative studies of gene expression, the number of reads mapping to a CDS should ideally be directly proportional to its level of transcription. Previous studies have shown that ds-cDNA sequencing is appropriate for quantitative studies of gene expression through comparisons against microarray data (Marioni *et al*., 2008; Mortazavi *et al*., 2008; Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008). In order to validate ss-cDNA sequencing against the results of ds-cDNA sequencing, the levels of gene expression inferred from the ss-cDNA and ds-cDNA samples were compared. For each annotated CDS, the level of transcription was quantified as the number of reads per kilobase of gene length per million mapped reads (RPKM) (Mortazavi *et al*., 2008). Across the genome, the RPKMs calculated from the two datasets correlated tightly (Pearson correlation, $R^2 = 0.84$, $p < 2.2\text{x}10^{-16}$; Figure 6.4), suggesting that the mechanism by which adapters are attached to ss-cDNA do not distort the proportion of sequence reads originating from each gene.

Hence, simply by not synthesizing the second cDNA strand, information regarding the direction of transcription is retained, without affecting the quantitative nature of the data.

6.3 Discussion

This method represents a novel approach for retaining directional fidelity in transcriptomic data. Sequencing ss-cDNA, a technique simpler than the original RNA-seq protocols as it eliminates the need for second strand cDNA synthesis, minimises the number of steps required to process bacterial RNA samples, which are typically more fragmentary than those of eukaryotes. Evaluation of this technique reveals that it maintains the quantitative aspect of sequencing ds-cDNA, crucial for use in gene expression studies. Hence, despite the requirement that cDNA strand anneal into a duplex to allow adapter to be attached, there does not appear to be an appreciable distortion of the relationship between a gene's level of transcription and the number of sequence reads mapping to it. This seems to be because there is little or no sequence dependence in the annealing of cDNA, which is likely to result from the high concentration of DNA in the end repair reaction and the low temperature at which it is conducted (~23°C). Therefore this approach is a simple way to accurately quantify the level, and direction, of expression across the pneumococcal genome.

7 Analysis of pneumococcal small interspersed repeats

7.1 Introduction

Small interspersed repeats, spatially separated genomic regions of similar sequence typically less than 200 bp in length, are frequently found in bacterial chromosomes (Delihas, 2008). These can be classified as either 'simple', when consisting of a single repeated unit, or 'composite', when comprised of a combination of different subsequences arranged in particular patterns (Bachellier *et al*., 1999). For example, a number of enterobacterial species harbour many instances of the simple 127 bp Enterobacterial Repetitive Intergenic Consensus (ERIC) sequence (Hulton *et al*., 1991) and hundreds of composite Bacterial Interspersed Mosaic Elements (BIMEs), which include multiple copies of the Palindromic Unit in a regular configuration (Gilson *et al*., 1991). Similarly, *N. meningitidis* genomes host simple 183 bp AT-rich Repeats and two families of more common, composite elements: 70-200 bp Neisserial Intergenic Mosaic Elements (NIMEs) and Correia Elements (CE), comprised of internal sequences up to 156 bp long delimited by 26 bp inverted repeats (Parkhill *et al*., 2000).

Many such repeat families are likely to be non-autonomous mobile parasitic elements, termed Miniature Inverted-repeat Transposable Elements (MITEs) (Delihas, 2008). These are characterized as being AT-rich, possessing terminal inverted repeats (TIR), having highly base-paired secondary structures and generating target site duplications (TSDs) on insertion. In a number of cases, it has been proposed that repeats are mobilized by the transposases encoded by IS elements within the same host, based on similarities between the TIR of the MITE and the IS sequence. For instance, the Nezha MITE found in cyanobacteria is proposed to be mobilized by IS*Npu*3-like elements (Zhou *et al*., 2008).

The tightly folded secondary structure characteristic of putative MITEs means they can impact on gene expression when they insert into transcribed regions. Some BIMEs, when inserted into operons, have been found to decrease the expression of downstream CDSs through acting as transcriptional attenuators (Espeli *et al*., 2001).

By contrast, regions upstream of ERIC elements integrated into operons may be destabilised by the presence of the repeat when in a specific orientation, as it appears to trigger transcript cleavage through introducing a putative RNase E target site (De Gregorio *et al*., 2005). Similarly, there is evidence that CE act as a target site for RNase III-mediated endoribonucleolytic cleavage when transcribed (Mazzone *et al*., 2001; De Gregorio *et al*., 2002). CE insertions have also been found to influence gene expression through generating functional promoters in *N. meningitidis* (Snyder *et al*., 2009). As well as affecting transcriptional regulation, repeat sequences can alter the sequences of genes without disrupting their function. For instance, in Rickettsia, repeat element insertions have been found in both coding and non-coding genes that appear still to be functional (Ogata *et al*., 2000; Ogata *et al*., 2002).

The first small interspersed repeat to be discovered in *S. pneumoniae* was the BOX element, a composite repeat consisting of boxA and boxC sequences usually separated by a variable number of boxB elements arranged in a tandem array (Martin *et al*., 1992). The variation in different strains' complements of these repeats has allowed them to form the basis of a PCR-based epidemiological typing scheme (van Belkum *et al*., 1996). An early hypothesised function of BOX elements, based on their proximity to a number of genes involved in competence and pathogenesis, was that they might act as regulatory motifs (Martin *et al*., 1992), and subsequent experiments have shown that boxA and boxC elements are able to stimulate the expression of downstream genes, although boxB elements can have an opposing inhibitory effect, depending on their orientation (Knutsen *et al*., 2006). A BOX element has also been hypothesised to increase the frequency of pneumococcal phase variation through affecting the regulation of neighbouring genes (Saluja and Weiser, 1995). Similarity between the TIR of BOX elements and IS*Spn*2, a transposon found in *S. pneumoniae*, has been proposed as the basis for mobilization of these elements. Likewise a second repeat also present in high copy number in the pneumococcal genome, the simple 107 bp long Repeat Unit of Pneumococcus (RUP), has TIR similar to those of IS630-*Spn*1, another transposon commonly found in *S. pneumoniae* (Oggioni and Claverys, 1999). RUP were proposed to preferentially insert into or near IS elements, based on their distribution in a draft of the *S. pneumoniae* TIGR4 genome (Tettelin *et al*., 2001), leading to the suggestion that these elements may

serve to limit the number of functional transposase genes in the chromosome (Delihas, 2008).

7.2 Analysis of small interspersed repeat sequences

7.2.1 Three families of pneumococcal repeats

The curated output of RepeatScout (see Materials and Methods) revealed the presence of three distinct repeat families in the genome of *S. pneumoniae* ATCC 700669. One of these corresponded exactly to the ~107 bp RUP element. Another represented the reverse complement of the 3' end of BOX elements; consequently, to fully define such repeats, independent models for each of the BOX modules were then constructed. The third is a novel repeat element, which shall be referred to as the *Streptococcus pneumoniae* Rho-Independent Terminator-like Element (SPRITE), on the basis of its sequence and predicted secondary structure (Figure 7.1C, Figure 7.2C).

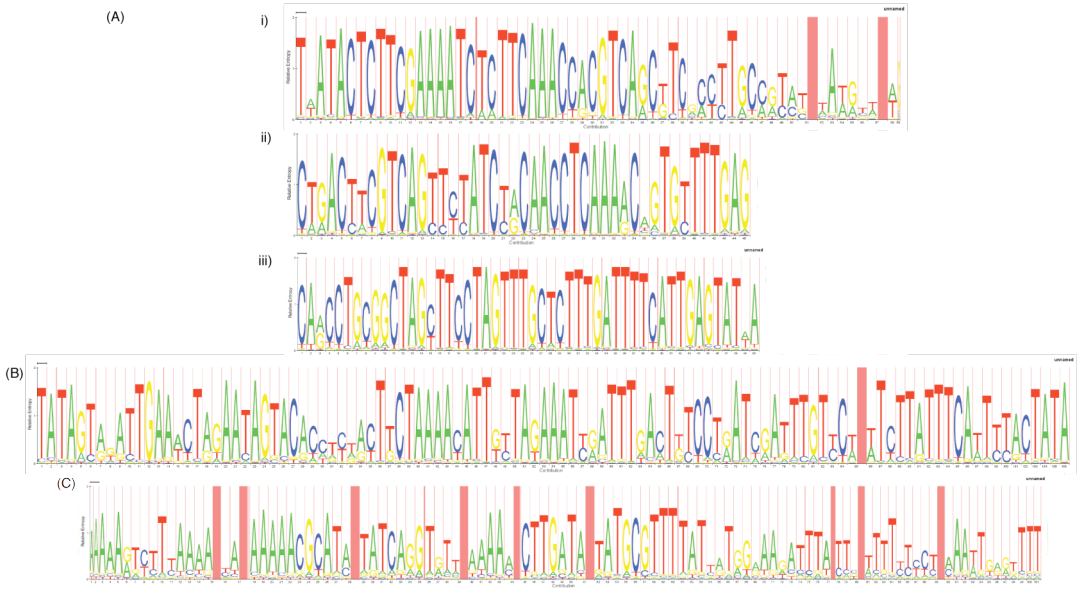


Figure 7.1 HMM logos representing pneumococcal interspersed repeat sequences. These images describe the HMMs used for sequence searches. Each column corresponds to a nucleotide in the repeat element. The total height of bases in each column represents how informative that position is in describing the element; the relative heights of the different bases indicate their respective emission probabilities in the model. Red shaded columns show positions where base insertions occur: the total width of these columns represents the expected number of inserted bases, whilst the dark shaded component indicates the probability that an insertion occurs. The repeats displayed are a) i) boxA, ii) boxB, iii) boxC, b) RUP and c) SPRITE.

Following refinement of the models (see Materials and Methods), the final HMMs used to identify the repeats are represented as logos in Figure 7.1. Overall, 125 BOX (composed of 422 modules), 110 RUP and 30 SPRITE elements were found in the ATCC 700669 genome; in addition, 17 lone box modules were found. All of the original examples used to define BOX and RUP elements were identified by this approach (Martin *et al*., 1992; Oggioni and Claverys, 1999). It seems likely that the lower frequency of the SPRITE repeat is the explanation as to why it was not characterised prior to the availability of complete genome sequences.

Each of the three families of repeats share at least some features of MITEs. All are typically less than 200 bp in length; unsurprisingly, the modular BOX elements are the most variable in size, ranging from 67 bp to 637 bp. Both RUP and SPRITE are AT-rich relative to the *S. pneumoniae* genome (GC content of 39.5%), with mean GC levels of 27.5% and 28.1% respectively. BOX and RUP have been previously shown to have TIR and cause TSDs on insertion (Martin *et al*., 1992; Oggioni and Claverys, 1999; Knutsen *et al*., 2006). SPRITE repeats have comparatively shorter and simpler TIR (the tetranucleotide AAAA and the complement TTTT; Figure 7.1C). Any TSD produced by SPRITE insertions could not be established from the current dataset, because no instances of the repeat with an easily comparable empty site could be found in the available collection of sequences, and no clear evidence could be identified by examining the regions flanking insertions.

All three elements are predicted to form stem-loop structures if transcribed into an RNA form (Figure 7.2). The structure of BOX elements was generated from those elements with a canonical $A_1B_1C_1$ sequence; notably, the folding of the boxB element is predicted to involve few interactions with the boxA and C elements that form the rest of the structure. If this folded RNA is functional, this characteristic may be permissive in allowing boxB to be absent, or present in multiple copies, without causing much disruption to the overall form of the transcript.

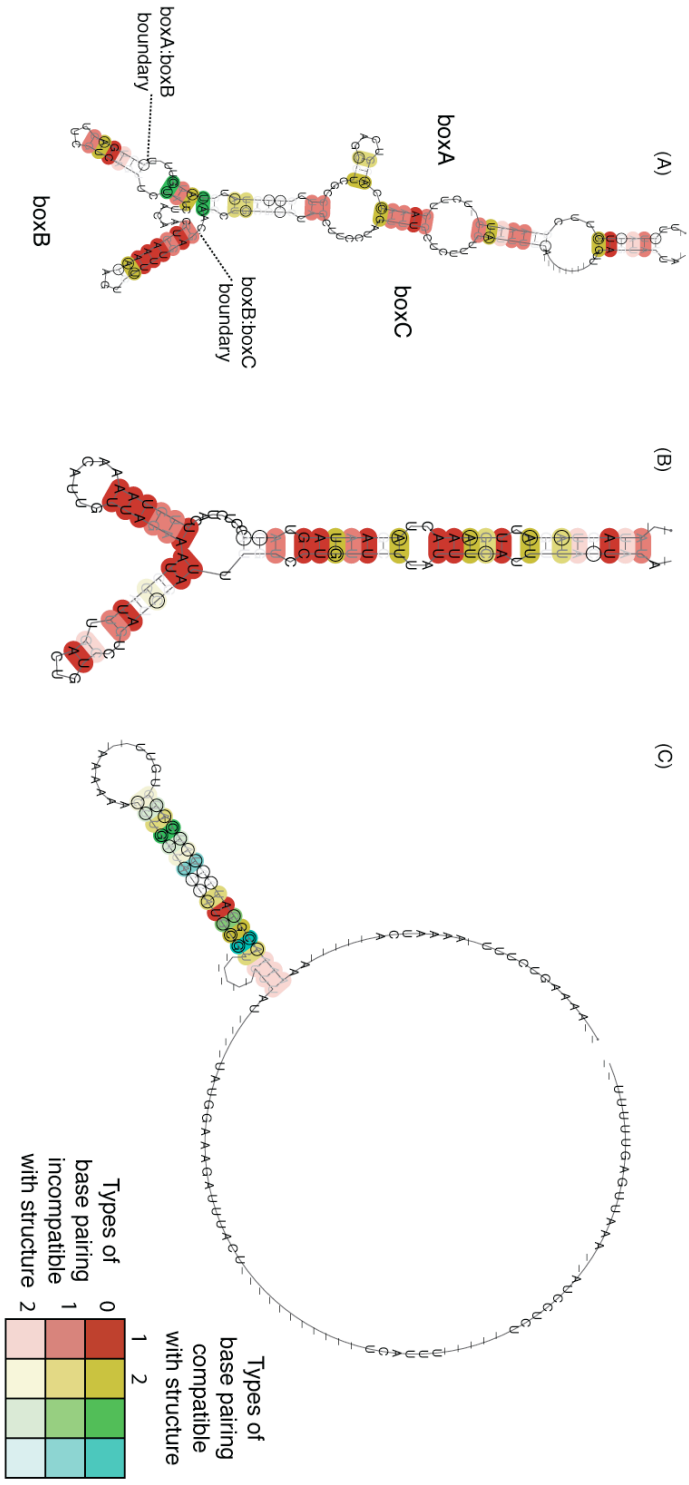


Figure 7.2 Predicted repeat sequence secondary structures. These images represent the predicted secondary structures of transcribed forms of the repeat sequence families: a) BOX, b) RUP and c) SPRITE. The structures were generated from an alignment of 30 sequences from the *S. pneumoniae* ATCC 700669 genome in each case; only BOX elements with a canonical $A_1B_1C_1$ structure were used to produce the structure in a). The boundaries between the different subsequences are marked on the image. Base pairings are coloured according to their conservation in the alignment: the bolder the colour, the more strongly conserved the pairing of the bases, with different colours indicating different numbers of compatible interactions at equivalent sites in the structure (see key).

The SPRITE structure is less tightly folded than that of BOX or RUP, and consists of an 18 bp duplex followed by a relatively uridine-rich (~48% uridine) tract, seeming likely to imbue it with the properties of a Rho-independent terminator. However, the repeat's structure is distinctive in that both the stem duplex and T-rich tract are much longer than the ~10 bp size of both these features in typical streptococcal Rho-independent terminators (de Hoon *et al*., 2005). Hence it appears that SPRITE are distinct from normal Firmicute terminators, although they may be able to function in such a capacity.

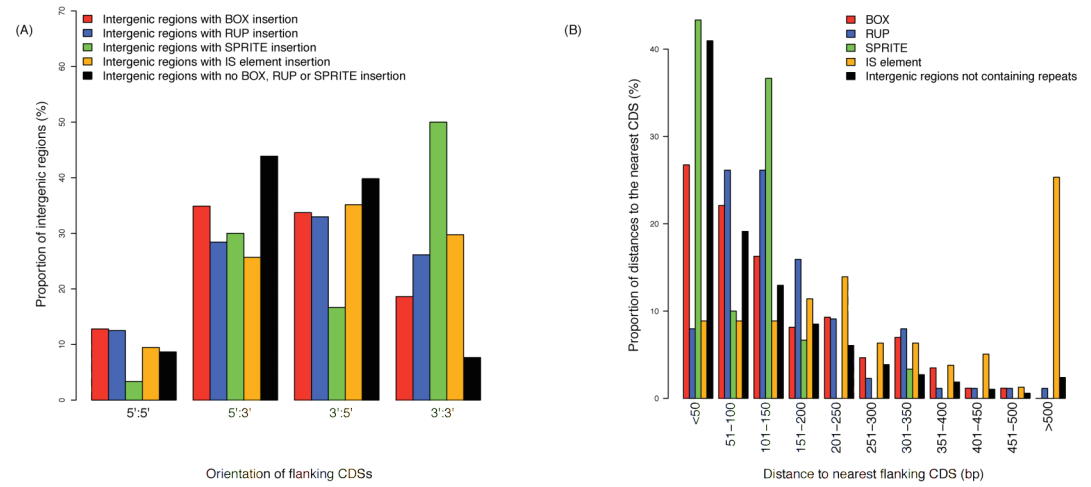


Figure 7.3 Distribution of repeat sequences within the *S. pneumoniae* ATCC 700669 genome. a) The orientation of the CDSs flanking different repeat elements. Data are shown for BOX, RUP, SPRITE and IS elements annotated in the chromosome; the black bars indicate the orientations of all neighbouring CDSs in the genome not separated by intervening small interspersed repeats. b) A comparison of the distance from repeat sequences, including IS elements, to the nearest CDS, and the distribution of lengths of intergenic sequences not containing repeats.

7.2.2 Genomic distribution of pneumococcal repeats

The distribution of these repeats relative to the protein coding genes of *S. pneumoniae* ATCC 700669 was examined. BOX, RUP and SPRITE were all found to mimic the coding bias of the sequence, with 60.8%, 60.9% and 63.3% of insertions on the leading strand of the genome, respectively. Although BOX elements have been found to affect gene regulation (Knutsen *et al*., 2006), they are only slightly overrepresented between divergently transcribed genes, and like RUP, SPRITE and IS elements, they are significantly overrepresented between convergently transcribed genes (Figure 7.3A; Table 7.1). This may be seen as evidence that these elements are mobile,

parasitic entities: the regions downstream of CDS are less likely to be under strong selection pressures, and hence more likely to tolerate repeat element insertions, than upstream regulatory regions or intergenic sequences between cotranscribed genes. Most strongly enriched in these regions are SPRITE, which, given their resemblance to terminator sequences, seem the most probable to disrupt transcription if inserted upstream or between genes.

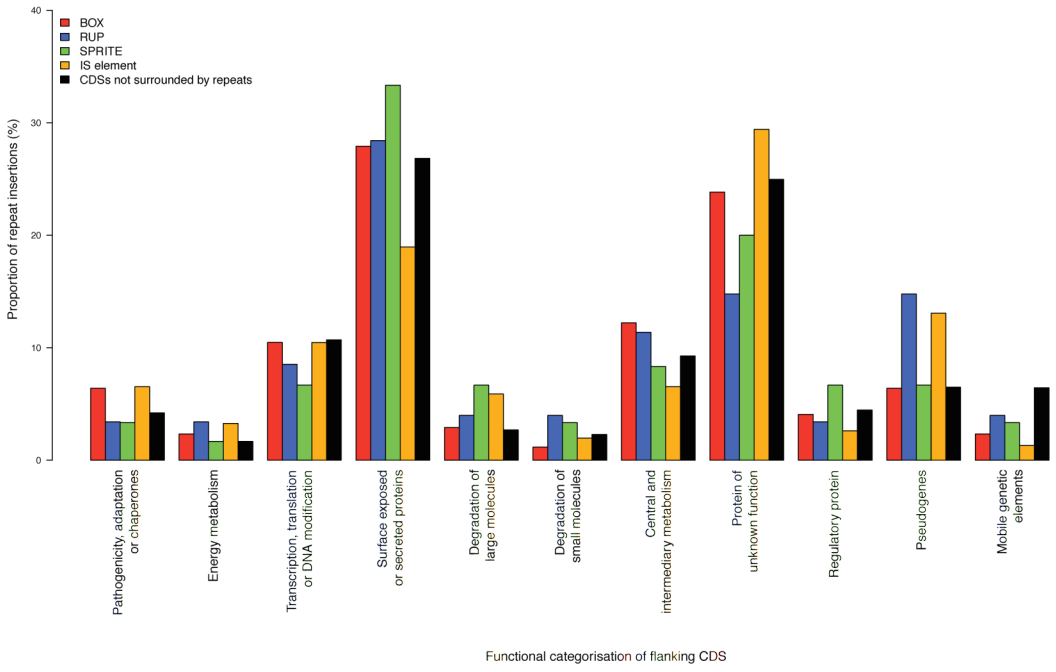


Figure 7.4 Distribution of repeat sequences relative to CDS function. Functional classification of CDSs adjacent to repeat sequences in the genome of *S. pneumoniae* ATCC 700669. For each type of repeat, both flanking CDSs were considered, and the proportions indicated by the graph. The black bars record the equivalent classification of CDSs with at least one associated intergenic region not containing a repeat sequence.

Across the pneumococcal chromosome, the size of intergenic distances follows a gradually decaying distribution (Figure 7.3B). A similar pattern is observed with the distances between BOX elements and the nearest gene, whereas the density of RUP elements is greatest 50-150 bp from the nearest gene. IS elements have an even more pronounced tendency to be distant from neighbouring CDSs; this may reflect the greater potential disruption to gene expression caused by these longer repeats should they insert within, or near, functional transcripts. SPRITE sequences tend to be close to adjacent CDSs, with only one SPRITE found less than 200 bp from the nearest

gene. This enrichment of SPRITE close to the 3' termini of CDS suggests they may have been co-opted by the pneumococcus into acting as functional transcriptional terminators.

Few clear relationships can be ascertained by looking at the association between repeats and the functional classes of their flanking CDSs (Figure 7.4). This again argues against a general role for these repeats as upstream regulatory elements coordinating transcriptional responses to stimuli, as has been previously suggested (Martin *et al*., 1992), because no informative overrepresentation of a repeat near CDSs with a particular function is observed. Furthermore, in agreement with the analysis of the *S. pneumoniae* TIGR4 genome (Tettelin *et al*., 2001), no support for the hypothesised association between IS elements and RUP insertions can be found (Oggioni and Claverys, 1999). There is also no evidence for the positioning of repeat arrays next to genes encoding surface-exposed proteins that may trigger a host response, proposed as a mechanism for promoting horizontal transfer of CDSs for antigenic proteins in *N. meningitidis* (Bentley *et al*., 2007). One apparent association, the preponderance of RUP elements and IS elements adjacent to pseudogenes, seems likely to reflect the tolerance of repeat insertions into regions of the genome that are no longer functional.

Nor is there evidence that the repeats play a role in the positioning of recombination events. Using the *in vitro* transformation data (Chapter 5), the density of BOX, RUP and SPRITE repeats in the aligned regions of the genome not found to participate in recombinations does not significantly differ from that within RSSs (Fisher's exact tests, $p = 0.65$, 1.0 and 0.35 respectively) or FRs (Fisher's exact tests, $p = 0.10$, 1.0 and 0.40 respectively; Table 7.2). Hence there is no evidence of a link between any of the repeats and the positioning of horizontal sequence transfer events. This would appear to be in contrast to the 9 nt DNA uptake sequences (DUS) of *H. parainfluenzae* and *H. influenzae* (Danner *et al*., 1980; Fitzmaurice *et al*., 1984; Smith *et al*., 1999) or 10 nt Uptake Signal Sequence (USS) of *N. meningitidis* and *N. gonorrhoeae* (Goodman and Scocca, 1988; Smith *et al*., 1999), which must be present on the DNA molecule for it to be efficiently passed into the cells of the respective species through the competence system.

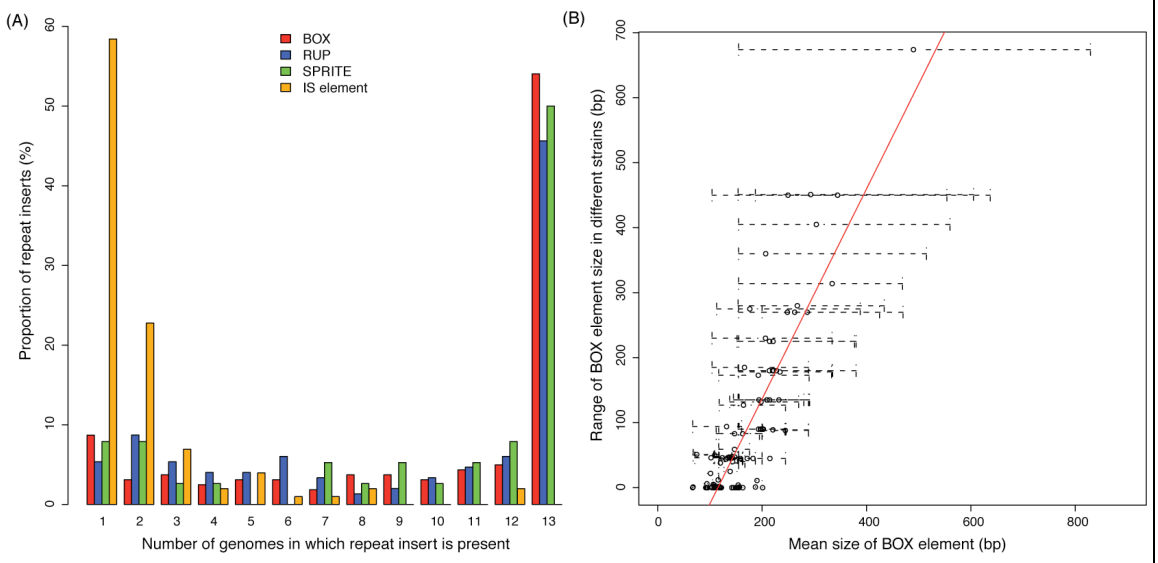


Figure 7.5 Distribution of repeat sequences between pneumococcal genomes. A) For each of the three families of small, interspersed repeat, and for IS elements, orthologous insertion events were defined between chromosomes (see Methods) and scored for presence or absence in the 13 complete pneumococcal clinical isolate chromosomes. The bar chart shows the proportion of insertions of each repeat element shared by a given number of pneumococcal sequences, ranging from insertions present in a single chromosome to those conserved among all strains. B) Variation in the length of BOX elements. For each BOX element identified in the chromosome sequences, the mean, minimum and maximum sizes of orthologous inserts in different strains was calculated. The graph shows the mean length plotted against the range in size, with the dotted bars showing the span of sizes detected.

7.2.3 Mobility of pneumococcal repeats

The level of variation in repeat insertions between all publicly available complete *S. pneumoniae* genomes was also studied (Figure 7.5A). For all three small interspersed repeats, approximately half of the insertions are 'core', *i.e.* present at the same location in all sequenced strains. This contrasts with the distribution of autonomously mobile IS elements, of which the majority of insertions are present only in a single strain. This is likely to reflect IS elements having a comparatively higher transposition rate, while also being removed more quickly by selection. Assuming that the frequency of IS elements in the pneumococcal population is relatively stable over time, this implies that they are much more mobile than the small interspersed repeats. Despite the hypothesized transposition of RUP in *trans* by IS630-*Spn*1 elements, there is no clear evidence from this distribution between genomes that it is more mobile than BOX, which has a lower level of similarity to the TIR of IS*Spn*2 (Knutsen *et al.*, 2006), or SPRITE, for which no significant similarity with pneumococcal IS TIR could be found.

One way in which BOX elements are observed to vary quite considerably is in their size (Figure 7.5B). Several mechanisms have been proposed to explain the fluctuation in the length of tandem repeat arrays, including slipped strand mispairing, unequal crossover during homologous recombination and circular excision followed by reinsertion (Achaz *et al*., 2002). Plotting the mean size of each BOX element insertion against the range of the lengths of the insertion in different genomes reveals a positive linear correlation (Pearson correlation, $R^2 = 0.74$, $p < 2.2 \times 10^{-16}$). This implies that the greater the average number of boxB repeats in a BOX element, the more likely that element is to vary by losing or acquiring these modules. Notably, all BOX elements with a large mean size exhibit considerable variation in length between strains, with none of them stably maintaining an extended form. This result indicates that at the disparate loci at which BOX elements are found, there is significant variation in the rate of mechanisms that change the number of boxB modules in these arrays, or greatly differing levels of selection pressure constraining the size of these composite repeats.

7.2.4 Repeat sequences in other streptococci

The application of the HMMs to the genomes of other nasopharyngeal commensals (*H. influenzae*, *N. meningitidis* and *Staph. aureus*) failed to identify any cases where the repeats had been horizontally transferred. A similar investigation of all publicly available complete streptococcal genomes, encompassing twelve species other than *S. pneumoniae*, also detected few instances of these repeat elements (Table 7.3). The sole representative genome of the most closely related species to *S. pneumoniae*, *S. mitis* B6 (Denapaite *et al*., 2010), contained 104 BOX elements (a mean density of 0.048 kb^{-1}), slightly lower than the mean of 122 in the pneumococcal chromosomes (a mean density of 0.057 kb^{-1}). By contrast, the density of SPRITE sequences in *S. mitis* is about half that of the pneumococcus, and there are only 9 detected instances of RUP in *S. mitis* B6. As *S. mitis* and *S. pneumoniae* are able to exchange DNA, it is not clear whether the repeats were present in their last common ancestor, or whether they have been acquired after speciation and subsequently spread horizontally. By contrast, all three repeat types are almost entirely absent from the genome of *S. sanguinis*, the only other mitis group streptococci to have been sequenced. Hence the most

parsimonious conclusion is that these elements have spread in the pneumococcal chromosome subsequent to the divergence of the more distantly related members of the mitis group.

The only other streptococcal species to have a comparatively high number of detected repeats was *S. suis*, all genomes of which had 11 boxC elements. These were found to coincide with previously discovered repeats, annotated as 'RepSU1', on the complementary strand of the genome in strains SC84, P1/7 and BM407 (Holden *et al*., 2009a). Further analysis revealed the presence of two novel families of BOX-type elements in these genomes, composed of a total of seven different subsequences in particular permutations. One is bounded by boxA and C modules, both of which are around 50 nt long, as are the pneumococcal equivalents. The RepSU1 elements accounted for only the smallest BOX-type repeats of this type, equivalent to A_1C_1 BOX sequences. The other family has a boxE sequence at the 5' end and a boxF module at the 3' end; these motifs are comparatively large, having mean sizes of 115 nt and 133 nt respectively. Both types are found surrounding the same type of intervening boxB modules; however, the boxAC-flanked elements are also sometimes found having boxD modules, always in addition to boxB modules. Hence the diversity of *S. suis* BOX elements appears to be greater than that of the *S. pneumoniae* equivalents.

7.2.5 Genes affected by repeat element insertions

BOX, RUP and SPRITE elements are frequently found together in clusters, and appear to have inserted into one another on a number of occasions. These spatial groupings may reflect a common preference for insertion sites, or a general tolerance of insertions in certain regions of the chromosome. However, repeats are also found interspersed within pseudogenes and regulatory sequences. It is known that BOX insertions can affect the expression of nearby genes (Saluja and Weiser, 1995; Knutsen *et al*., 2006); another example where they might impact on the transcription of an operon is upstream of the *trp* gene cluster. In many Gram positive species, this operon is regulated by two copies of the T box riboswitch, which binds uncharged tRNA. Whilst streptococci have previously been thought to only have a single copy (Gutierrez-Preciado *et al*., 2007), in fact the pneumococcus has two, separated by a

$A_1B_2C_1$ BOX element. This results in the formation of a compound 5' untranslated region nearly a kilobase long, composed of three elements that, given their individually stable structures, seem likely to fold largely independently.

A number of protein coding genes are disrupted by repeat insertions. Instances found in genome annotations include orthologues of the *S. pneumoniae* TIGR4 CDS SP_0243, encoding the extracellular binding protein for a putative iron ABC transporter, which is disrupted by the insertion of a RUP element in all the other pneumococcal genomes except *S. pneumoniae* AP200, 670-6B and TIGR4 itself. However, another CDS encoding part of the same ABC transporter (SP_0241 in TIGR4) is disrupted through frameshift mutations in these three strains. Both of these CDSs appear to be intact in several incompletely sequenced *S. mitis* strains, which lack the alternative *pit2* iron transport system found on Pneumococcal Pathogenicity Island 1 (Brown *et al*., 2001). SPN23F05190 (TIGR4 orthologues SP_0574 and SP_0575), encoding a restriction endonuclease in *S. pneumoniae* ATCC 70069, has a RUP insertion in *S. pneumoniae* TIGR4 and D39, whilst the orthologous gene in *S. pneumoniae* AP200 has been disrupted through the insertion of an IS element. Further examination of the repeat insertions reveals a RUP insertion that has knocked out a serine/threonine protein kinase, previously annotated as two separate CDSs (*e.g.* SPN23F18490 and SPN23F18500 in *S. pneumoniae* ATCC 700669; SP_1831 and SP_1832 in *S. pneumoniae* TIGR4), in all strains except *S. pneumoniae* Taiwan 19F-14 and TCH8431/19A. BOX elements can also cause gene disruption through insertion: a gene encoding a DNA alkylation repair protein is disrupted by a BOX insertion in all the available pneumococcal sequences, whilst an $E_1B_1F_1$ element appears to have inserted into an acetyltransferase pseudogene in the sequenced *S. suis* genomes. Hence the mobility of these repeats has the potential to contribute to phenotypic polymorphism in the *S. pneumoniae* and *S. suis* populations.

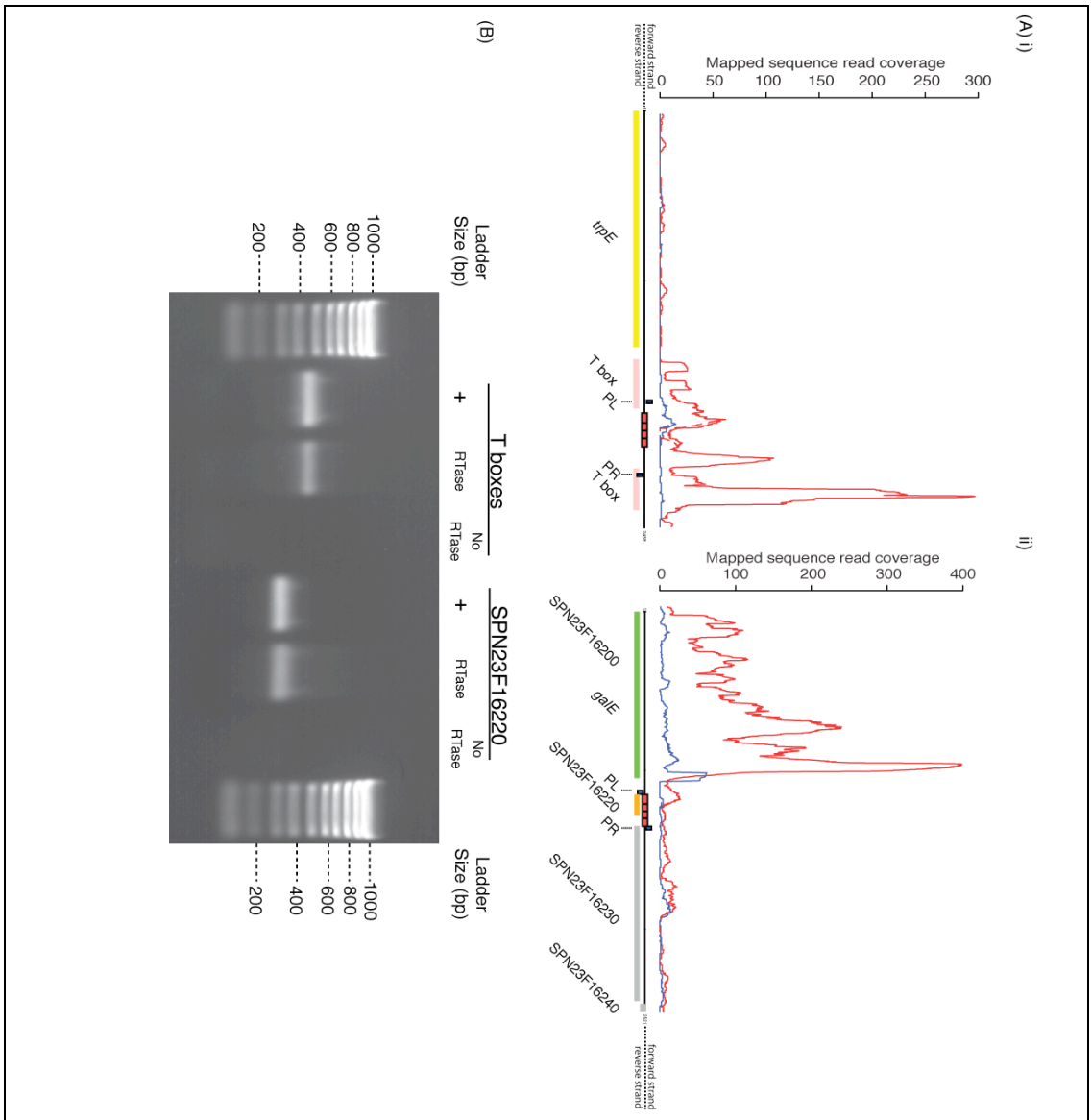


Figure 7.6 Repeat sequence expression congruent with genome annotation. (A) All RNA-seq data is shown as plots of read coverage against the annotation of the represented genomic locus. Along the bottom of these panels, CDSs and non-coding RNAs, coloured according to function as in Chapter 3, are represented as blocks above or below the scale line, depending on their orientation. BOX repeats are shown as red blocks on the scale line. Primer binding sites are indicated by blue blocks labelled using dashed lines. Above the annotation, as part of the coverage plots, blue lines indicate transcription of the upper strand of the genome, while red lines show transcription of the reverse strand. Solid lines represent the result of fully redundant mapping, where reads mapping to multiple sites on the chromosome are randomly distributed between them. Dashed lines represent locally redundant mapping, where reads that might map to regions outside the displayed locus are excluded from the graph (see Materials and Methods). i) The region upstream of the *trp* operon. The *trpE* gene is adjacent to two T box riboswitch motifs separated by an intervening BOX element, represented as four adjacent red boxes representing the $A_1B_2C_1$ structure of the repeat. The RNA-seq data suggests the T box motifs and BOX element are cotranscribed as a composite element, repressing the transcription of the downstream biosynthetic operon. ii) Locus surrounding SPN23F16220. This small CDS is annotated as being encompassed by a BOX element. RNA-seq data suggested it was cotranscribed with SPN23F16230, present on the other side of the repeat relative to the more highly expressed *galE* gene. (B) RT-PCR to confirm transcription of these BOX elements. The positions of the primers used in these reactions are indicated by the blue boxes labelled PL (left primer) and PR (right primer) in (A) i) and ii). In each case, the three lanes correspond to a positive control reaction using a genomic DNA (+), a test using cDNA produced through reverse transcription of an RNA sample (RTase) and a negative control using a non-reverse transcribed RNA sample (No RTase). The bands indicate that these BOX elements are expressed, as suggested by the RNA-seq data.

7.2.6 Expressed open reading frames generated by large BOX elements

Fifty-eight CDSs in the *S. pneumoniae* ATCC 700669 annotation overlap with BOX elements. In 36 cases, this corresponds to the extreme 3' end of a gene, with the BOX repeat encoding the stop codon; in some cases, these correspond to well-characterised genes such as *folE*, *mtlD*, *dnaJ* and *glgP*. However, alignments with non-pneumococcal orthologues do not provide strong evidence for truncation of the encoded polypeptide in any case, especially when the relatively weak conservation of the extreme C terminal portion of proteins is taken into account.

A further 19 sequences, which appear to encode proteins on the basis of GC frameplot and correlation scores (Parkhill, 2002), with little or no functional annotation were found to be mostly, or wholly, encoded by BOX elements. Pneumococcal BOX repeats can extend to over 500 bp in length, and these larger elements tend to encode an open reading frame on both strands. Of the CDSs encoded mainly by BOX sequence, all but two (SPN23F00880 and SPN23F08320) were annotated on the opposite strand of genome to that on which the BOX elements are marked. None of the translated BOX-encoded CDSs exhibited significant similarity with any sequence in the public databases other than matches to hypothetical proteins annotated in mitis group streptococcal genomes.

Directional RNA sequencing data was used to determine whether these genes are expressed. In the case of SPN23F16220 (Figure 7.6A, ii), the transcription follows the direction expected from the annotation, with the BOX element forming a 3' extension to the upstream three CDS operon, as confirmed by RT-PCR (Figure 7.6B). Entirely encompassed within this PCR product is a 42 aa predicted protein encoded by an $A_1B_2C_1$ BOX. Also confirmed to conform to the genome annotation is the BOX element lying between the T box motifs upstream of the *trp* operon (Figure 7.6A, i). The pneumococcal culture from which the RNA was extracted was grown in nutrient-rich conditions, hence the T box motifs are expressed, but the downstream *trp* operon is not. It appears that the riboswitches are still able to function as a regulatory structure, despite the intervening BOX element. Therefore, as anticipated from the genome sequence, BOX elements can be transcribed as extensions to both the 5' and 3' regions of operons.

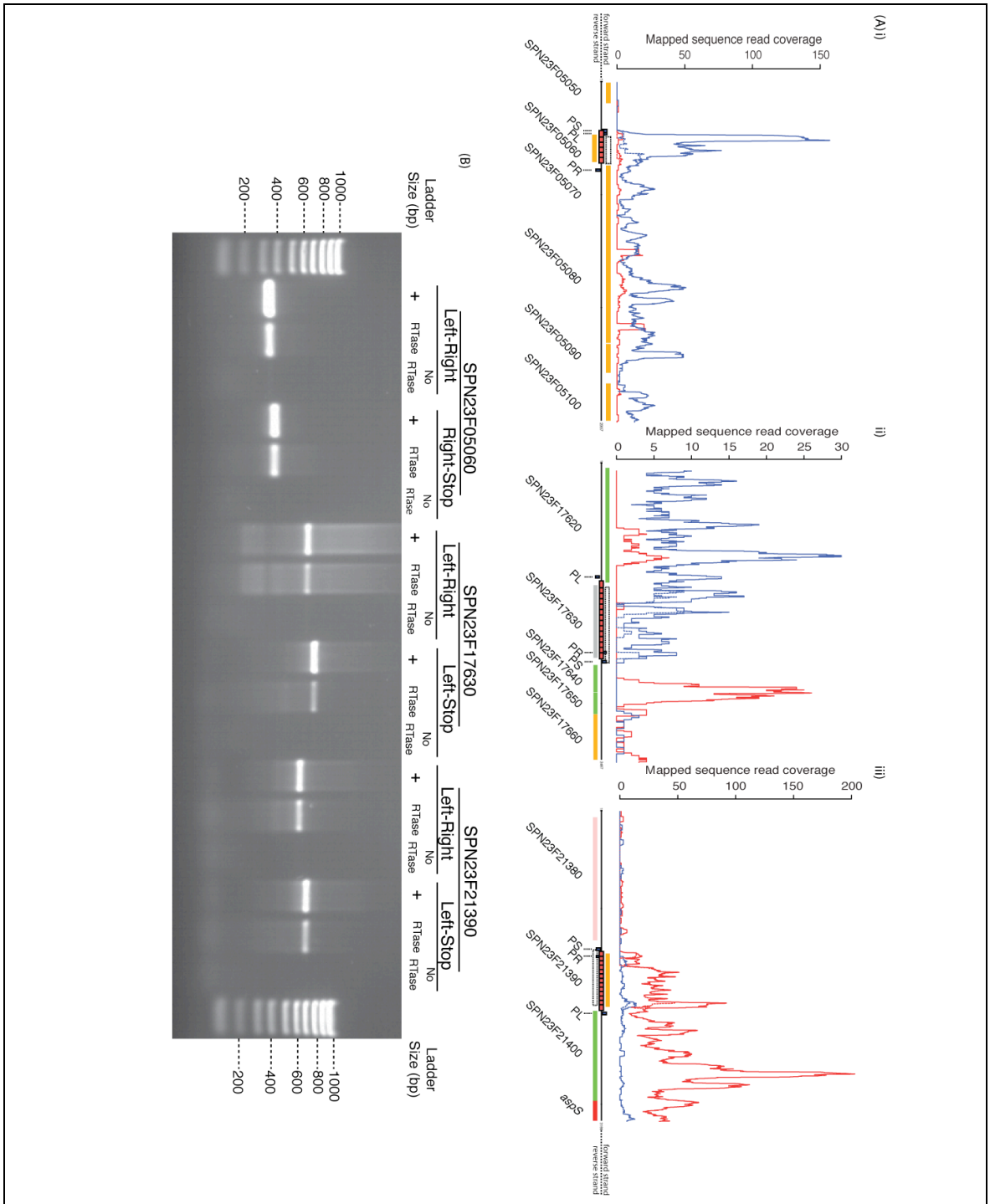


Figure 7.7 Potential misannotation of the *S. pneumoniae* ATCC 700669 genome. RNA-seq data is displayed as described in Figure 7.6. In addition to the published annotation of *S. pneumoniae* ATCC 700669, dashed boxes indicate alternative open reading frames encoded by BOX elements. (A) i) The locus around SPN23F05060, encoded by a BOX element. The CDS is annotated on the bottom strand, but the RNA-seq results indicate it is co-transcribed with the operon on the top strand. ii) This BOX element appears to be cotranscribed with the upstream SPN23F17620 CDS at a low level, rather than encoding the 603 bp putative CDS SPN23F17630. iii) The BOX element encompassing putative CDS SPN23F21390 appears to transcribed on the reverse strand, along with the neighbouring CDSs. (B) For each of the three loci displayed in a), two experiments were performed, each as described in Figure 7.6. One, using PL (left primer) and PR (right primer) tested for expression of the BOX element itself. The second used PL or PR and PS (stop codon primer), which tested whether the full-length open reading frame on the transcribed strand was expressed. At all three loci, both reactions were positive using a cDNA sample as template.

However, in three cases, (SPN23F005060, SPN23F17630 and SPN23F21390), the direction of transcription indicated by the RNA-seq data contradicted the predicted CDS, appearing instead to be continuing from the adjacent operon (Figure 7.7A). SPN23F005060 is contained within a small 289 bp repeat likely to form a 5' extension to the downstream operon. The relatively high density of reads mapping to this BOX element may reflect mismapping of sequences that correspond to a different, more highly expressed repeat (as the level of locally redundant mapping is lower, and hence more congruent with the level of transcription of the rest of the operon), or indicate that the repeat functions as a transcriptional attenuator due to its highly folded structure. The BOX-encoded putative CDSs SPN23F17630 and SPN23F21390 form long (649 bp and 604 bp, respectively) 3' structures. The cotranscription of these elements in the direction indicated by the RNA-seq data was confirmed by RT-PCR in all three examples (Figure 7.7B), implying the annotation is likely to be erroneous.

However, in all three cases, there is also an ORF in the transcribed direction; rather than the start codon being in boxC and boxA encoding the stop codon, as predicted, boxC instead encodes the start codon and the stop codon lies beyond the BOX element. These expressed, BOX-encoded potential CDSs are indicated as dashed boxes in Figure 7.7A. Further RT-PCR confirmed that the RNA extended not just to the end of these BOX elements, but extended as far as the stop codon of these ORFs (Figure 7.7B). However, the proteins encoded by these ORFs also failed to significantly match any sequences other than hypothetical CDSs from mitis group streptococci and lacked good candidate Shine-Dalgarno sequences. Nevertheless, this confirmed that these 5' and 3' operon adducts, formed by BOX elements, have the potential to become nascent protein coding sequences.

7.3 Discussion

The three families of small interspersed repeats found in the pneumococcal chromosome are found, albeit at a reduced frequency, in the closely related species, *S. mitis*, and very infrequently in other streptococci. These include the previously unidentified SPRITE repeat, which resembles a Rho-independent terminator element

in its secondary structure. This is quite unlike the structures of the BOX and RUP elements, which are much more tightly folded and include their TIR hybridised to one another as parts of duplexes. A likely consequence of this form is the observed strong enrichment of this element close to the 3' ends of convergently transcribed CDSs, such that it does not disrupt normal gene expression patterns.

Even the naturally transformable oral streptococcus *S. sanguinis*, also part of the mitis group, lacks these elements. This implies that the repeats are unlikely to fulfil any of the possible important functions that might be ascribed to repeated sequences: for instance, chromosome packaging, aiding with replication or incorporation of horizontally transferred DNA. Furthermore, their distribution within the *S. pneumoniae* ATCC 700669 chromosome, resembling as it does the pattern of IS elements in being enriched between convergently transcribed CDSs, is suggestive of the main alternative explanation of their prevalence: that they are parasitic, non-autonomously mobile elements.

Based on their distribution between different streptococci, it appears that the repeats are likely to have been acquired subsequent to the divergence of the mitis group species. Two possible hypotheses may be advanced to explain the current distribution of repeats in the pneumococcus; one is that they may have been present in the last common ancestor of *S. pneumoniae*, and the position of some repeat insertions in this progenitor subsequently conserved amongst all pneumococcal strains. Alternatively, the repeats may have been acquired by *S. pneumoniae* and then spread horizontally through the population, resulting in the repeats being fixed at certain chromosomal loci over time. This second scenario is likely to be more sensitive to negative selection against the repeat insertions. In either case, a period of relatively rapid spread seems to have occurred in the population's past, which now seems to have abated. The proportion of repeats that are 'core' is similar to the proportion of 'core' CDSs in the pneumococcal pan-genome (Donati *et al*., 2010), and there are few insertions unique to any given chromosome that would indicate recent transposition events, contrasting with the distribution of IS elements between chromosomes.

The only other sequenced streptococcal species to have acquired BOX-type repeats is *S. suis*, which is also able to colonise the human nasopharynx, suggesting there may

be a common source of these sets of elements. Although the *S. suis* BOX elements are present at a lower density in the chromosome, they are more diverse. It is difficult to assess how 'active' these elements are in this species, given the closely related nature of the currently sequenced *S. suis* genomes (Chen *et al*., 2007; Holden *et al*., 2009a), but in the current sample there is little evidence that they are more mobile than in *S. pneumoniae*. Hence in both species, these elements appear to be currently dormant.

One reason to suggest there may be selection against any mechanism that mobilises such elements is the disruption of CDSs by repeat insertion, which is evident in both *S. pneumoniae* and *S. suis*. However, there is also the potential for the formation of novel ORFs by BOX elements. Again, this is observed in both species; as well as the pneumococcal instances, there are two CDSs in the *S. suis* genomes that appear to be intact despite containing box modules (SSUSC84_0055 and 0899 in *S. suis* SC84) and three that are mostly, or entirely, encoded by BOX elements (SSUSC84_0048, 0112 and 0453 in *S. suis* SC84). The RNA-seq and RT-PCR data suggest that in some cases in *S. pneumoniae* such elements are transcribed, and have the potential to become nascent CDSs. Such instances appear to represent the consequences of three proposed properties of BOX elements: firstly, their mobility allowing them to insert into transcribed regions of the genome; secondly, the formation of an open reading frame on both strands of the element, and thirdly, their modular nature allowing them to expand to longer forms.

Whether the polypeptides they encode are actually expressed is not clear; it seems more likely that they are transcribed as untranslated regions. If so, they may influence the levels of expression of co-transcribed genes; those elements forming 3' adducts to operons are likely to form stem-loop structures that may impede the action of 3'→5' exonucleases, the primary RNA degradation pathway in bacteria, thereby stabilising the transcript. However, ERICs are capable of triggering endoribonucleolytic cleavage of transcripts, depending on the orientation of the element and the sequence of the operon, and CE can also trigger cleavage of mRNA. Hence the overall impact of a repeat insertion into an operon is difficult to predict, and is liable to change with the variation in the length of the BOX element and the context of the insertion site. Unfortunately, the sequence read coverage across operons with current RNA-seq

techniques is too inconsistent to make any firm inferences about the impact of these BOX elements.

The simplest mechanism by which these repeats may affect transcription is through acting as terminators, especially given the resemblance of SPRITE sequences to such structures. Such a function has been previously been proposed to be performed by a BOX element (Saluja and Weiser, 1995). There is also a precedent for repeats having a similar potential impact in another nasopharyngeal commensal and pathogen: the USS of *N. meningitidis* which, when found in close proximity to one another, tend to be inversely orientated, allowing them to form a stem loop structure predicted to act as a terminator (Ambur *et al*., 2007). *S. pneumoniae*, although naturally transformable, lacks the selectivity in its uptake of DNA exhibited by *N. meningitidis* and *H. influenzae* (Smith *et al*., 1999), and partial SPRITE sequences were not sufficiently abundant to suggest the element described here is a composite of pairs of motifs analogous to DUS or USS. It seems likely, in fact, that the prevalence of the repeat families present in the pneumococcal chromosome exemplifies a potential disadvantage of the intrinsically competent lifestyle these three respiratory pathogens have adopted: the risk of acquiring genomic parasites that may cause considerable disruption whilst they remain mobile.

Table 7.1 Distribution of repeat elements relative to CDSs. This table shows the results of testing for overrepresentation of repeat sequences in intergenic regions between convergently transcribed CDSs. For each repeat type, the number of insertions in the two different contexts were tested against the number of intergenic sites containing no short interspersed repeats in the same contexts (bottom row). The displayed *p* values were calculated from these 2x2 contingency tables using a two-tailed Fisher exact test.

Feature	**No. upstream of ≥1 CDS**	**No. between convergently transcribed CDSs**	***p* value**
BOX	70	16	0.0017
RUP	65	23	3.4×10^{-7}
SPRITE	15	15	1.7×10^{-9}
IS element	52	22	5.1×10^{-8}
Intergenic sequence	1800	149	-

Table 7.2 Association of repeat sequences with *in vitro* recombination events. The positioning of pneumococcal repeats relative to transformation events observed *in vitro*. Excluding the primary locus, the length of sequence encompassed by RSSs and FRs in the first *in vitro* transformation experiment (Chapter 5), and the length of sequence not found to be part of either are displayed. The frequency of each of the three pneumococcal repeats in these categories is also noted, with Fisher's exact test used to identify any significant enrichment of the elements within, or adjacent to, the recombination events. However, none of the calculated *p* values were significant.

	Outside recombinations	**Within RSSs**	**Within FRs**	***p* value, within RSS**	***p* values, within FR**
Sequence length (bp)	1,646,830	200,193	101,784	-	-
BOX	101	14	11	0.65	0.10
RUP	85	10	5	1.0	1.0
SPRITE	25	1	0	0.35	0.40

Table 7.3 Frequency of repeats in streptococcal genome sequences. This table shows the number of *S. pneumoniae* BOX, RUP and SPRITE repeats, and the number of *S. suis* BOX repeats, found in each of the publicly available complete streptococcal genome sequences.

	***S. pneumoniae* Repeats**					***S. suis* Repeats**					
Genome	RUP	SPRITE	boxA	boxB	boxC	boxA	boxB	boxC	boxD	boxE	boxF
Streptococcus agalactiae 2603V/R	0	0	0	0	0	0	0	0	0	0	0
Streptococcus agalactiae A909	0	0	0	0	0	0	0	0	0	0	0
Streptococcus agalactiae NEM316	0	0	0	0	0	0	0	0	0	0	0
Streptococcus dysgalactiae subsp. equisimilis GGS_124	0	0	0	0	0	0	0	0	0	0	0
Streptococcus equi subsp. equi 4047	0	0	0	0	0	0	0	0	0	0	0
Streptococcus equi subsp. zooepidemicus H70	0	0	0	0	0	0	0	0	0	0	0
Streptococcus equi subsp. zooepidemicus MGCS10565	0	0	0	0	0	0	0	0	0	0	0
Streptococcus gallolyticus UCN34	0	0	0	0	0	1	1	2	0	0	0
Streptococcus gordonii str. Challis substr. CH1	0	1	1	0	1	0	0	1	0	0	0
Streptococcus mitis B6	9	15	104	103	103	0	0	94	0	0	0
Streptococcus mutans NN2025	0	0	1	1	2	0	0	4	0	0	0
Streptococcus mutans UA159	0	0	1	1	2	0	0	3	0	0	0
Streptococcus pneumoniae 670-6B	105	28	123	200	121	0	0	94	0	0	0
Streptococcus pneumoniae 70585	111	31	124	196	121	0	0	97	0	0	0
Streptococcus pneumoniae AP200	101	26	119	163	117	0	0	89	0	0	0
Streptococcus pneumoniae ATCC 700669	110	30	127	183	122	0	0	93	0	0	0
Streptococcus pneumoniae CGSP14	103	29	128	195	122	0	0	93	0	0	0
Streptococcus pneumoniae D39	106	28	117	160	110	0	0	85	0	0	0
Streptococcus pneumoniae G54	102	29	119	170	116	0	0	92	0	0	0
Streptococcus pneumoniae Hungary19A-6	105	30	125	187	121	0	0	95	0	0	0
Streptococcus pneumoniae JJA	102	30	126	189	124	0	0	95	0	0	0
Streptococcus pneumoniae P1031	109	28	124	186	120	0	0	97	0	0	0
Streptococcus pneumoniae R6	106	28	117	160	110	0	0	85	0	0	0
Streptococcus pneumoniae Taiwan19F-14	101	29	118	168	116	0	0	89	0	0	0
Streptococcus pneumoniae TCH8431/19A	100	28	119	172	117	0	0	88	0	0	0
Streptococcus pneumoniae TIGR4	108	25	128	195	127	0	0	97	0	0	0
Streptococcus pyogenes M1	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes Manfredo	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS10270	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS10394	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes	0	0	0	0	0	0	0	0	0	0	0

MGAS10750											
Streptococcus pyogenes MGAS2096	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS315	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS5005	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS6180	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS8232	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes MGAS9429	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes NZ131	0	0	0	0	0	0	0	0	0	0	0
Streptococcus pyogenes SSI-1	0	0	0	0	0	0	0	0	0	0	0
Streptococcus sanguinis SK36	0	1	0	1	1	0	0	0	0	0	0
Streptococcus suis 05ZYH33	0	0	1	0	11	46	43	49	15	33	17
Streptococcus suis 98HAH33	0	0	1	0	11	47	44	49	15	34	17
Streptococcus suis BM407	0	0	1	0	11	47	43	49	15	34	16
Streptococcus suis GZ1	0	0	1	0	11	46	44	49	14	33	17
Streptococcus suis P1/7	0	0	1	0	11	47	44	49	15	34	17
Streptococcus suis SC84	0	0	1	0	11	47	44	49	15	34	17
Streptococcus thermophilus CNRZ1066	0	0	0	0	0	0	0	0	0	0	0
Streptococcus thermophilus LMD-9	0	0	0	0	0	0	0	0	0	0	0
Streptococcus thermophilus LMG 18311	0	0	0	0	0	0	0	0	0	0	0
Streptococcus uberis 0140J	0	0	0	0	0	0	0	0	0	0	0

8 The evolution of serotype 3 ST180 pneumococci

8.1 Introduction

Serotype 3 was one of the earliest pneumococcal capsule types to be identified, first described in 1913 (Dochez and Gillespie, 1913). Bacterial colonies of this serotype have a characteristic mucoid phenotype when grown on an agar plate, due to the long polymeric chains of cellobiuronic acid units that comprise the thick capsule (Heidelberger and Goebel, 1927). For some time, this serotype was treated as a distinct species, named *Pneumococcus mucosus* (Watson *et al.*, 1993), and whilst such a separation cannot be justified on the basis of genetic divergence it does have distinctive epidemiological traits. Unusually for *S. pneumoniae*, the risk of serotype 3 disease increase with age (Gransden *et al.*, 1985; Scott *et al.*, 1996; Inostroza *et al.*, 2001; Harboe *et al.*, 2009), which may relate to the high immunogenicity of the capsule antigen in young children (Douglas *et al.*, 1983). This is likely to explain why serotype 3 has been found to have a high odds ratio for causing disease when comparing infections in adults to carriage in children (Sandgren *et al.*, 2004), whereas this association was not found in similar studies of paediatric disease (Smith *et al.*, 1993; Brueggemann *et al.*, 2003; Sleeman *et al.*, 2006), although a recent similar study in a community vaccinated with PCV7 found serotype 3 to be one of the most invasive capsular types (Yildirim *et al.*, 2010). Furthermore, serotype 3 disease is associated with an increased relative risk of mortality in humans (Gransden *et al.*, 1985; Henriques *et al.*, 2000; Martens *et al.*, 2004; Harboe *et al.*, 2009), and correspondingly strains of this serotype are amongst the quickest to cause death in a mouse model of bacteraemia (Briles *et al.*, 1992).

It has been suggested that these observations may stem from the high frequency with which serotype 3 isolates cause unusual modes of IPD, such as infections of extra-pulmonary sites and broncho-, rather than pulmonary, pneumonia (Finland and Barnes, 1977). Whether these characteristics are the consequence of the capsule or the genetic background itself is difficult to study, because serotype 3 isolates are unusually clonal when compared to other *S. pneumoniae* capsule types. The thick capsule is hypothesized to inhibit the uptake of exogenous DNA by these

pneumococci (Hsieh *et al.*, 2006), thereby enforcing a degree of genetic isolation upon this population. Of the 323 serotype 3 isolates in the multi-locus sequence typing database (Aanensen and Spratt, 2005), the majority belong to a single clonal complex founded by sequence type ST180. This lineage has also been designated as the Netherlands 3-31, or PMEN31, clone (McGee and Klugman, 2011); although it is not associated with penicillin resistance, (Hsieh *et al.*, 2006), many representatives of the lineage are resistant to macrolides (Isozumi *et al.*, 2008). Furthermore, the lineage is geographically highly widespread, having been found across Europe (Enright and Spratt, 1998), Japan (Isozumi *et al.*, 2008) and North and South America (Enright and Spratt, 1998; Beall *et al.*, 2006; Reis *et al.*, 2008). Additionally, it has been observed to increase in prevalence in the USA following the introduction of the heptavalent conjugate vaccine, which does not protect against serotype 3 pneumococci (Beall *et al.*, 2006).

8.2 Genetics of the serotype 3 ST180 population

8.2.1 The genome of *S. pneumoniae* OXC141

The complete genome of *S. pneumoniae* OXC141, a serotype 3 ST180 carriage isolate from a child in Oxford, was generated using a combination of 454 and capillary sequence data. The chromosome was found to be 2,036,867 bp long and contains 1,974 CDSs (including 150 pseudogenes and 58 IS elements), 122 BOX elements, 106 RUP elements and 29 SPRITE repeats (Figure 8.1). The 34-kb prophage ϕOXC141 is the only clearly autonomously mobile genetic element detectable in the chromosome, although there is also a 6.3 kb island that could be a small ICE. Two large genomic islands also distinguish *S. pneumoniae* OXC141 from ATCC 700669: an AT-rich ~22 kb region directly upstream of *pspA* that encodes multiple bacteriocin systems, and a ~25 kb section of PPI-1 (see Chapter 1) which contains a cryptic set of metabolic genes.

In order to perform an analysis analogous to that of the PMEN1 population, six further representatives of ST180 were sequenced using a combination of 454 and capillary sequencing, so as to assess the diversity of the accessory genome, and 77 ST180 and ST505 (a double locus variant of ST180) strains were sequenced as

multiplexed libraries on the Illumina platform to study the demographics of the population (Appendix IV: Serotype 3 strains).

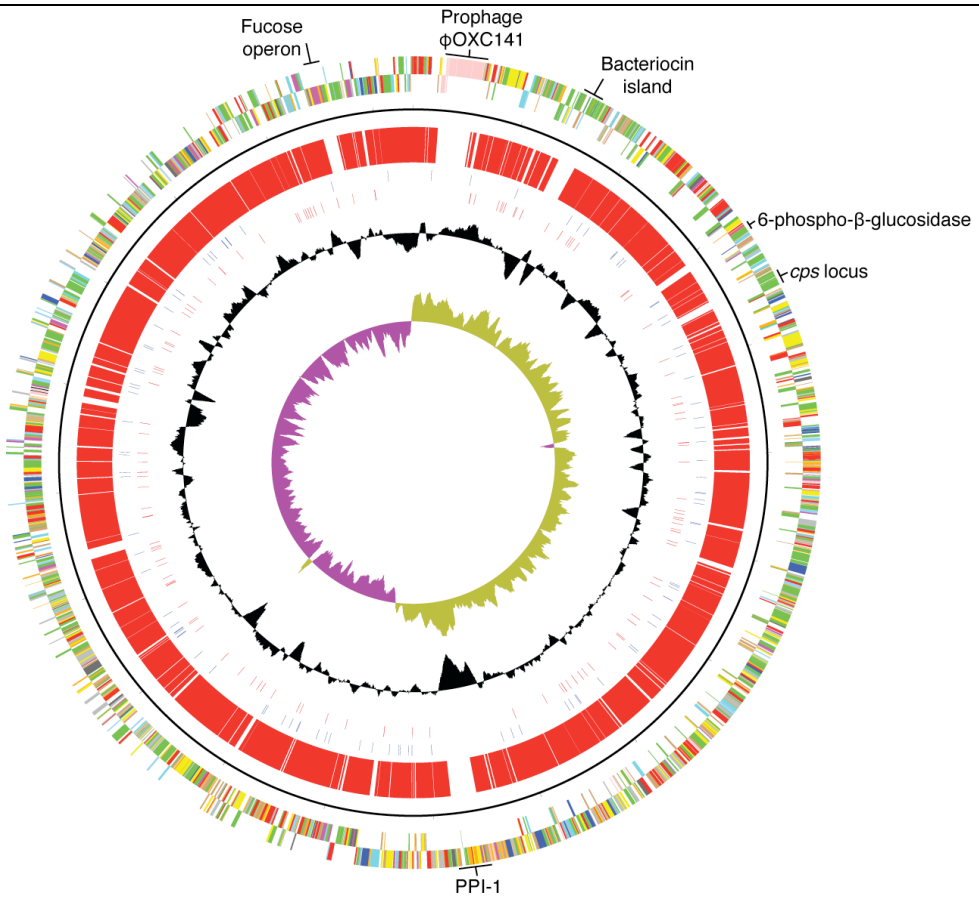


Figure 8.1 Circular diagram representing the *S. pneumoniae* OXC141 genome. The outer rings show the arrangement of coding sequences on the two strands of the genome, coloured according to annotated function (see Chapter 1). The first inner ring shows BLASTN matches to the *S. pneumoniae* ATCC 700669 genome. Moving clockwise from the origin of replication, the major regions of difference are the prophage ϕOXC141, an island encoding bacteriocin systems, the *cps* locus, PPI-1, and the fucose metabolism operon. Moving inward, the rings show the position of RUP repeats (blue), and BOX elements (red). The black graph indicates sequence GC content, and the innermost green and purple graph shows the GC skew of the sequence.

8.2.2 The phylogeny of ST180

The Illumina sequenced strains had a mean mapping depth of 145-fold coverage, each strain having a mean depth of at least 56-fold coverage. Constructing a phylogeny and identifying recombinations for the isolates, as described in Chapter 4, reveals a very different pattern of evolution from that observed for PMEN1 (Figure 8.2). Within ST180, 8,814 polymorphic sites could be identified, dividing the lineage into two groups (clades I and II), with one isolate (BHN609) in an intermediate position.

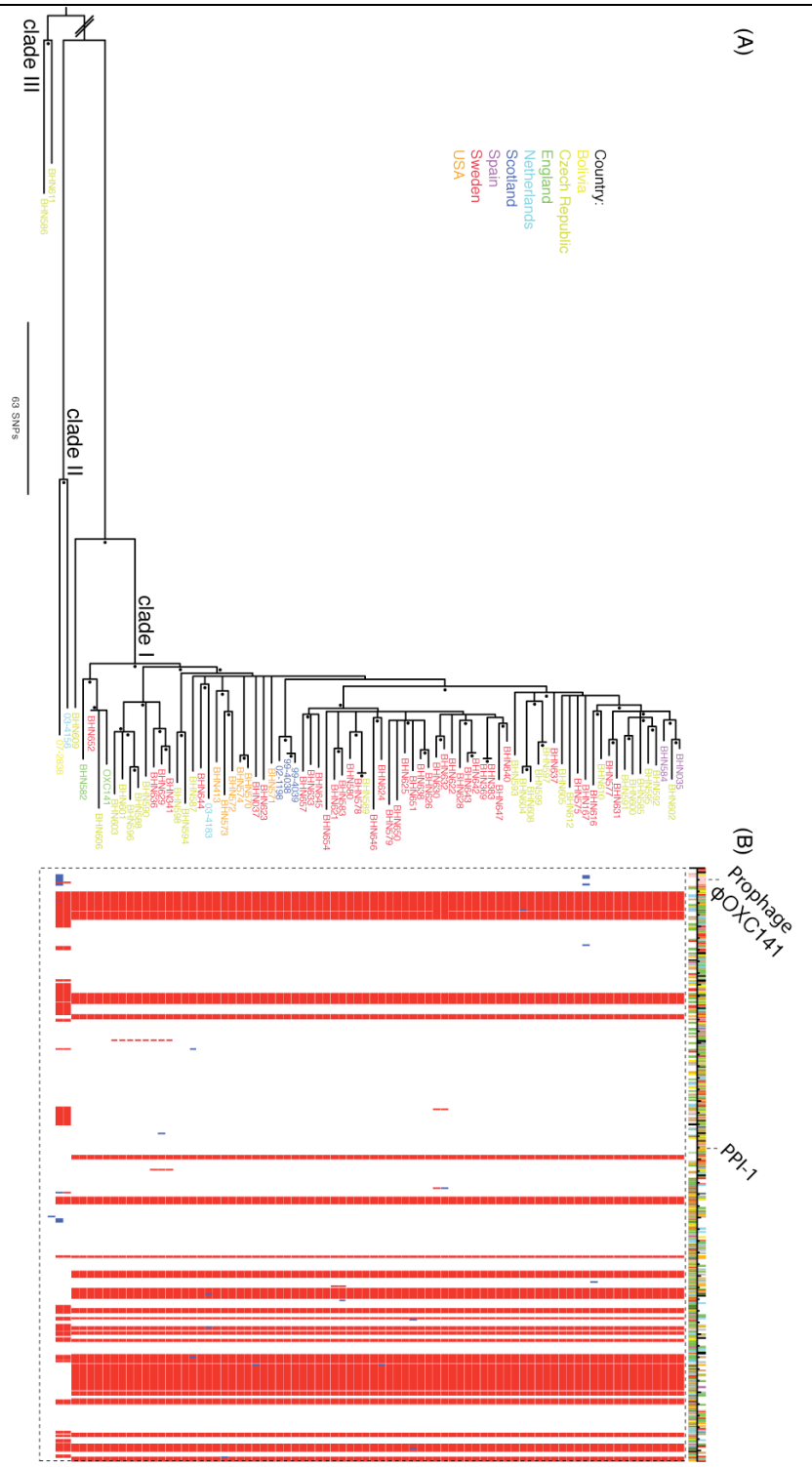


Figure 8.2 Phylogenetic analysis of the serotype 3 isolates. (A) A maximum likelihood phylogeny of the serotype 3 isolates based on the vertically inherited substitutions not introduced through recombination events. The strains are coloured according to their site of geographic origin. Nodes of the tree with a bootstrap support value of greater than 75, from 100 replicates, are marked with a dot. (B) The annotation of the *S. pneumoniae* OXC141 genome is shown, as displayed in Figure 8.1, with prophage ϕOXC141 and PPI-1 marked. The putative recombination events in each taxon, detected by the algorithm described in Chapter 4, are marked relative to this annotation in the panel underneath outlined by the dotted line. Red blocks indicate recombination events reconstructed as occurring on an internal branch, hence are shared by more than one strain, while blue blocks indicate recombinations unique to one particular strain.

The ST505 outgroup, forming the more distant clade III, predicts that the root of the ST180 phylogeny lies between clades I and II. Reconstructing the history of the lineage suggests that 9,025 SNPs have occurred within ST180, 73% of which have been introduced by 92 recombination events. The ratio of transformation events to point mutations is highly heterogeneous throughout the phylogeny, hence within each clade there is little diversity: the branches separating clades I and II account for 15% of the substitutions occurring within the ST180 tree, but over 84% of the polymorphisms imported by recombination events. Despite their infrequency, the 88 homologous recombinations outside the prophage region appear to have a similar size distribution (λ_R = 1.2x10^{-4} bp^{-1}, 95% confidence interval 9.6x10^{-5}-1.4x10^{-4} bp^{-1}; Figure 8.3) to those identified in PMEN1 (λ_R for PMEN1 = 1.6x10^{-4} bp^{-1}, 95% confidence interval 1.5x10^{-4}-1.7x10^{-4} bp^{-1}). Hence the overall *r/m* for ST180 is 2.6, just over a third of that of PMEN1.

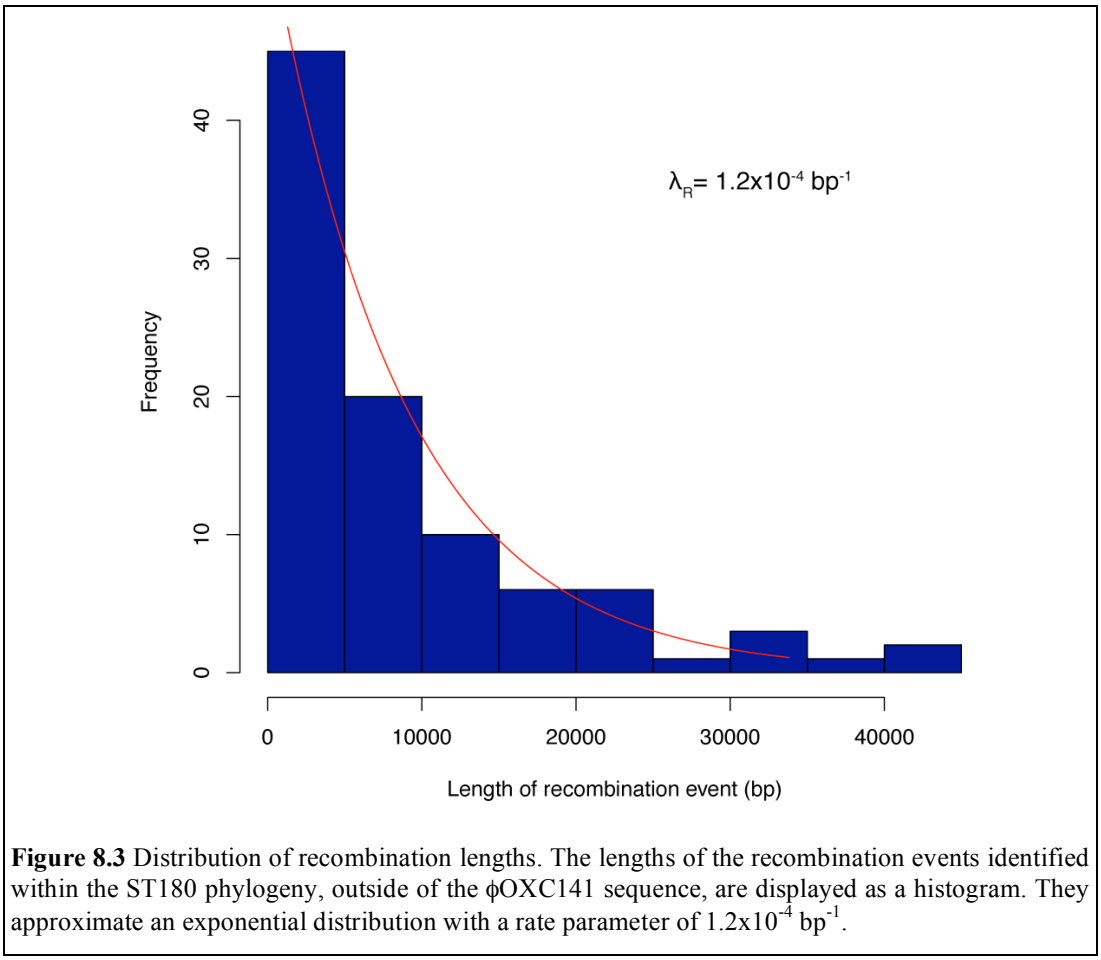


Figure 8.3 Distribution of recombination lengths. The lengths of the recombination events identified within the ST180 phylogeny, outside of the ϕOXC141 sequence, are displayed as a histogram. They approximate an exponential distribution with a rate parameter of 1.2x10^{-4} bp^{-1}.

The range of sampling dates relative to the age of ST180 was insufficient for root-to-tip analyses to provide precise estimates of the mutation rates or time since the last common ancestor, hence a Bayesian analysis of the phylogeny was required to assess these parameters. The last common ancestor of ST180 was estimated to have existed about 144 years ago (95% credibility interval 51-274 years ago), with the emergence of clade I predicted to be associated with a dramatic increase in population size around 51 years ago (95% credibility interval 24-90 years). This implies a mutation rate, across ST180 and ST505, of 8.2×10^{-7} substitutions per site per year (95% credibility interval 2.6×10^{-7}-1.4×10^{-6} substitutions per site per year), somewhat slower than that estimated for PMEN1 (1.6×10^{-6} substitutions per site per year). This may reflect the greater age of the clone allowing more time for selection to purge older disadvantageous mutations from the genotype.

There is no obvious reason for the reduced rate of transformation relative to PMEN1 evident in the genomes of most of the isolates, implying that the mucoid capsule itself is sufficient to enforce a degree of genetic isolation. However, individual isolates are observed to have frameshift mutations in the *comD* sensor kinase, *comFA* helicase and *comEA* transport protein, suggesting the selection pressure for the retention of the machinery may have been weakened (although a *comFA* frameshift mutation was observed in the PMEN1 population). Overall, 178 mutations causing the truncation of CDSs were observed. Based on the Poisson model of mutation events described in Chapter 4, the only functional CDSs to suffer a significantly high number of such disruptive mutations were 1,551 bp SPNOXC10420, encoding a voltage-gated Ca^{2+} channel (five disruptions, $p = 7.5 \times 10^{-7}$), and 1,245 bp SPNOXC12950, encoding a putative flavin mononucleotide reductase (four disruptions, $p = 9.52 \times 10^{-6}$).

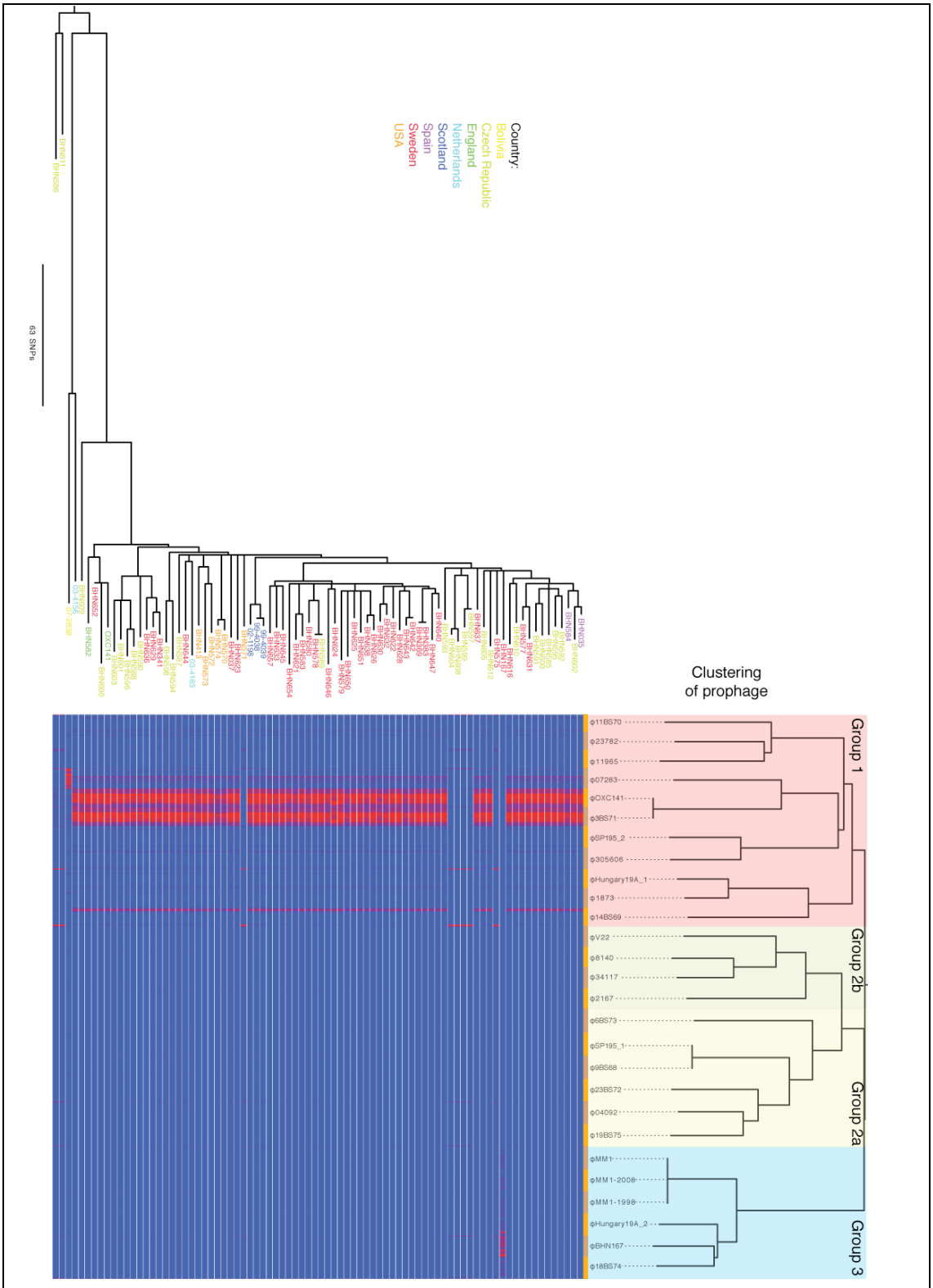


Figure 8.4 Heatmap showing the distribution of prophage sequences between serotype 3 strains. (A) The maximum likelihood phylogeny of the serotype 3 isolates is displayed as in Figure 8.2. (B) All available *S. pneumoniae* prophage genomes, including ϕOXC141, ϕBHN167 and ϕ072838 derived from genomes sequenced as part of this study, are clustered and grouped based on gene content as described in Materials and Methods. The alternating orange and brown bars underneath the clustering indicate the extent of each prophage. (C) A heatmap of Illumina sequence read coverage of the prophage sequences is displayed for each taxon. Blue indicates low coverage, and red indicates high coverage.

Clades I and II also differ quite extensively in their accessory genome content. Notable in clade I is the widespread presence of prophage ϕOXC141, which appears to have been acquired by an ancestor of the clade and subsequently deleted on three independent occasions in the sampled strains (Figure 8.4). Heightened coverage of this element in the Illumina sequence data suggests it may be capable of active replication (Figure 8.5), which, along with the homoplasic nature of its removal, indicates the element is likely to be detrimental to the host. One of the taxa to have lost ϕOXC141, BHN167, is the only strain in clade I showing evidence of having acquired a novel prophage. In clade II, *S. pneumoniae* 03-4156 has a prophage sufficiently closely related to ϕOXC141 that it may represent *in situ* modification of that element, while *S. pneumoniae* 07-2838 has gained a very distinct prophage. Overall, the flux of such elements in the ST180 population appears to be much slower than in PMEN1.

Also found to vary between the two clades are some metabolic and bacteriocin synthesis operons, and the protein antigens PspA and PspC (Figure 8.6). The most parsimonious explanation of the patterns of variation appears to be acquisition of accessory loci during the diversification between the two clades, followed by infrequent deletion of islands over the shorter timescales within each clade, although only clade I is sampled with sufficient density to study this pattern in detail. Just one instance of antibiotic gene acquisition is evident in the population. This is an insertion of Tn*916*, carrying the *tetM* tetracycline resistance gene, into the Bolivian strain *S. pneumoniae* 07-2838. There is no evidence for import of macrolide resistance cassettes, so frequent in PMEN1, or acquisition of other antibiotic resistance genes. Furthermore, the fluoroquinolone resistance polymorphisms within the topoisomerase genes *gyrA*, *gyrB*, *parC* and *parE* that are observed to be homoplasic in the phylogenies of PMEN1 (Chapter 4), *Staph. aureus* ST239 (Harris *et al*., 2010) and *Salmonella* Typhi (Holt *et al*., 2008) are not evident at all in ST180.

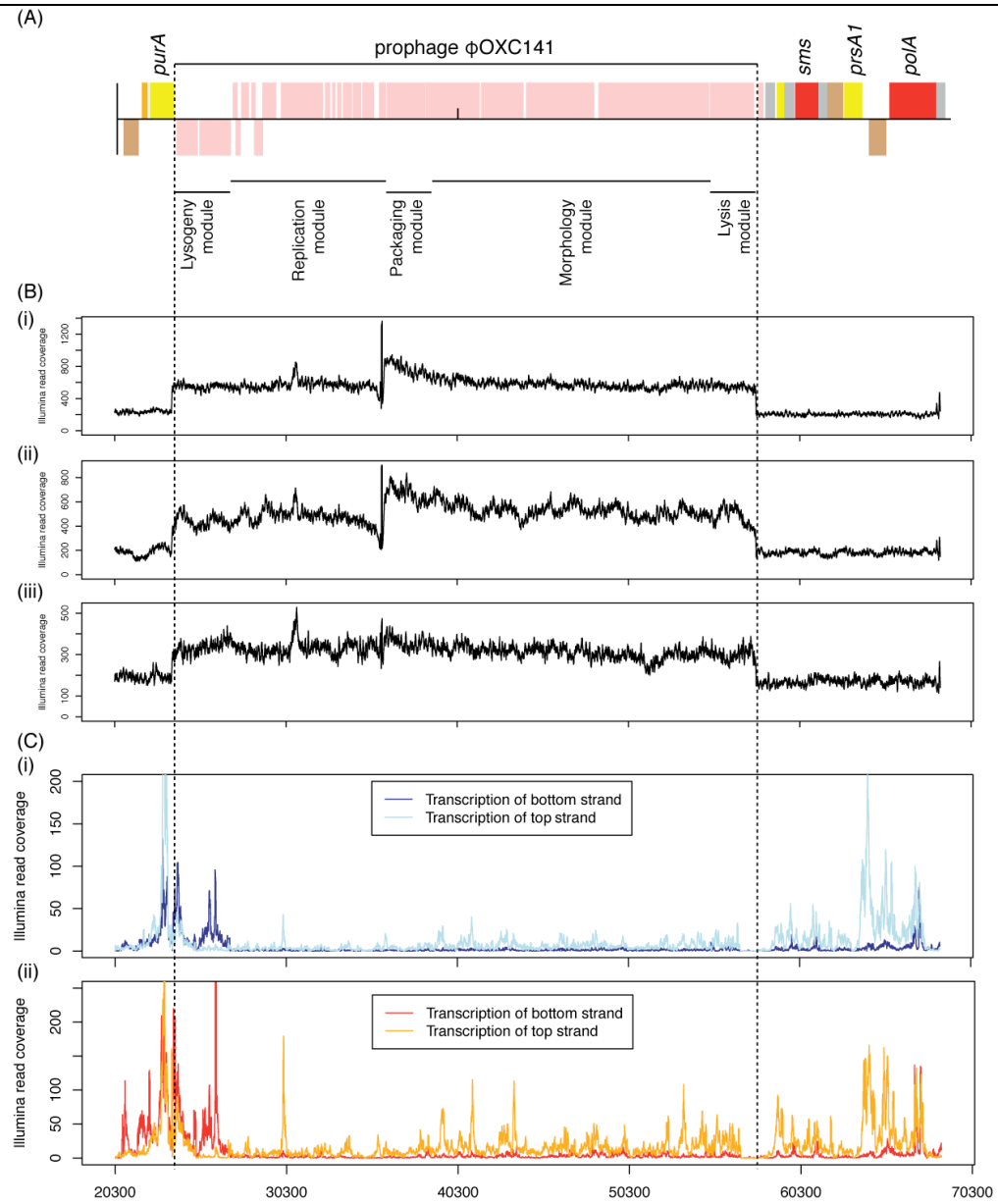


Figure 8.5 Activity of prophage ϕOXC141. (A) Detailed annotation of ϕOXC141 and the flanking genomic regions. The modular organisation typical of pneumococcal prophage is marked relative to the viral sequence. (B) Mapping of Illumina genome sequencing data from strains (i) BHN640 (ii) BHN035 and (iii) BHN598, phylogenetically disparate within clade I. The two- to three-fold increased coverage of ϕOXC141 suggests it is actively replicating. (C) Prophage transcription. The RNA-seq data from (i) 99-4039 and (ii) 99-4038 indicate that in some cells, the lysogeny module is active, keeping the prophage dormant, while in others, the modules required for replication and host cell lysis are expressed.

One SNP that does cause antibiotic resistance, a polymorphism upstream of the ABC-type efflux pump *patAB*, is found distinguishing the closely related sister taxa *S. pneumoniae* 99-4038 and 99-4039. These strains were isolated from a single case of meningitis: 4038 was taken from the bloodstream, and 4039 subsequently from the CSF. The maximum likelihood phylogeny predicts that these are distinguished by just

three polymorphisms, all of which are of the ancestral allele in the isolate from the bloodstream, and of the derived form in the CSF strain. This suggests the mutations have arisen within the patient during, or prior to, the invasive disease. Manual inspection of the assembled sequences confirmed the only other differences between the strains were non-synonymous changes to a LysR-domain regulator and a haloacid dehalogenase-type hydrolase in addition to the mutation upstream of *patAB*.

8.3 Comparison of *S. pneumoniae* 99-4038 and 99-4039

8.3.1 Transcriptional profiles of *S. pneumoniae* 99-4038 and 99-4039

Preliminary work comparing the transcriptional profiles of 99-4038 and 99-4039 using a microarray based on the genome of *S. pneumoniae* TIGR4 was performed by Donald Inverarity of the University of Glasgow (Inverarity, 2009); this revealed significant differences in their patterns of gene expression (Table 8.1). Sequencing of three paired RNA samples from the two strains grown *in vitro* were then used to characterise these differences more precisely. This experiment broadly agreed with the conclusions from the microarray work (Table 8.2, Table 8.3). Analysis of differential gene expression revealed that 53 CDSs had a significantly altered pattern of transcription. These could be grouped into 11 gene clusters and 17 singleton CDSs. Both the microarray and RNA-seq data found the *patAB* genes to be upregulated in *S. pneumoniae* 99-4039, with both CDSs apparently cotranscribed despite the intervening degenerate transposase sequence being encoded on the complementary strand of the genome (Figure 8.7).

The RNA-seq data also showed that the *pur* purine biosynthesis operon, adenylate kinase and guanine synthase were also more highly expressed in *S. pneumoniae* 99-4039, indicating a change in nucleotide metabolism. A similar rise in transcription is observed for the translation initiation protein genes *infA* and *infC*. By contrast, the translation elongation factor Ts gene *tsf* and the cotranscribed *rpsB* ribosomal protein CDS are expressed at a lower level in *S. pneumoniae* 99-4039, according to both RNA-seq and microarray data. RNA-seq data also indicates that the chaperones *dnaK*, *grpE* and *clpL* are transcribed at a lower level in 99-4039. Congruent with the DNA sequence coverage mapping, active transcription of both the ϕOXC141 lysogeny

module, and the prophage's genes for the replication and host cell lysis, was observed in the two strains (Figure 8.5C), suggesting a heterogenous state in which the phage is induced in some cells and dormant in others.

8.3.2 Phenotypic differences between *S. pneumoniae* 99-4038 and 99-4039

Surveys of clinical isolates have found variable levels of *patAB* expression (Garvey *et al.*, 2010). Those strains with elevated levels of expression are observed to have reduced susceptibility to fluoroquinolones and other antimicrobials, including linezolid (Feng *et al.*, 2009), reserpine (Garvey and Piddock, 2008), acriflavine, berberine, ethidium bromide (Robertson *et al.*, 2005) and the dye Hoescht 33352 (Garvey *et al.*, 2010). Expression of the efflux pump is normally induced by DNA damaging agents, including fluoroquinolones and mitomycin C, suggesting they may be part of a wider stress response that includes the induction of competence (El Garch *et al.*, 2010). In order to test whether such phenotypic differences between *S. pneumoniae* 99-4038 and 99-4039, the strains' resistance to a variety of antimicrobial compounds was tested using phenotype microarrays using the Omnilog platform (see Materials and Methods).

Nine arrays were used to test for susceptibility to a range of antimicrobial compounds, each at four different concentrations, and a tenth examined responses to a number of osmolytes. All 21 cases where a statistically significant difference between the strains could be detected resulted from *S. pneumoniae* 99-4039 exhibiting greater resistance to antimicrobial agents (Table 8.4), with the increased expression of *patAB* not appearing to affect the bacterium's membrane integrity, based on the osmolyte assays. The greatest difference was observed in the presence of the purine analogue 6-mercaptopurine, which caused a significantly greater inhibition of *S. pneumoniae* 99-4038's metabolism at three of the concentrations tested. As well as confirming the protection offered against acriflavine, the upregulation of *patAB* also appears to protect against the related compound proflavine and the nucleotide analogue 5,7-dichloro-8-hydroxyquinoline. Other planar, aromatic molecules with nitrogen substitutents causing differential respiration included three uncouplers (pentachlorophenol, crystal violet and 2,4-dinitrophenol) and three drugs prescribed as antidepressants (atropine, orphenadrine and amitriptyline).

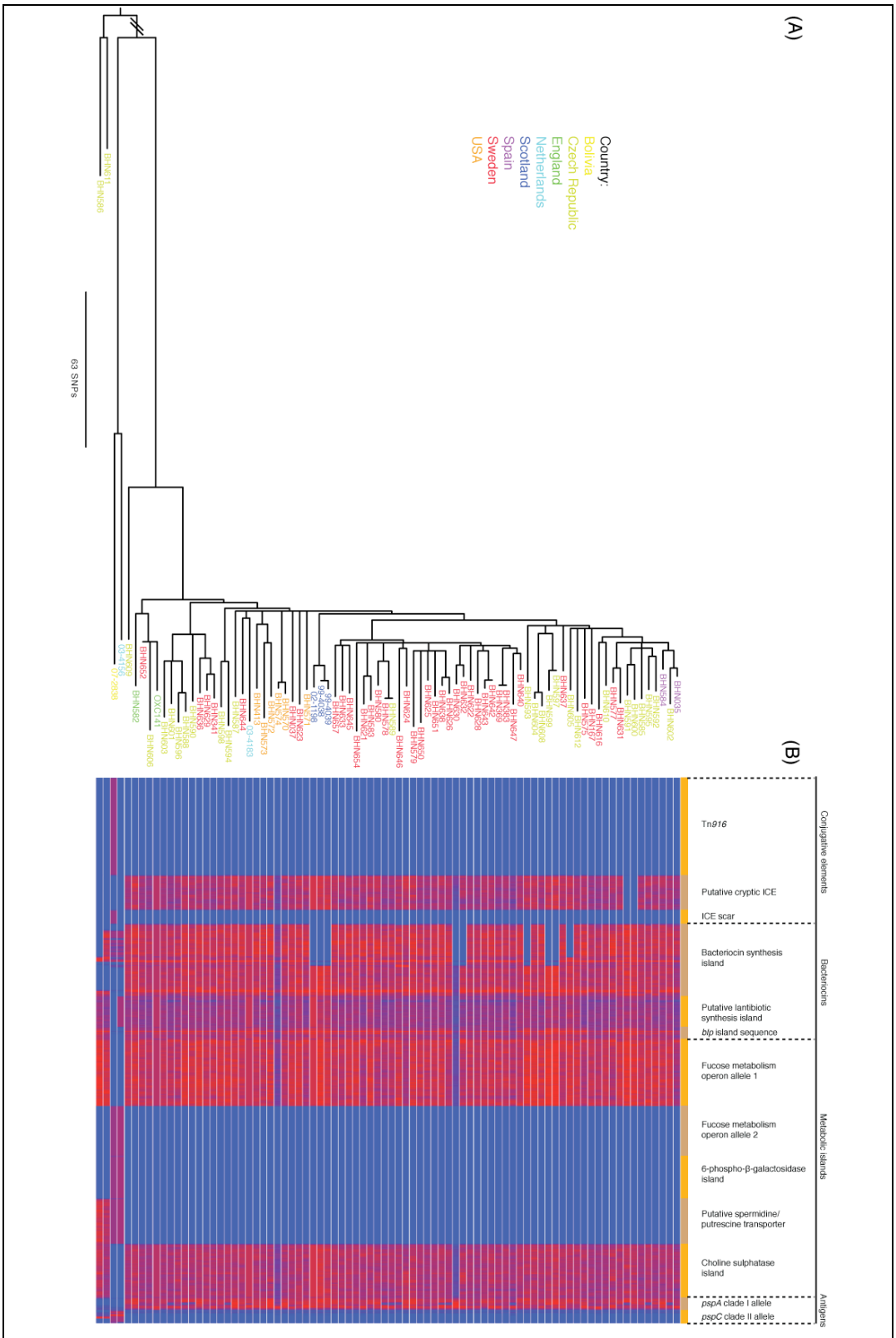


Figure 8.6 Distribution of accessory genome loci. (A) The maximum likelihood phylogeny, as displayed in Figure 8.2. (B) The accessory genome loci, identified through alignment of the strains sequenced using 454 and capillary data, are displayed as alternating orange and brown blocks, grouped according to function. Underneath, a heatmap for each taxon indicates the level of Illumina sequence read mapping along these sequences. Blue indicates low levels of mapping, while red suggests higher levels of mapping.

The long chain quaternary ammonium compounds dodecyltrimethyl ammonium bromide, domiphen bromide and cetylpyridinium chloride were also found to cause lower respiration by the blood isolate by the screen. In this context, it is interesting to note that the one CDS highlighted as being co-conserved among genomes with *patAB*, SPNOXC06720 (Szklarczyk *et al*., 2011), which is also upregulated in the CSF strain (Table 8.2), contains a DegV domain. This motif appears to bind fatty acid chains, such as those present in these quaternary ammonium cations, and, from its similarity with related domains, may interact with the transporter. Hence this may explain the structural difference between these molecules and the others found to cause significant differences in metabolism between *S. pneumoniae* 99-4038 and 99-4039.

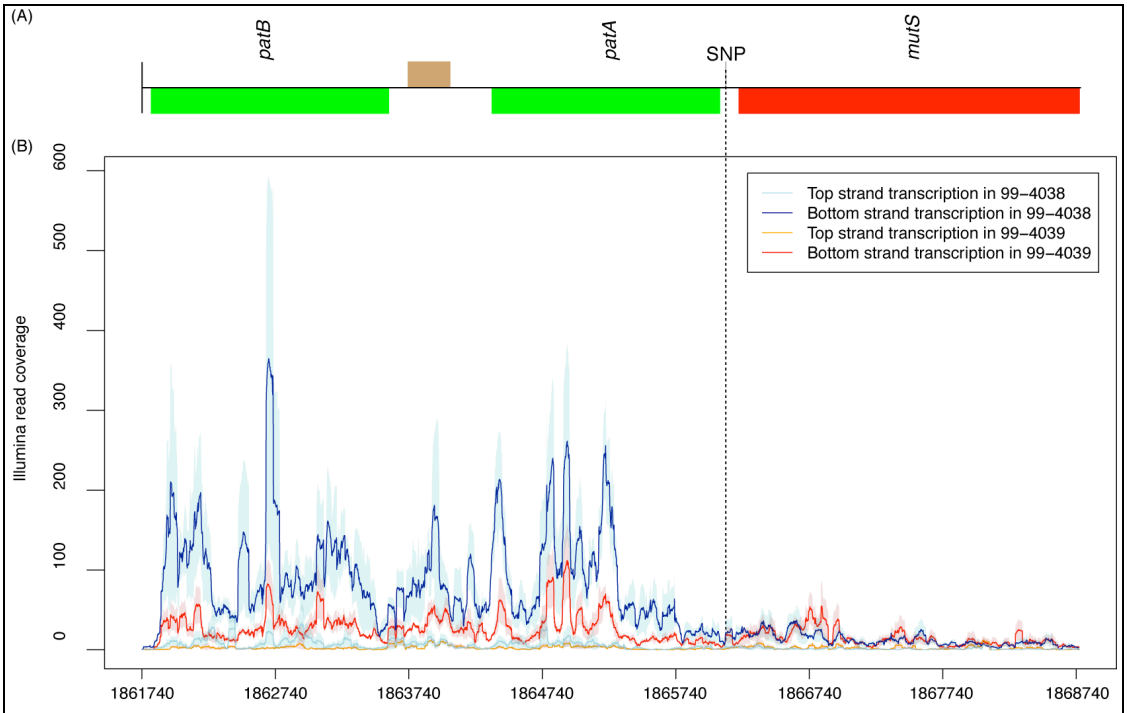


Figure 8.7 Expression of the *patAB* locus. (A) Annotation of the *patAB* gene cluster, with the intervening IS element remnant on the opposite strand, and the upstream *mutS* gene. The position of the polymorphism distinguishing 99-4038 and 99-4039 is indicated. (B) Graph showing Illumina read coverage from RNA-seq experiments. Each line shows the mean value from three replicates surrounded by a shaded area representing one standard error either side. The red and orange lines represent transcription of the two strands of the genome in 99-4038, and the blue lines represent the transcription of the two strands of the genome in 99-4039.

8.3.3 The mechanism of *patAB* upregulation

Analysis of the region upstream of the *patAB* operon revealed a strong promoter appearing to initiate transcription 69 nt upstream of the *patA* start codon. This 5' untranslated region is predicted to fold into a hairpin followed by a run of uridine residues, indicating that it could function as a terminator (Figure 8.8). This suggests a simple transcriptional attenuation mechanism; any compound that destabilizes this hairpin will increase the transcription of the downstream CDSs. The transcription of these genes is known to be increased by compounds that can intercalate nucleic acids, which have been found to disrupt RNA helices (Berman and Young, 1981). Hence this suggests a simple mechanism whereby any intercalating antimicrobial causes upregulation of an efflux pump capable of removing such planar compounds from the cell.

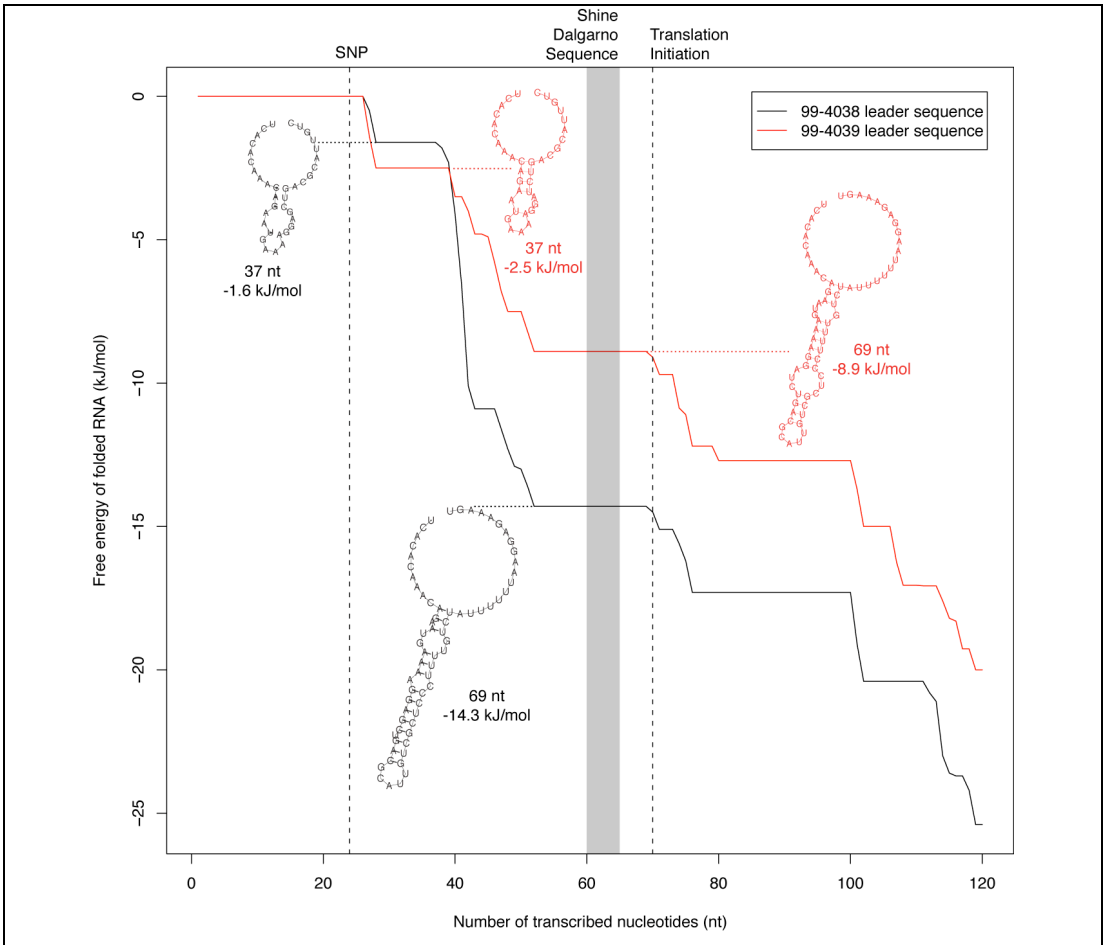


Figure 8.8 Leader sequence upstream of *patAB*. The graph represents the free energy of the most stable folded form of the 5' region of the *patAB* transcript, as it extends from the promoter, in 99-4038 (black line) and 99-4039 (red line). The positions of the polymorphism distinguishing the two strains, the Shine-Dalgarno sequence and translation initiation site are indicated by the dotted lines. The terminator-like hairpin structures formed by both transcripts as the RNA reaches the point at which translation initiation begins is displayed; it is evident that the mutation in 99-4039 is predicted to greatly weaken the stem of this secondary structure.

The *patAB* upregulatory SNP (PUS) distinguishing *S. pneumoniae* 99-4038 and 99-4039 is predicted to destablise this hairpin (Figure 8.8), thereby reducing any transcriptional attenuation and providing an explanation for the observed difference in expression levels. To test this hypothesis, the ~500 bp region around the PUS was cloned from each of *S. pneumoniae* 99-4038 and 99-4039, and transformed in *S. pneumoniae* TIGR4. These transformations were then grown in BHI broth with no selection, or containing 2 μg mL^{-1} ciprofloxacin. Only the culture transformed with the intergenic region from the CSF isolate generated mutants able to grow in the presence of ciprofloxacin (Figure 8.9). Sequencing the regions upstream of *patAB* in nine ciprofloxacin-resistant colonies, picked from three independent transformations with the leader sequence from the CSF isolate, revealed they had all acquired the PUS, whereas nine colonies picked from the same reactions with no selection all lacked it. The transcriptional profile of a transformant carrying the PUS (*S. pneumoniae* TIGR4^{PUS}) was compared to that of the otherwise isogenic parental strain under the same conditions as the comparison between *S. pneumoniae* 99-4038 and 99-4039. This confirmed the mutation caused the upregulation of the *patAB* operon; however, no other significant differences in expression were observed (Table 8.5).

8.4 Discussion

The broad spectrum of antimicrobial compounds against which PatAB appears to afford protection would seem to indicate the specificity of the efflux pump is relatively relaxed, removing a quite broad range of molecules from the cell. That 6-mercaptopurine appears to be the compound most effectively ejected from the cell suggests purine-like molecules (including fluoroquinolones) are the most efficient substrates of the pump; given the low specificity of its action, it seems likely purines themselves would also be pumped out of the cell. A hypothetical model for explaining the observed differences between *S. pneumoniae* 99-4038 and 99-4039 is outlined in Figure 8.10. Selection against this efflux of crucial metabolites may be the reason *patAB* expression is normally tightly regulated. The variability in expression of these genes observed in clinical isolates may reflect different abilities to adapt to such a detrimental side-effect of expression.

This hypothesis is supported by the apparently different responses of the clade I ST180 genotype and *S. pneumoniae* TIGR4 to the upregulation of *patAB*, although it is important to note that *S. pneumoniae* 99-4038 and 99-4039 are distinguished by a further two SNPs that may also impact on the differences between them. One possible reason for the increased sensitivity of the ST180 genotype to purine efflux may be deduced from Figure 8.5: the transcription of the lysogeny module of the ϕOXC141 prophage appears to extend across the adjacent *purA* purine synthesis gene in the antisense direction. This may inhibit *purA* expression and thereby limit the maximum possible rate of purine biosynthesis in the ST180 strains. However, in the light of the success of clade I, it would seem unlikely that such a mechanism would affect the fitness of the lineage under typical conditions.

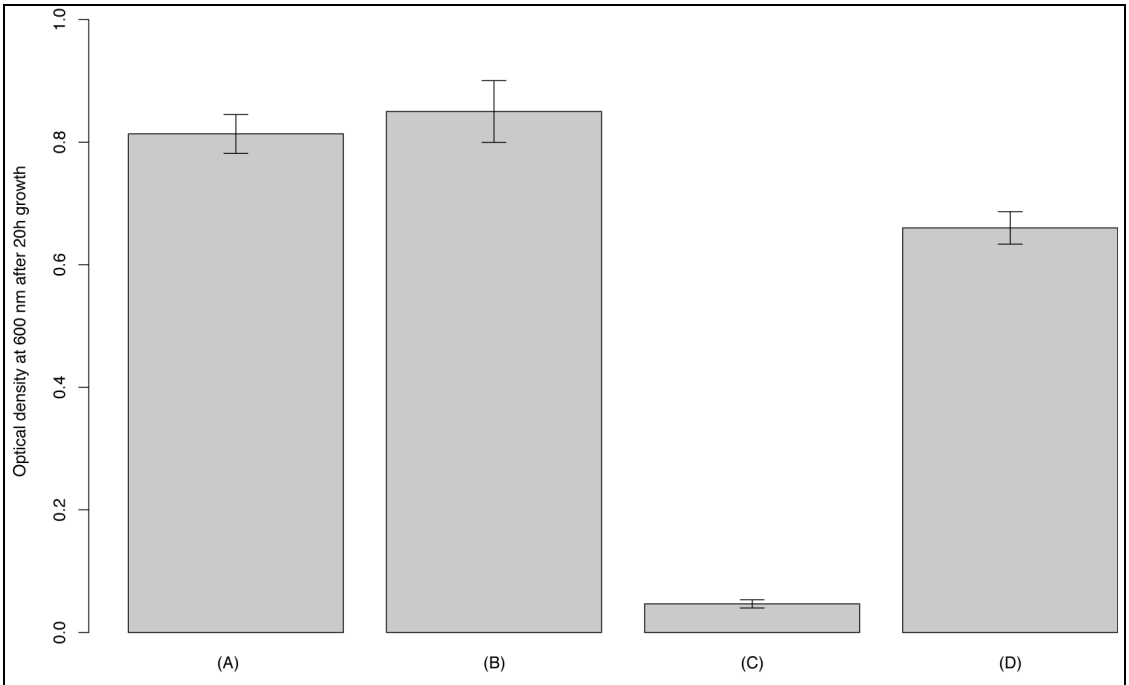


Figure 8.9 The effect of the PUS. A sample of *S. pneumoniae* TIGR4 was transformed with the region upstream of *patAB* from either (A) and (C) *S. pneumoniae* 99-4038 or (B) and (D) *S. pneumoniae* 99-4039. The transformed cultures were then grown in (A) and (B) BHI or (C) and (D) BHI supplemented with 2 μg mL^{-1} ciprofloxacin. The bars show the mean optical density at 600 nm after 20 h growth at 37 °C, with the error bars showing one standard error of the mean.

Such a putative cost of a resistance mechanism exemplifies the potential selective advantage of the continued antibiotic susceptibility of ST180. It seems clear that the mucoid capsule causes some degree of genetic isolation, causing a barrier to transformation. Furthermore, based on comparisons with PMEN1 and the observation

that capsule inhibits phage binding to teichoic acid receptors, the mucoid polysaccharide is likely to block virus and conjugative element acquisition as well. However, rare instances of insertions in this dataset, and the existence of macrolide-resistant ST180 strains (Isozumi *et al*., 2008), suggest the barrier is not impermeable. Furthermore, it seems unlikely that the capsule would be the reason for the paucity of observed fluoroquinolone resistance mutations, given that such substitutions were observed in smaller surveys of the non-transformable species *Salmonella* Typhi (Holt *et al*., 2008) and *Staph. aureus* ST239 (Harris *et al*., 2010). A potentially important caveat is that the sample of ST180 strains is heavily biased towards isolates from Europe, where low levels of fluoroquinolone resistance have been observed (Reinert *et al*., 2005a), and there has been a previous report of an ST180 isolate developing fluoroquinolone resistance through such mutations during treatment (Perez-Trallero *et al*., 2003). However, it is notable that these resistant variants have not displaced their universally susceptible progenitors. Hence it seems that a different response to selection pressures, rather than a lack of opportunity, shapes the contrasting resistance profiles of PMEN1 and ST180.

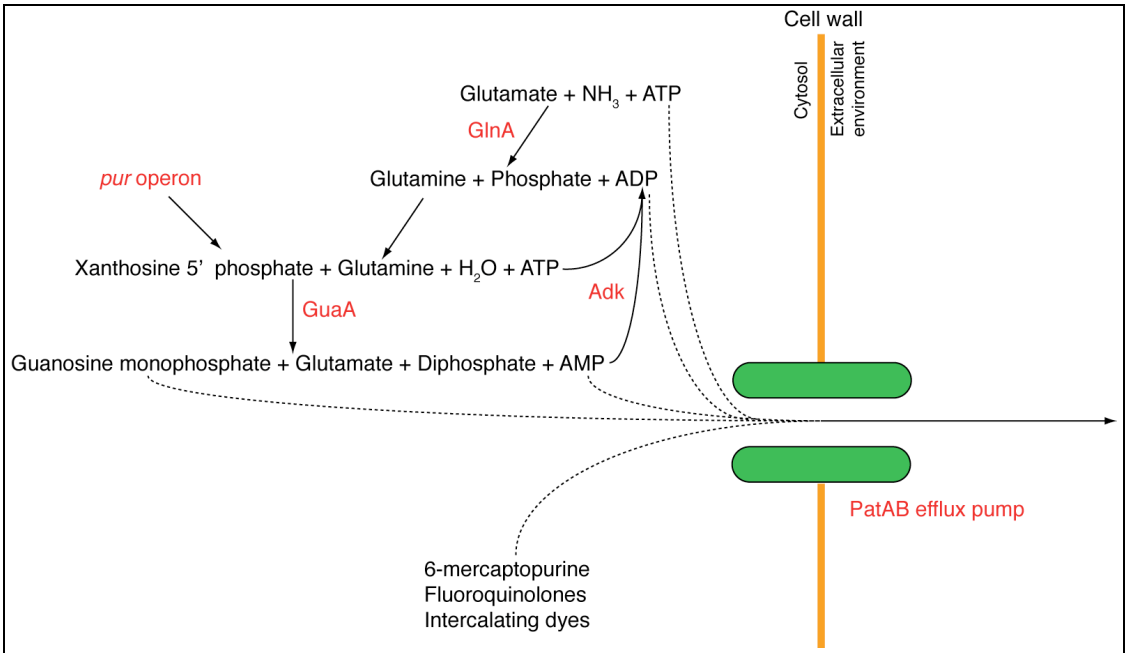


Figure 8.10 Hypothetical model to explain the changes in gene expression resulting from the PUS. With increased expression of *patAB*, there is predicted to be increased efflux of purines, based on the results of the phenotype microarray results. In turn, this would lead to upregulation of the pathways that replenish the purine levels within the cell, explaining many of the observed results from the RNA-seq comparison between *S. pneumoniae* 99-4038 and 99-4039.

Table 8.1 Microarray analysis of differential expression between *S. pneumoniae* 99-4038 and 99-4039. CDSs found to be expressed at significantly different levels in *S. pneumoniae* 99-4038 and *S. pneumoniae* 99-4039 using a microarray based on *S. pneumoniae* TIGR4. The fold change value represents the ratio of expression level in *S. pneumoniae* 99-4039 to that in *S. pneumoniae* 99-4038.

TIGR4 CDS	OXC141 Orthologue	Gene	Gene Product	Fold Change	Corrected *p* Value
SP_2075	SPNOXC18290	*patB*	ABC transporter ATP-binding membrane protein	5.09	0.00880
SP_2073	SPNOXC18270	*patA*	ABC transporter ATP-binding membrane protein	4.79	0.00880
SP_0493	SPNOXC04590	*rpoE*	putative DNA-directed RNA polymerase, delta subunit	0.38	0.0385
SP_0492	-	-	-	0.30	0.0354
SP_0373	SPNOXC03690		putative RNA methylase family protein	0.28	0.0496
SP_0490	SPNOXC04570		putative uncharacterized protein	0.28	0.0496
SP_0649	-	-	-	0.27	0.0318
SP_0489	SPNOXC04560		PAP2 superfamily protein	0.27	5.62×10^{-5}
SP_1293	SPNOXC11570	*rplS*	50S ribosomal protein L19	0.25	1.00×10^{-4}
SP_0487	-	-	-	0.21	1.21×10^{-4}
SP_0488	SPNOXC04550		putative membrane protein	0.16	2.11×10^{-5}
SP_2214	SPNOXC19560	*tsf*	elongation factor Ts	0.11	2.66×10^{-6}
SP_2215	SPNOXC19570	*rpsB*	30S ribosomal protein S2	0.08	4.62×10^{-6}

Table 8.2 CDSs found to be significantly more highly expressed in *S. pneumoniae* 99-4039 than in *S. pneumoniae* 99-4038 using RNA-seq.

OXC141 ID	TIGR4 Orthologue	Gene	Product	Fold Change	Corrected *p* value
SPNOXC03970	SP_0409	-	putative decarboxylase	2.65	2.29×10^{-15}
SPNOXC18260	SP_2072	-	putative peptidase	2.99	4.54×10^{-7}
SPNOXC00860	SP_0046	*purF*	putative amidophosphoribosyltransferase precursor	2.94	2.81×10^{-6}
SPNOXC00850	SP_0045	-	putative phosphoribosylformylglycinamidine synthase protein	2.68	2.81×10^{-6}
SPNOXC00880	SP_0048	*purN*	phosphoribosylglycinamide formyltransferase	3.00	8.75×10^{-6}
SPNOXC18270	SP_2073	*patA*	ABC transporter ATP-binding membrane protein	1.95	2.40×10^{-5}
SPNOXC00900	SP_0050	*purH*	bifunctional purine biosynthesis protein	2.75	2.04×10^{-4}
SPNOXC05670	SP_0620	-	putative extracellular solute-binding protein	2.70	3.01×10^{-4}
SPNOXC18290	SP_2075	*patB*	ABC transporter ATP-binding membrane protein	2.04	4.01×10^{-4}
SPNOXC02530	SP_0231	*adk*	adenylate kinase	1.69	0.00143
SPNOXC08640	SP_0962	*gloA*	putative lactoylglutathione lyase	1.52	0.00143
SPNOXC03060	SP_0288	-	putative CAAX amino terminal protease family membrane protein	2.90	0.00210
SPNOXC00840	SP_0044	*purC*	phosphoribosylaminoimidazole-succinocarboxamidesynthase	2.98	0.00497
SPNOXC02580	SP_0237	*rplQ*	50S ribosomal protein L17	1.56	0.00497
SPNOXC00890	SP_0049	*vanZ*	putative VanZ-family resistance protein	2.42	0.00794
SPNOXC03050	SP_0287	-	putative permease	1.67	0.0101
SPNOXC11890	SP_1355	*rplJ*	50S ribosomal protein L10	1.17	0.011
SPNOXC08630	SP_0961	*rplT*	50S ribosomal protein L20	1.45	0.011
SPNOXC06720	SP_0742	-	putative fatty-acid binding protein	1.38	0.011
SPNOXC19520	SP_2210	*cysM*	putative cysteine synthase	2.21	0.014
SPNOXC04660	SP_0502	*glnA*	putative glutamine synthetase	1.36	0.017
SPNOXC00930	SP_0054	*purK*	putative phosphoribosylaminoimidazole carboxylase ATPase subunit	1.51	0.018
SPNOXC02540	SP_0232	*infA*	translation initiation factor IF-1	1.27	0.021
SPNOXC12710	SP_1445	*guaA*	GMP synthase [glutamine-hydrolyzing]	1.26	0.022
SPNOXC13390	SP_1527	-	putative extracellular oligopeptide-binding protein	1.28	0.033
SPNOXC08610	SP_0959	*infC*	translation initiation factor IF-3	1.34	0.039
SPNOXC02570	SP_0236	*rpoA*	DNA-directed RNA polymerase alpha chain	1.53	0.044

Table 8.3 CDSs found to be significantly less highly expressed in *S. pneumoniae* 99-4039 than in *S. pneumoniae* 99-4038 using RNA-seq.

OXC141 ID	TIGR4 ID	Gene	Product	Fold Change	Corrected *p* Value
SPNOXC05950	SP_2314	-	putative uncharacterized protein	0.27	1.46×10^{-7}
SPNOXC03690	SP_0373	-	putative RNA methylase family	0.38	1.46×10^{-7}
SPNOXC03700	SP_0374	-	putative membrane protein	0.41	1.46×10^{-7}
SPNOXC19560	SP_2214	*tsf*	elongation factor Ts	0.37	2.81×10^{-6}
SPNOXC04610	SP_0496	-	putative Na^{+}/Pi-cotransporter protein	0.44	3.10×10^{-6}
SPNOXC19570	SP_2215	*rpsB*	30S ribosomal protein S2	0.27	5.37×10^{-6}
SPNOXC04550	SP_0488	-	putative membrane protein	0.44	5.86×10^{-6}
SPNOXC04560	SP_0489	-	PAP2 superfamily protein	0.42	1.06×10^{-5}
SPNOXC04770	SP_0517	*dnaK*	chaperone protein DnaK (heat shock protein 70)	0.44	3.21×10^{-5}
SPNOXC05940	SP_2313	-	putative uncharacterized protein	0.26	6.60×10^{-5}
SPNOXC16430	SP_1872	-	siderophore uptake periplasmic	0.40	2.25×10^{-4}
SPNOXC16410	SP_1870	-	putative iron compound ABC	0.40	0.00111
SPNOXC04570	SP_0490	-	putative uncharacterized protein	0.47	0.00111
SPNOXC14710	SP_1674	-	putative transcription regulator	0.55	0.00139
SPNOXC16420	SP_1871	-	siderophore uptake ATP-binding	0.40	0.00210
SPNOXC14290	SP_1626	*rpsO*	30S ribosomal protein S15	0.52	0.00210
SPNOXC04780	-	-	putative membrane protein	0.56	0.00210
SPNOXC07270	SP_0800	-	putative membrane protein	0.49	0.00407
SPNOXC18140	SP_2058	*tgt*	queuine tRNA-ribosyltransferase	0.65	0.00448
SPNOXC04580	SP_2283	-	acetyltransferase (GNAT) family	0.47	0.00855
SPNOXC04590	SP_0493	*rpoE*	putative DNA-directed RNA	0.63	0.0101
SPNOXC03470	SP_0338	*clpL*	putative ATP-dependent protease	0.66	0.0111
SPNOXC16320	SP_1859	-	nicotinamide mononucleotide	0.41	0.0165
SPNOXC05800	SP_0631	*rplA*	50S ribosomal protein L1	0.65	0.0216
SPNOXC04760	SP_0516	*grpE*	GrpE protein (HSP-70 cofactor)	0.79	0.0216
SPNOXC11570	SP_1293	*rplS*	50S ribosomal protein L19	0.64	0.0281

Table 8.4 Comparison of *S. pneumoniae* 99-4038 and 99-4039 using phenotype microarrays. Antimicrobial compounds found to result in significantly higher levels of respiration with *S. pneumoniae* 99-4039 than with *S. pneumoniae* 99-4038 on the Omnilog platform.

Antimicrobial	Mechanism	*p* value
Atropine	Acetylcholine receptor agonist	0.015
Aminotriazole	Amino acid analogue	0.011
Dodecyltrimethyl ammonium bromide	Cationic detegent	0.081
Domiphen bromide	Cationic detergent	0.046
Cetylpyridinium chloride	Cationic detergent	0.049
Cefsulodin	Cephalosporin	0.017
Orphenadrine	Cholinergic agonist	0.017
Nordihydroguaiaretic acid	Glutathione depletion	0.011
Acriflavine	Intercalator	0.011
5,7-Dichloro-8-hydroxyquinoline	Intercalator	0.011
Proflavine	Intercalator	0.011
3,4-Dimethoxybenzylalcohol	Oxidising agent	0.011
6-Mercaptopurine	Purine analogue	0.011
6-Mercaptopurine	Purine analogue	0.018
6-Mercaptopurine	Purine analogue	0.018
Amitriptyline	Serotonin reuptake inhibitor	0.018
Boric acid	Toxic anion	0.011
Sodium metaborate	Toxic anion	0.011
Sodium bromate	Toxic anion	0.011
D,L-Methionine hydroxymate	tRNA synthetase inhibitor	0.018
Pentachlorophenol	Uncoupler	0.011
Crystal violet	Uncoupler	0.011
2,4-Dinitrophenol	Uncoupler	0.024

Table 8.5 Microarray analysis of differential expression between *S. pneumoniae* TIGR4 and TIGR4^{PUS}. CDSs found to be significantly more highly expressed in *S. pneumoniae* TIGR4^{PUS} relative to the parental strain.

TIGR4 CDS	OXC141 Orthologue	Gene	Gene Product	Fold Change	Corrected *p* Value
SP_2075	SPNOXC18290	*patB*	ABC transporter ATP-binding membrane protein	9.64	0.00147
SP_2073	SPNOXC18270	*patA*	ABC transporter ATP-binding membrane protein	6.30	0.00147
SP_2074	SPNOXC18280	-		3.60	0.00481
SP_0455	-	-	Hypothetical protein	2.40	0.0291

9 Discussion

9.1 Pneumococcal transformation

9.1.1 Biases in the detection of recombination

The detection of recombination through genetic approaches requires the transfer of polymorphisms from the donor to the recipient; biochemical approaches, such as using isotopically labelled DNA, are required to overcome such limitations. This introduces an inherent degree of uncertainty into demarcating the extent of sequence transfers. At the simplest level, this can be quantified as the FRs in the work in Chapter 5: the position of RSS boundaries can only be measured to a level of precision corresponding to the density of SNPs in the region. This uncertainty is compounded by the mosaic nature of recombinations, alternating between recipient and donor segments: just as it cannot be assumed that the sequence between two neighbouring recipient allele polymorphisms is not affected by a small recombination that does not import any SNPs, within MRs it is not possible to tell whether there has been a reversion to recipient sequence in the distance between two polymorphisms converted to the respective donor alleles.

Nevertheless, the exponential distribution of RSS sizes detected in Chapter 5 suggests that a reasonable approximation of the underlying situation has been observed. Such a distribution of lengths suggests that very small events will be the most common, although there is likely to be a minimum enforced by the requirements for strand exchange on at least one side of the recombination. Unfortunately, the low sensitivity for detecting such short events places limits on the resolution at which this lower bound can be defined. The power to detect short recombinations is further reduced when the sequence of the donor bacterium is not known, as when examining the collection of PMEN1 and ST180 isolates (Chapters 4 and 8). Hence although the distribution of detected homologous recombination sizes is similar to those found in the *in vitro* work, there is a notable under representation of short events less likely to cause a statistically significant increase in SNP density. This leads to the differences in the estimated mean and median sizes of such events in the two studies. More

accurate estimates will require the improved precision achievable through the study of exchanges between more divergent bacteria, which would require good reference sequences to be generated for strains with appropriate selectable markers from other mitis group species.

As indicated by the distribution of RSS lengths spanning the selected locus in Chapter 5, studying a single genetic locus exacerbates the problem of overrepresentation of larger recombination events through 'length biased sampling' (Cox, 1969). Just as bigger events are more likely to contain polymorphisms that make them detectable, they are also more probable to affect any given subset of a chromosome, an issue only overcome by genome-wide analyses.

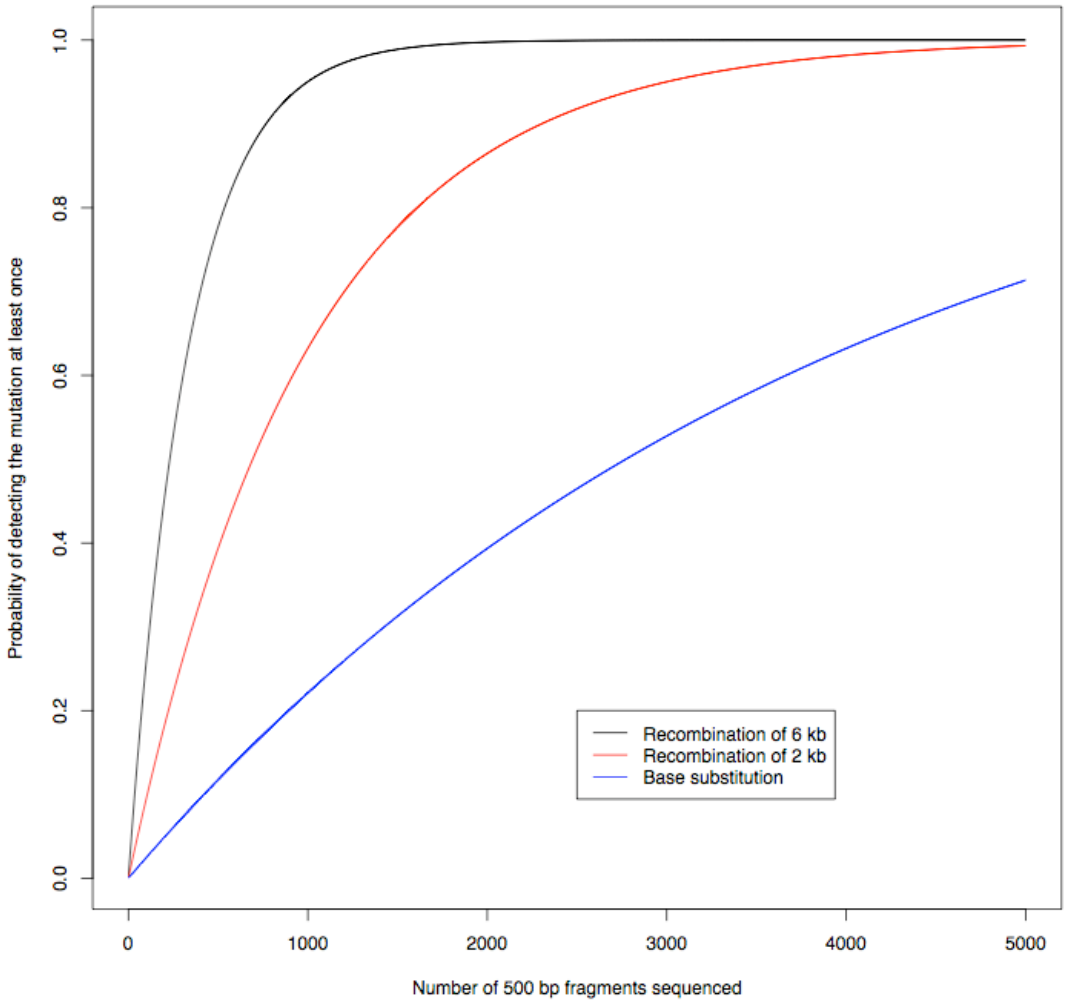


Figure 9.1 Modelling the ability to detect recombinations and mutations through sequencing multiple loci. The probability of detecting a point mutation or recombination at least once through the sequencing of n multiple loci is displayed for two different values of the recombination length, x (Equation 9.1). This shows how, the larger a mutation is, the more likely it is to be detected by incomplete sequencing of a genome. All three lines approach an asymptote of $p = 1$ as a greater proportion of the genome is sequenced.

9.1.2 Impact of biases on *r/m* estimates

This simple principle has an important consequence in resolving the discrepancy between the estimates of *r/m* for *S. pneumoniae* from the PMEN1 and ST180 analyses and the MLST database. By studying seven defined regions of the genome, MLST is more likely to detect larger mutations than smaller ones, because the greater the extent of a mutation, the more likely it is to overlap with one of the studied loci. Hence, given the same sample, the calculated *r/m* value will change as different proportions of the chromosome are sequenced. The effect can be quantified through a simple model. If a single point mutation and recombination of length *x* bp are present in a 2 Mb genome, then when sequencing a 500 bp locus there is a probability of (500 ÷ 2,000,000 =) $2.5\text{x}10^{-4}$ of detecting the point mutation, but a probability of approximately $(2.5\text{x}10^{-4})x$ of identifying the recombination, depending on the density of SNPs throughout the event. Using values of *x* corresponding to 2 kb (close to the mean length from the transformation experiment) and 6 kb (close to the mean length estimated from the PMEN1 dataset), the probabilities of detecting a single point mutation (p_{mut}) and single recombination event (p_{rec}) at least once using data from *n* multiple loci can be estimated from a binomial distribution.

$$p_{mut} \sim Binom\left(n, \frac{1}{2000000}\right)$$
$$p_{rec} \sim Binom\left(n, \frac{x}{2000000}\right)$$

Equation 9.1

These outcomes of these models are displayed graphically in Figure 9.1. When only a few loci are sequenced, the bias towards the detection of larger recombinations is greatest. When $n = 7$ and $x = 6$ kb, the ratio *r/m* would be estimated as 11.9 rather than one. This approximates to the observed 9.2-fold and 25.4-fold difference between the *r/m* ratios observed for PMEN1 and ST180, respectively, and the MLST data (Feil *et al*., 2000). Hence the *r/m* statistics derived from whole and partial genome sequences are not easily comparable.

9.1.3 The advantage of being transformable

A number of different hypotheses have been proposed as the main function of the competence system in *S. pneumoniae* and other bacterial species. These fall into two categories: the advantages of generating genetic diversity, and the use of exogenous DNA as a source of metabolic substrates.

Hypotheses of the former category were originally formulated for sexually reproducing eukaryotic species. These suggest the main advantage of exchange is the acquisition of beneficial genetic material. The first model of this type, the Fisher-Muller hypothesis (Fisher, 1930; Muller, 1932), proposed that recombining populations evolve more quickly than asexual populations because advantageous alleles at different loci can accumulate into a single genotype more rapidly. However, this explanation was criticised for being formulated in terms of group selection (Maynard Smith, 1978), the validity of which is very doubtful (Maynard Smith, 1964; Williams, 1966). Subsequent descriptions have focussed on demonstrating advantages on the level of selection of individuals (Hill and Robertson, 1966; Felsenstein, 1974; Felsenstein and Yokoyama, 1976).

When applied to bacteria, this category of hypotheses is further subdivided. One set of extensions to this concept has focussed on the increased speed with which populations can respond to diversifying or balancing selection pressures (Michod *et al*., 2008; Vos, 2009). Given the nature of the regulation of the pneumococcal competence system, this has lead to the proposal that *S. pneumoniae* is adapted to diversify under stress, when it is maladapted to its environment (Claverys *et al*., 2006; Prudhomme *et al*., 2006). A second set of hypotheses suggest that transformation allows an improved response to purifying selection pressures: these focus on the possibility of recombination repairing genes afflicted by disruptive mutations (Michod *et al*., 2008; Vos, 2009), and thereby reversing Muller's ratchet (Muller, 1964; Felsenstein, 1974).

The main problem with such explanations is that the competence system, *a priori*, cannot distinguish beneficial and deleterious mutations, hence particular circumstances are required for transformation events to be of an overall selective advantage to an individual cell. For instance, the nature of epistatic interactions

between different loci around the chromosome has a bearing on the fitness of competence (de Visser and Elena, 2007). Positive epistasis, occurring when different advantageous alleles in a genotype adapt to one another, selects against transformation because mutually beneficial interactions are disrupted. Conversely, negative epistasis, involving deleterious loci reinforcing one another's negative impact, favours the evolution of transformation. Hence competence is beneficial when there is strong selection for the repair of deleterious mutations, while there are relatively small fitness costs when adaptive polymorphisms are lost. An alternative explanation is that the system can be biased towards the uptake of adaptive mutations through triggering competence in response to stress (Redfield, 1988). An increase in fitness following the acquisition of DNA is most likely when the host genotype has been damaged, or is maladapted to its environment, through the simple principle of regression to the mean. One criticism levelled at this model is that free DNA in the environment is most likely to have been released by cells that have already lysed, and hence this exogenous gene pool is likely to overrepresent genotypes poorly adapted to the environment (Redfield, 1988). A more direct response to genetic damage is the proposed triggering of transformation events at the site of lesions created by genotoxic stress (Hoelzer and Michod, 1991). However, this is not congruent with the observation that an increase in external DNA concentration can cause a rise in the amount of sequence imported without a commensurate increase in the level of host DNA damage (*e.g.* Chapter 5), suggesting only a small proportion of transformation events, if any, are directly instigated by genomic lesions.

Such couterpoints, and the observation that *H. influenzae* and *B. subtilis* do not regulate competence in response to DNA damage (Redfield, 1993a), lead to an alternative suggestion for the purpose of competence. Rather than a source of genetic information, imported DNA may serve as a source of nucleotides for DNA and RNA synthesis. Exogenous DNA is abundant in the nasopharynx, estimated to be present at a concentration of ~300 mg L^{-1} mucus (Matthews *et al*., 1963) and, in the case of the purine and pyrimidine auxotroph *H. influenzae*, transcription of the competence genes is repressed by an abundance of nucleotides in the growth medium (MacFadyen *et al*., 2001). However, several aspects of the competence machinery imply it is unlikely to have been optimised for the acquisition of nucleotides. Firstly, in the case of the

pneumococcal system, only one strand is imported through the pore while the other is degraded, making its constituents available to other cells (Dubnau, 1999; Johnsborg *et al*., 2007). Secondly, once inside the cell, the ssDNA is protected from degradation through the loading of ssDNA binding proteins, the binding and disassociation cycle of which requires ATP hydrolysis. Thirdly, species such as *H. influenzae* and *N. meningitidis* do not take up all DNA available to them, but rather selectively uptake homologous sequences containing distinctive repeats, hence limiting their ability to acquire nucleotides (Dubnau, 1999; Smith *et al*., 1999).

Whatever the primary function of the competence system, it must be sufficiently advantageous to counteract the inherent instability of such a mechanism. If transformation events only rarely prove to be beneficial, then the competence system will become defunct comparatively rapidly relative to purely metabolic genes. This results from an the asymmetric situation in which transformable bacteria are able to horizontally acquire mutations, or genes that inhibit transformation, that render the competence system non-functional, while non-competent cells cannot acquire sequences that cause them to revert to a competent state (Redfield, 1993b).

9.1.4 Inferences from PMEN1 and ST180

The 'hotspots' of recombination in the PMEN1 population, such as the capsule and proteinaceous antigens *pspA* and *pspC*, would appear to support the hypothesis that the competence system benefits the population through allowing it to respond to diversifying selection. However, this perception may be due to the bias of studying a single lineage: both *pspA* and *pspC* are very diverse genes within the species, hence are likely to be identified as an import each time they are transferred, whereas there is less sensitivity to detect transfer of other, more conserved, regions of the genome. Furthermore, it is likely that the acquired alleles of these antigens only appear diverse relative to the PMEN1 background and are actually sequences already frequent in the species. This is because transformation will sample alleles in proportion to their frequency in the population, thereby making such recombination events an inefficient mechanism for responding to balancing selection, which drives an increase in the frequency of rare alleles.

Furthermore, the mean rate of import from other lineages was approximately one event per three years in PMEN1 and one event per 30 years in ST180, although these rates were highly heterogeneous across the phylogenies. Analysis of MLST data from *H. influenzae* suggest that such horizontal transfer of divergent sequence is even rarer in that species (Feil *et al*., 2001). Hence it seems likely that transformation within a clonal population of bacteria, rather than between lineages, is likely to be the main selective pressure maintaining the competence system; this would suggest repair is a more important function than diversification. However, the relatively low recombination rate in ST180 does not appear to result in a particularly elevated rate of pseudogene formation: 6.7% of the CDSs in *S. pneumoniae* ATCC 700669 are pseudogenes, compared to 7.6% in *S. pneumoniae* OXC141, while 0.077 gene disruptions per SNP were observed in the PMEN1 sample, relative to 0.025 per SNP in ST180. However, this dissimilarity is likely to result from the differences in sequence data quality and indel identification, as well as the longer time period over which selection has been able to act to remove deleterious mutations from the ST180 lineage. Furthermore, the most frequent spontaneous mutations observed in *S. pneumoniae*, transitions and small indels, are efficiently repaired by the MMR system when imported in low numbers (Claverys *et al*., 1981; Lacks *et al*., 1982; Claverys *et al*., 1983), making this an ineffective use of the transformation machinery.

9.2 Site-specific and homologous recombination

9.2.1 The characteristics of horizontal sequence transfer

Transduction and conjugation, like transformation, both require the donor and recipient bacteria to co-colonise the same environment. However, the donor cell is usually lysed in order to release genomic DNA or phage particles for transformation and transduction respectively, whereas the donor cell remains intact following conjugation. Another trait of conjugation is that the donor may be distantly related to the recipient, given the wide host range of ICE (Chapter 3). By contrast, heterospecific transfers via phage or transformation are relatively rare (Chapter 5).

While the main evolutionary advantage of transformation is still debated, conjugation and phage infection are both driven by mobile genetic elements (MGEs), which, unless they carry selectively advantageous cargo, are likely to be detrimental to the host. This necessity of transmitting an entire MGE means that conjugation and phage infection events are large, spanning tens of kilobases, which contrasts with the observed range of transformation events, which have a median length of around 2.3 kb (Chapter 4). These differences suggest a novel hypothesis for the role of transformation in bacterial evolution.

9.2.2 Transformation as bacterial 'gene therapy'

As stated in Chapter 5, given the exponential distribution of RSS sizes and the apparent exclusion of donor insertions from such regions, it is evident that at any given polymorphic locus in a transformable pneumococcal population the smallest allele will transfer most quickly among the population and hence, in the absence of selection, will drift to fixation. Therefore transformation tends to reduce the size of the genome, as opposed to site-specific recombination, which inserts material directly into the chromosome. The insertions that are removed the most quickly will be those that are large and lack any selective advantage, the most obvious example of which are prophage and cryptic ICE. Hence it may be that homologous recombination serves primarily as a mechanism for removing parasites from the genome. Such DNA exchanges would be effective at eliminating any recently acquired parasite that was polymorphic within the population, allowing transformation to act as a post-integrative MGE defence mechanism.

Unlike the import of SNPs and small indels, where deleterious alleles are as likely to be acquired as beneficial ones, the system would be adapted to specifically importing short sequences in order to remove large insertions, and the transfers would be more efficient because long indels are much less frequently repaired through MMR (Claverys *et al*., 1980). Given this hypothesis, it may be expected that selection for transformation events at the prophage integration sites in the pneumococcal genome would lead to them being identified as 'hotspots' in the PMEN1 population. However, more detailed consideration suggests this is unlikely. Transformation will only be effective at removing prophage shortly after infection, when the element is

polymorphic in a clonally-related population; during this period, the co-colonising strain would almost certainly have been the origin of the prophage, hence could not donate non-lysogenic sequence that would lead to its removal. Only sequence donated by non-lysogenic members of the same clonally-related population could be used for repair, and these would not be identified as imports in the analysis outlined in Chapter 4. Once the prophage has been fixed through transmission bottlenecks, its subsequent elimination would be unlikely, as reinfection would be highly probable to undo any repair events.

Is MGE transmission within clonally-related bacteria a sufficiently strong selection pressure to maintain the competence system? The availability of prophage sequences within pneumococcal genomes permitted surveys of clinical isolates to be conducted; these have found that the majority of strains are lysogenic (Ramirez *et al*., 1999; Romero *et al*., 2009). When present at such a high level within the population, phage infections must be more frequent than imports of sequence from other lineage. While the physical distance to be traversed between pneumococcal communities is the same in each case, the phage is adapted for efficient and targeted transmission between pneumococci, protected against DNAse activity and is not competing for uptake with high concentrations of heterologous host DNA. One difference is that several restriction systems are present in the pneumococcal genome to protect against the acquisition of MGE as dsDNA, but will not act on ssDNA entering as a RecA- or DprA-coated nucleoprotein filament. One such system, displayed in Chapter 6, is the *hsdS* hypervariable restriction enzyme locus, which alters in its specificity over very short timescales; this seems likely to be importing in inhibiting the spread of recently acquired MGEs between clonally-related pneumococci, as this model proposes the competence system may do. The rapid alternation between forms means an MGE that was previously integrated into, and methylated as part of, a clonally related bacterium can be recognised and degraded by the cell, which would only be advantageous if the cell itself was not yet hosting the element itself.

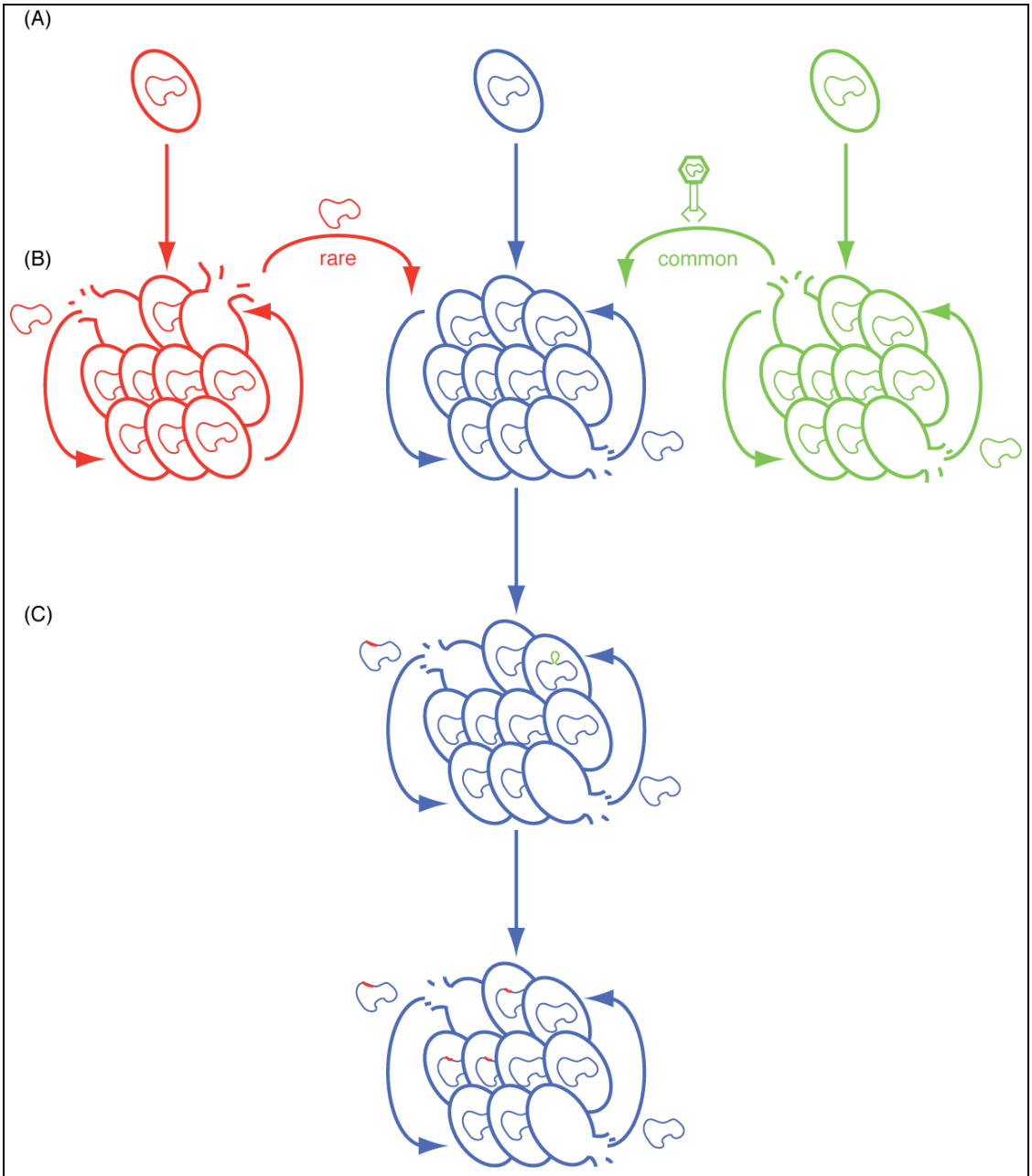


Figure 9.2 Model for the role of transformation in pneumococci. (A) Three different co-colonising *S. pneumoniae* genotypes are represented in red, green and blue, each of which gives rise to a spatially separated clonal population. (B) DNA exchange is rapid within each clonal population. MGEs, adapted to transmission between bacteria, are frequently passed between lineages. The movement of genomic DNA between different strains is likely to be much less common. (C) Once diversity has arisen in the originally clonal population through DNA import and MGE acquisition, its prevalence is determined by exchange within the population. Alleles of equal size are likely to be exchanged at equal rates, whereas the transfer of smaller alleles will be more rapid than that of larger alleles at the same locus, leading to elimination of recently acquired MGEs unless selection acts to retain them.

9.2.3 The interaction of mobile elements and transformation

A line of evidence that would support the hypothesis that recombination acts to remove mobile elements from the genome is the mechanisms that parasitic entities in

pneumococcal genomes appear to employ to counter this host defence tactic. The simplest mechanism, observed to occur with the group 2b prophage in the PMEN1 population, was complete abrogation of the host's competence system. This resulted from the insertion of the lysogen into the *comYC* gene, encoding a subunit of the competence pseudopilus essential for transformation. Similarly, the non-transformable equine pathogen *S. equi* is distinguished from its progenitor, the transformable streptococcus *S. zooepidemicus*, by the insertion of a prophage into a homologue of *comFA*, essential for competence (Holden *et al*., 2009b). Such targeting of the viral integration event immediately removes the chance of the element being removed by a homologous recombination.

Another putative defence against recombinatorial removal is seen with the *Tn*5252-type ICE common in multidrug-resistant *S. pneumoniae* strains. Notable among their cargo is the presence of *uvrD* or *umuCD* DNA repair genes, the *E. coli* orthologues of which are both upregulated as part of the SOS response to genotoxic stress (Munoz-Najar and Vijayakumar, 1999). The diverse sources of these sequences suggest they are cargo genes that have been acquired in parallel, rather than constituting part of the much more strongly conserved conjugative machinery. In *E. coli*, RecA filaments recruit UmuCD to ssDNA lesions within the genome to trigger repair via translesion synthesis; as part of this function, the UmuCD complex appears to be able to bind and disrupt RecA nucleoprotein filaments (Sommer *et al*., 1993; Rehrauer *et al*., 1998). Similarly, mutant *E. coli* strains found to perform RecA-mediated integration of conjugatively-transferred DNA at an increased rate frequently have a disrupted *uvrD* gene (Arthur and Lloyd, 1980; Feinstein and Low, 1986; Bierne *et al*., 1997), with overexpression of *uvrD* inhibiting RecA-mediated recombination (Petranovic *et al*., 2001). This appears to be a consequence of UvrD's ability to dismantle RecA nucleoprotein filaments (Veaute *et al*., 2005) and inhibit RecA-facilitated pairing of homologous sequence (Morel *et al*., 1993). Hence these ICE-borne genes would be able to reduce the rate of recombination by disrupting RecA nucleoprotein filaments before they invade the genomic DNA duplex. Such an advantage would also apply *in trans* to any other element sharing the same genome; this may explain the observed association of the shorter, independently mobile Tn*916* element with the larger

transposon, and the consequent association of chloramphenicol and tetracycline resistance.

Lastly, the small interspersed repeat sequences so prevalent in *S. pneumoniae* (Chapter 7) seem likely to be exactly the type of parasites that could propagate effectively despite transformation. While transformation seems likely to be effective at removing large, single copy parasites, smaller elements would be more easily horizontally acquired and more slowly removed. Also, their high copy number would mean that they could not be eliminated effectively by a small number of horizontal sequence transfer events.

The high rate of MGE turnover in the population is likely to have the consequence that examples of such elements retained in the host are likely to be both selectively advantageous and actively mobile. While the prophage population in PMEN1 undergoes rapid flux, ICE*Sp*23FST81 remains relatively stable, presumably due to the selection for the antibiotic resistance genes it carries. However, the presence of ICE*Sp*11876 and ICE*Sp*11930 in the population appear to represent cases where ICE*Sp*23FST81 has been lost and alternative ICEs, carrying similar resistance elements, rapidly acquired in its place. In conjunction with the 'scars' observed adjacent to *rplL*, this supports the hypothesis that there is a rapid flux of conjugative elements through the pneumococcal population, which are swiftly eliminated by transformation unless there is a selection pressure to retain them.

9.2.4 Criticism of this model

Transformation would appear to be a poor method of responding to diversifying selection, because it would sample the most common alleles at any given locus most frequently. It also appears to be an inefficient way of repairing the most common forms of genome damage, because disruptive mutations would be as frequently acquired as advantageous ones, and also the MMR system is adept at inhibiting such transfers anyway. However, in order for this hypothesis to account for the advantages of transformation, it is necessary that large insertions in the recipient, relative to the donor strand, do not inhibit recombination events. Evidence for this was presented in Chapter 5, but more precise experiments focussing on this question would provide a

more definitive answer. Additionally, it is necessary that exchange within a clonally descended population is sufficiently fast to overcome the spread of MGEs through infection. While competence is upregulated under stressful conditions, when lysogens are most likely to be induced, phage have an intrinsic advantage of targeting a specific locus in the genome, whereas transformation necessarily has to cover the entire genome at random with short DNA fragments, hence must be a very common occurrence in order to compete with prophage integration at any one site. *In vitro* systems that more closely mimic the *in vivo* niche of the pneumococcus will be required to accurately test these relative rates. However, even if the kinetics of transformation are only sufficient to retard the spread of prophage, rather than eliminate them, then this would still lead to a greater chance of a non-lysogenic member of the population being transmitted to another host.

9.2.5 Comparing the evolutionary dynamics of PMEN1 and ST180

The comparison of PMEN1 and ST180 reveals starkly contrasting patterns of variation, even though both appear to have dramatically expanded in their population size over the past few decades. Assuming these differences do not represent a much stronger purifying selection pressure on ST180, all three mechanisms of horizontal DNA transfer seem to be more rapid in PMEN1, hence its more frequent acquisition of antibiotic resistance and different capsule types.

The success of ST180, despite its infrequent imports of divergent sequence, would seem to contradict the hypothesis that transformation increases a bacterium's fitness through enabling it to respond to diversifying or balancing selection. Furthermore, there is no evidence for a significantly increased proportion of pseudogenes arising in the population, implying transformation is unlikely to have a significant role in repair. It could be considered that ST180 may have switched from using imported DNA for genetic purposes to using it as a metabolic substrate instead, which could make sense in the light of the lineage's increased requirement for nucleotide sugars for the biosynthesis of the extensive mucoid capsule. However, given such a pressing need of the cell's biochemistry, it would not be expected to find the observed mutants with disruptions to key competence genes.

Rather, these observations are more in keeping with the lack of selection pressure for the retention of transformability in ST180, in turn congruent with the hypothesis that transformation acts to remove mobile elements. The thick capsular layer is likely to inhibit phage adhesion (Bernheimer and Tiraby, 1976), which would correspond with the very low rate of prophage integration observed. The capsule is likely to prevent conjugation as well, given the rarity of ICE in the sequenced ST180 isolates. With this physical barrier protecting against such parasites, the selection pressure to maintain competence would be weakened. The low rate of transformation in ST180 may account for the resilience of ΦOXC141, despite DNA and RNA sequencing revealing it to be active, and rare homoplasic deletions suggesting its loss may be advantageous. This comparison between PMEN1 and ST180 suggests that the thickness of the capsular layer may link the rates of phage infection, conjugative transfer and transformation, but these lineages are likely to represent extremes of the species. Whether this relationship holds when considering a greater number of serotypes will require the sequencing of further lineages.

9.3 Concluding remarks

Horizontal sequence transfer through transformation, conjugation and viral infection all contribute to the genetic diversity of *S. pneumoniae*. The flux of mobile elements appears to be rapid in lineages such as PMEN1, where site-specific integration leads to rapid acquisition of MGEs, while homologous recombination frequently acts to remove such features from the genome. In ST180, the mucoid capsule inhibits mobile element acquisitions but also slows the rate of transformation correspondingly. This leads to a much more static genotype, hence reducing the incidence of antibiotic resistance gene acquisition and serotype switching. However, the lower level of variation alone is not sufficient to explain the low prevalence of resistance mutations in this lineage, suggesting that selection maintains the universally susceptible phenotype. Improved understanding of the evolutionary dynamics of lineages in between these two extremes of the population should inform our knowledge of how *S. pneumoniae* populations are likely to react to clinical interventions in the future.

References

Aanensen, D. M. and B. G. Spratt (2005). The multilocus sequence typing network: mlst.net. *Nucleic Acids Res* **33**(Web Server issue):W728-33.

Achaz, G., *et al*. (2002). Origin and fate of repeats in bacteria. *Nucleic Acids Res* **30**(13):2987-94.

Adams, W. G., *et al*. (1993). Decline of childhood *Haemophilus influenzae* type b (Hib) disease in the Hib vaccine era. *JAMA* **269**(2):221-6.

Aguiar, S. I., *et al*. (2010). Serotypes 1, 7F and 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7 years of heptavalent conjugate vaccine use. *Vaccine* **28**(32):5167-73.

Ahronheim, G. A., *et al*. (1979). Penicillin-insensitive pneumococci. Case report and review. *Am J Dis Child* **133**(2):187-91.

Ajdic, D., *et al*. (2002). Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci U S A* **99**(22):14434-9.

Alloing, G., *et al*. (1998). Development of competence in *Streptococcus pneumoniae*: pheromone autoinduction and control of quorum sensing by the oligopeptide permease. *Mol Microbiol* **29**(1):75-83.

Almirall, J., *et al*. (2000). Epidemiology of community-acquired pneumonia in adults: a population-based study. *Eur Respir J* **15**(4):757-63.

Altschul, S. F., *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17):3389-402.

Ambur, O. H., *et al*. (2007). New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. *J Bacteriol* **189**(5):2077-85.

Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**(10):R106.

Angiuoli, S. V. and S. L. Salzberg (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**(3):334-42.

Aniansson, G., *et al*. (1992). Nasopharyngeal colonization during the first year of life. *J Infect Dis* **165 Suppl 1**:S38-42.

Ardanuy, C., *et al*. (2009). Emergence of a multidrug-resistant clone (ST320) among invasive serotype 19A pneumococci in Spain. *J Antimicrob Chemother* **64**(3):507-10.

Arrecubieta, C., *et al*. (1995). Sequence and transcriptional analysis of a DNA region involved in the production of capsular polysaccharide in *Streptococcus pneumoniae* type 3. *Gene* **167**(1-2):1-7.

Arrecubieta, C., *et al*. (1994). Molecular characterization of *cap3A*, a gene from the operon required for the synthesis of the capsule of *Streptococcus pneumoniae* type 3: sequencing of mutations responsible for the unencapsulated phenotype and localization of the capsular cluster on the pneumococcal chromosome. *J Bacteriol* **176**(20):6375-83.

Arthur, H. M. and R. G. Lloyd (1980). Hyper-recombination in *uvrD* mutants of *Escherichia coli* K-12. *Mol Gen Genet* **180**(1):185-91.

Ash, R. and M. Solis-Cohen (1929). Contrasted behaviour of pneumococci toward quinin and optochin in relation to drug fastness. *J Infect Dis* **45**(6):457-462.

Aspa, J., *et al*. (2006). Impact of initial antibiotic choice on mortality from pneumococcal pneumonia. *Eur Respir J* **27**(5):1010-9.

Assefa, S., *et al*. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**(15):1968-9.

Auburtin, M., *et al*. (2002). Pneumococcal meningitis in the intensive care unit: prognostic factors of clinical outcome in a series of 80 cases. *Am J Respir Crit Care Med* **165**(5):713-7.

Austrian, R. (1960). The Gram stain and the etiology of lobar pneumonia, an historical note. *Microbiol Mol Biol Rev* **24**(3):261-265.

Austrian, R. (1978). The Jeremiah Metzger Lecture: Of gold and pneumococci: a history of pneumococcal vaccines in South Africa. *Trans Am Clin Climatol Assoc* **89**:141-61.

Austrian, R. (2000). Pneumococcal otitis media and pneumococcal vaccines, a historical perspective. *Vaccine* **19 Suppl 1**:S71-7.

Austrian, R., *et al*. (1976). Prevention of pneumococcal pneumonia by vaccination. *Trans Assoc Am Physicians* **89**:184-94.

Austrian, R. and R. Rosenblum (1953). The relative efficacy of erythromycin (ilotycin) and of penicillin in the treatment of pneumococcal lobar pneumonia. *Am J Med Sci* **226**(5):487-90.

Avadhanula, V., *et al*. (2006). Respiratory viruses augment the adhesion of bacterial pathogens to respiratory epithelium in a viral species- and cell type-dependent manner. *J Virol* **80**(4):1629-36.

Avery, O., *et al*. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med* **79**(2):137-158.

Avery, O. T. and R. Dubos (1931). The protective action of a specific enzyme against type III pneumococcus infection in mice. *J Exp Med* **54**(1):73-89.

Avery, O. T. and W. F. Goebel (1931). Chemo-immunological studies on conjugated carbohydrate-proteins : V. The immunological specifity of an antigen prepared by combining the capsular polysaccharide of type III pneumococcus with foreign protein. *J Exp Med* **54**(3):437-47.

Ayoubi, P., *et al*. (1991). Tn*5253*, the pneumococcal omega (*cat tet*) BM6001 element, is a composite structure of two conjugative transposons, Tn*5251* and Tn*5252*. *J Bacteriol* **173**(5):1617-22.

Bachellier, S., *et al*. (1999). Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res Microbiol* **150**(9-10):627-39.

Bagnoli, F., *et al*. (2008). A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. *J Bacteriol* **190**(15):5480-92.

Bakir, M., *et al*. (2001). Asymtomatic carriage of *Neisseria meningitidis* and *Neisseria lactamica* in relation to *Streptococcus pneumoniae* and *Haemophilus influenzae* colonization in healthy children: apropos of 1400 children sampled. *Eur J Epidemiol* **17**(11):1015-8.

Balganesh, T. S. and S. A. Lacks (1985). Heteroduplex DNA mismatch repair system of *Streptococcus pneumoniae*: cloning and expression of the *hexA* gene. *J Bacteriol* **162**(3):979-84.

Ball, P. (1994). Bacterial resistance to fluoroquinolones: lessons to be learned. *Infection* **22 Suppl 2**:S140-7.

Barbour, M. L., *et al*. (1995). The impact of conjugate vaccine on carriage of *Haemophilus influenzae* type b. *J Infect Dis* **171**(1):93-8.

Barcus, V. A., *et al*. (1995). Genetics of high level penicillin resistance in clinical isolates of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **126**(3):299-303.

Barnes, D. M., *et al*. (1995). Transmission of multidrug-resistant serotype 23F *Streptococcus pneumoniae* in group day care: evidence suggesting capsular transformation of the resistant strain *in vivo*. *J Infect Dis* **171**(4):890-6.

Barocchi, M. A., *et al*. (2006). A pneumococcal pilus influences virulence and host inflammatory responses. *Proc Natl Acad Sci U S A* **103**(8):2857-62.

Barrett-Connor, E. (1971). Bacterial infection and sickle cell anemia. An analysis of 250 infections in 166 patients and a review of the literature. *Medicine (Baltimore)* **50**(2):97-112.

Beall, B., *et al*. (2006). Pre- and postvaccination clonal compositions of invasive pneumococcal serotypes for isolates collected in the United States in 1999, 2001, and 2002. *J Clin Microbiol* **44**(3):999-1017.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to mutliple testing. *J R Statist Soc* **57**(1):289-300.

Bentley, D. R., *et al*. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218):53-9.

Bentley, S. D., *et al*. (2006). Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* **2**(3):e31.

Bentley, S. D., *et al*. (2007). Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet* **3**(2):e23.

Beres, S. B., *et al*. (2010). Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc Natl Acad Sci U S A* **107**(9):4371-6.

Berge, M., *et al*. (2003). Transformation of *Streptococcus pneumoniae* relies on DprA- and RecA-dependent protection of incoming DNA single strands. *Mol Microbiol* **50**(2):527-36.

Berk, S. L., *et al*. (1985). Type 8 pneumococcal pneumonia: an outbreak on an oncology ward. *South Med J* **78**(2):159-61.

Berman, H. M. and P. R. Young (1981). The interaction of intercalating drugs with nucleic acids. *Annu Rev Biophys Bioeng* **10**:87-114.

Bernhart, S. H., *et al*. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**:474.

Bernheimer, H. P. and J. G. Tiraby (1976). Inhibition of phage infection by pneumococcus capsule. *Virology* **73**(1):308-9.

Berriman, M. and K. Rutherford (2003). Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4**(2):124-32.

Berry, A. M., *et al*. (1989). Presence of a small plasmid in clinical isolates of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **53**(3):275-8.

Berry, A. M. and J. C. Paton (2000). Additive attenuation of virulence of *Streptococcus pneumoniae* by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. *Infect Immun* **68**(1):133-40.

Bibb, M. J., *et al*. (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* **30**(1-3):157-66.

Bierne, H., *et al*. (1997). *uvrD* mutations enhance tandem repeat deletion in the *Escherichia coli* chromosome via SOS induction of the RecF recombination pathway. *Mol Microbiol* **26**(3):557-67.

Bito, H., *et al*. (1994). Cloning, expression and tissue distribution of rat platelet-activating-factor-receptor cDNA. *Eur J Biochem* **221**(1):211-8.

Block, S. L., *et al*. (2004). Community-wide vaccination with the heptavalent pneumococcal conjugate significantly alters the microbiology of acute otitis media. *Pediatr Infect Dis J* **23**(9):829-33.

Blomberg, C., *et al*. (2009). Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. *J Infect Dis* **199**(7):1032-42.

Bluestone, C. and J. Klein (2007). *Otitis media in infants and children*. USA, People's Medical Publishing House.

Bochner, B. R. (2009). Global phenotypic characterization of bacteria. *FEMS Microbiol Rev* **33**(1):191-205.

Bogaert, D., *et al*. (2004a). *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *Lancet Infect Dis* **4**(3):144-54.

Bogaert, D., *et al*. (2001). Pneumococcal carriage in children in The Netherlands: a molecular epidemiological study. *J Clin Microbiol* **39**(9):3316-20.

Bogaert, D., *et al*. (2005). Epidemiology of nasopharyngeal carriage of *Neisseria meningitidis* in healthy Dutch children. *Clin Infect Dis* **40**(6):899-902.

Bogaert, D., *et al*. (2004b). Colonisation by *Streptococcus pneumoniae* and *Staphylococcus aureus* in healthy children. *Lancet* **363**(9424):1871-2.

Bolan, G., *et al*. (1986). Pneumococcal vaccine efficacy in selected populations in the United States. *Ann Intern Med* **104**(1):1-6.

Bonfield, J. K., *et al*. (1995). A new DNA sequence assembly program. *Nucleic Acids Res* **23**(24):4992-9.

Brandt, J., *et al*. (2002). Invasive pneumococcal disease and hemolytic uremic syndrome. *Pediatrics* **110**(2 Pt 1):371-6.

Bratcher, P. E., *et al*. (2010). Identification of natural pneumococcal isolates expressing serotype 6D by genetic, biochemical and serological characterization. *Microbiology* **156**(Pt 2):555-60.

Briles, D. E., *et al*. (1992). Strong association between capsular type and virulence for mice among human isolates of *Streptococcus pneumoniae*. *Infect Immun* **60**(1):111-6.

Brochet, M., *et al*. (2008). Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* **105**(41):15961-6.

Brown, J. S., *et al*. (2001). A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol Microbiol* **40**(3):572-85.

Brown, J. S., *et al*. (2004). A locus contained within a variable region of pneumococcal pathogenicity island 1 contributes to virulence in mice. *Infect Immun* **72**(3):1587-93.

Brueggemann, A. B., *et al*. (2003). Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis* **187**(9):1424-32.

Brueggemann, A. B., *et al*. (2007). Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathog* **3**(11):e168.

Brueggemann, A. B., *et al*. (2004). Temporal and geographic stability of the serogroup-specific invasive disease potential of *Streptococcus pneumoniae* in children. *J Infect Dis* **190**(7):1203-11.

Brugger, S. D., *et al*. (2010). Multiple colonization with *S. pneumoniae* before and after introduction of the seven-valent conjugated pneumococcal polysaccharide vaccine. *PLoS One* **5**(7):e11638.

Brundish, D. E. and J. Baddiley (1968). Pneumococcal C-substance, a ribitol teichoic acid containing choline phosphate. *Biochem J* **110**(3):573-82.

Bryan, J. P., *et al*. (1990). Etiology and mortality of bacterial meningitis in northeastern Brazil. *Rev Infect Dis* **12**(1):128-35.

Buchanan, R. and N. Gibbons, Eds. (1974). Bergey's Manual of Determinative Bacteriology. Baltimore, Williams & Wilkins Co.

Burman, L. A., *et al*. (1985). Invasive pneumococcal infections: incidence, predisposing factors, and prognosis. *Rev Infect Dis* **7**(2):133-42.

Burman, L. A., *et al*. (1991). Diagnosis of pneumonia by cultures, bacterial and viral antigen detection tests, and serology with special reference to antibodies against pneumococcal antigens. *J Infect Dis* **163**(5):1087-93.

Buu-Hoi, A. and T. Horodniceanu (1980). Conjugative transfer of multiple antibiotic resistance markers in *Streptococcus pneumoniae*. *J Bacteriol* **143**(1):313-20.

Byington, C. L., *et al*. (2006). Impact of the pneumococcal conjugate vaccine on pneumococcal parapneumonic empyema. *Pediatr Infect Dis J* **25**(3):250-4.

Byington, C. L., *et al*. (2002). An epidemiological investigation of a sustained high rate of pediatric parapneumonic empyema: risk factors and microbiological associations. *Clin Infect Dis* **34**(4):434-40.

Calverley, P. M., *et al*. (2007). Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* **356**(8):775-89.

Camilli, R., *et al*. (2006). Zinc metalloproteinase genes in clinical isolates of *Streptococcus pneumoniae*: association of the full array with a clonal cluster comprising serotypes 8 and 11A. *Microbiology* **152**(Pt 2):313-21.

Camou, T., *et al*. (1998). The apparent importation of penicillin-resistant capsular type 14 Spanish/French clone of *Streptococcus pneumoniae* into Uruguay in the early 1990s. *Microb Drug Resist* **4**(3):219-24.

Campbell, E. A., *et al*. (1998). A competence regulon in *Streptococcus pneumoniae* revealed by genomic analysis. *Mol Microbiol* **27**(5):929-39.

Canvin, J. R., *et al*. (1995). The role of pneumolysin and autolysin in the pathology of pneumonia and septicemia in mice infected with a type 2 pneumococcus. *J Infect Dis* **172**(1):119-23.

Capdevila, O., *et al*. (2001). Pneumococcal peritonitis in adult patients: report of 64 cases with special reference to emergence of antibiotic resistance. *Arch Intern Med* **161**(14):1742-8.

Cardozo, D. M., *et al*. (2008). Prevalence and risk factors for nasopharyngeal carriage of *Streptococcus pneumoniae* among adolescents. *J Med Microbiol* **57**(Pt 2):185-9.

Carver, T., *et al*. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**(23):2672-6.

Carver, T. J., *et al*. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics* **21**(16):3422-3.

Casey, J. R. and M. E. Pichichero (2004). Changes in frequency and pathogens causing acute otitis media in 1995-2003. *Pediatr Infect Dis J* **23**(9):824-8.

Castaneda, E., *et al*. (1998). Penicillin-resistant *Streptococcus pneumoniae* in Colombia: presence of international epidemic clones. Colombian pneumococcal study group. *Microb Drug Resist* **4**(3):233-9.

Chen, C., *et al*. (2007). A glimpse of streptococcal toxic shock syndrome from comparative genomics of *S. suis* 2 Chinese isolates. *PLoS One* **2**(3):e315.

Chen, F. M., *et al*. (1998). Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive *Streptococcus pneumoniae* infections. *Am J Epidemiol* **148**(12):1212-8.

Chen, I., *et al*. (2005). The ins and outs of DNA transfer in bacteria. *Science* **310**(5753):1456-60.

Chen, K. and A. W. Ravin (1966). Heterospecific transformation of pneumococcus and streptococcus. *J Mol Biol* **22**:123-134.

Chen, Z., *et al*. (2008). Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature* **453**(7194):489-4.

Chilcote, R. R., *et al*. (1976). Septicemia and meningitis in children splenectomized for Hodgkin's disease. *N Engl J Med* **295**(15):798-800.

Chotpitayasunondh, T. (1994). Bacterial meningitis in children: etiology and clinical features, an 11-year review of 618 cases. *Southeast Asian J Trop Med Public Health* **25**(1):107-15.

Claverys, J. P., *et al*. (2000). Is the Ami-AliA/B oligopeptide permease of *Streptococcus pneumoniae* involved in sensing environmental conditions? *Res Microbiol* **151**(6):457-63.

Claverys, J. P. and L. S. Havarstein (2002). Extracellular-peptide control of competence for genetic transformation in *Streptococcus pneumoniae*. *Front Biosci* **7**:d1798-814.

Claverys, J. P., *et al*. (1980). Transformation of *Streptococcus pneumoniae* with *S. pneumoniae*-lambda phage hybrid DNA: induction of deletions. *Proc Natl Acad Sci U S A* **77**(6):3534-8.

Claverys, J. P., *et al*. (2007). Competence-induced fratricide in streptococci. *Mol Microbiol* **64**(6):1423-33.

Claverys, J. P., *et al*. (1981). Base specificity of mismatch repair in *Streptococcus pneumoniae*. *Nucleic Acids Res* **9**(10):2267-80.

Claverys, J. P., *et al*. (1983). Mismatch repair in *Streptococcus pneumoniae*: relationship between base mismatches and transformation efficiencies. *Proc Natl Acad Sci U S A* **80**(19):5956-60.

Claverys, J. P., *et al*. (2006). Induction of competence regulons as a general response to stress in Gram-positive bacteria. *Annu Rev Microbiol* **60**:451-75.

Cloonan, N., *et al*. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**(7):613-9.

Cloonan, N. and S. M. Grimmond (2008). Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biol* **9**(9):234.

Cochetti, I., *et al*. (2008). *erm*(B)-carrying elements in tetracycline-resistant pneumococci and correspondence between Tn*1545* and Tn*6003*. *Antimicrob Agents Chemother* **52**(4):1285-90.

Cochetti, I., *et al*. (2007). New Tn*916*-related elements causing *erm*(B)-mediated erythromycin resistance in tetracycline-susceptible pneumococci. *J Antimicrob Chemother* **60**(1):127-31.

Coffey, T. J., *et al*. (1993). Horizontal spread of an altered penicillin-binding protein 2B gene between *Streptococcus pneumoniae* and *Streptococcus oralis*. *FEMS Microbiol Lett* **110**(3):335-9.

Coffey, T. J., *et al*. (1991). Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae*. *Mol Microbiol* **5**(9):2255-60.

Coffey, T. J., *et al*. (1998a). Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. *Mol Microbiol* **27**(1):73-83.

Coffey, T. J., *et al*. (1998b). Serotype 19A variants of the Spanish serotype 23F multiresistant clone of *Streptococcus pneumoniae*. *Microb Drug Resist* **4**(1):51-5.

Cohen, R., *et al*. (1997). Change in nasopharyngeal carriage of *Streptococcus pneumoniae* resulting from antibiotic therapy for acute otitis media in children. *Pediatr Infect Dis J* **16**(6):555-60.

Coles, C. L., *et al*. (2001). Pneumococcal nasopharyngeal colonization in young South Indian infants. *Pediatr Infect Dis J* **20**(3):289-95.

Collatz, E., *et al*. (1984). Characterization of high-level aminoglycoside resistance in a strain of *Streptococcus pneumoniae*. *J Gen Microbiol* **130**(7):1665-71.

Collins, H. S., *et al*. (1948). Aureomycin in treatment of pneumococcal pneumonia and meningococcemie. *Proc Soc Exp Biol Med* **69**(2):263-5.

Connell, S. R., *et al*. (2003). Ribosomal protection proteins and their mechanism of tetracycline resistance. *Antimicrob Agents Chemother* **47**(12):3675-81.

Corriere, J. N., Jr. and L. I. Lipshultz (1974). Pneumococcal nephritis. *Urology* **3**(5):557-61.

Corso, A., *et al*. (1998). Molecular characterization of penicillin-resistant *Streptococcus pneumoniae* isolates causing respiratory disease in the United States. *Microb Drug Resist* **4**(4):325-37.

Cortaza, G., *et al*. (1983). A plasmid in a drug-resistant clinical isolate of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **17**:55-57.

Cortese, M. M., *et al*. (1992). High incidence rates of invasive pneumococcal disease in the White Mountain Apache population. *Arch Intern Med* **152**(11):2277-82.

Courvalin, P., *et al*. (1985). Multiplicity of macrolide-lincosamide-streptogramin antibiotic resistance determinants. *J Antimicrob Chemother* **16 Suppl A**:91-100.

Cox, D. (1969). Some sampling problems in technology. New Developments in Survey Sampling. N. Johnson and H. Smith. New York, Wiley**:** 506-527.

Cundell, D. R., *et al*. (1995a). *Streptococcus pneumoniae* anchor to activated human cells by the receptor for platelet-activating factor. *Nature* **377**(6548):435-8.

Cundell, D. R., *et al*. (1995b). Relationship between colonial morphology and adherence of *Streptococcus pneumoniae*. *Infect Immun* **63**(3):757-61.

Cvitkovitch, D. G., *et al*. (1997). Role of the citrate pathway in glutamate biosynthesis by *Streptococcus mutans*. *J Bacteriol* **179**(3):650-5.

Cybulska, J., *et al*. (1970). Prevalence of types of *Diplococcus pneumoniae* and their susceptibility to 30 antibiotics. *Chemotherapy* **15**(5):304-16.

Dagan, R. and K. P. Klugman (2008). Impact of conjugate pneumococcal vaccines on antibiotic resistance. *Lancet Infect Dis* **8**(12):785-95.

Dagkessamanskaia, A., *et al*. (2004). Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae*: competence triggers stationary phase autolysis of *ciaR* mutant cells. *Mol Microbiol* **51**(4):1071-86.

Danecek, P., *et al*. (2011). The variant call format and VCFtools. *Bioinformatics* **27**(15):2156-2158.

Dang-Van, A., *et al*. (1978). Chloramphenicol resistance in *Streptococcus pneumoniae*: enzymatic acetylation and possible plasmid linkage. *Antimicrob Agents Chemother* **13**(4):577-83.

Danner, D. B., *et al*. (1980). An eleven-base-pair sequence determines the specificity of DNA uptake in *Haemophilus* transformation. *Gene* **11**(3-4):311-8.

Darling, A. E., *et al*. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**(6):e11147.

Dave, S., *et al*. (2004). Dual roles of PspC, a surface protein of *Streptococcus pneumoniae*, in binding human secretory IgA and factor H. *J Immunol* **173**(1):471-7.

Dawid, S., *et al*. (2007). The *blp* bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both *in vitro* and *in vivo*. *Infect Immun* **75**(1):443-51.

De Gregorio, E., *et al*. (2002). The abundant class of nemis repeats provides RNA substrates for ribonuclease III in *Neisseriae*. *Biochim Biophys Acta* **1576**(1-2):39-44.

De Gregorio, E., *et al*. (2005). Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J Bacteriol* **187**(23):7945-54.

de Hoon, M. J., *et al*. (2005). Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* **1**(3):e25.

de Saizieu, A., *et al*. (2000). Microarray-based identification of a novel *Streptococcus pneumoniae* regulon controlled by an autoinduced peptide. *J Bacteriol* **182**(17):4696-703.

de Visser, J. A. and S. F. Elena (2007). The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat Rev Genet* **8**(2):139-49.

Del Grosso, M., *et al*. (2006). The *mef*(E)-carrying genetic element (mega) of *Streptococcus pneumoniae*: insertion sites and association with other genetic elements. *Antimicrob Agents Chemother* **50**(10):3361-6.

Del Grosso, M., *et al*. (2007). The macrolide resistance genes *erm*(B) and *mef*(E) are carried by Tn*2010* in dual-gene *Streptococcus pneumoniae* isolates belonging to clonal complex CC271. *Antimicrob Agents Chemother* **51**(11):4184-6.

Delcher, A. L., *et al*. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**(6):673-9.

Delihas, N. (2008). Small mobile sequences in bacteria display diverse structure/function motifs. *Mol Microbiol* **67**(3):475-81.

Denapaite, D., *et al*. (2010). The genome of *Streptococcus mitis* B6--what is a commensal? *PLoS One* **5**(2):e9426.

Didelot, X. and D. Falush (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**(3):1251-66.

Dillard, J. P., *et al*. (1995). Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. *J Exp Med* **181**(3):973-83.

Ding, F., *et al*. (2009). Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant *Streptococcus pneumoniae* serotype 14. *BMC Genomics* **10**:158.

Dixon, J. M., *et al*. (1977). Detection and prevalence of pneumococci with increased resistance to penicillin. *Can Med Assoc J* **117**(10):1159-61.

Dochez, A. and L. Gillespie (1913). A biologic classification of pneumococci by means of immunity reactions. *JAMA* **61**(10):727-732.

Dochez, A. R. and O. T. Avery (1917). The elaboration of specific soluble substance by pneumococcus during growth. *J Exp Med* **26**(4):477-93.

Doit, C., *et al*. (2010). Epidemiology of pediatric community-acquired bloodstream infections in a children hospital in Paris, France, 2001 to 2008. *Diagn Microbiol Infect Dis* **66**(3):332-5.

Donaldson, S. S., *et al*. (1978). Bacterial infections in pediatric Hodgkin's disease: relationship to radiotherapy, chemotherapy and splenectomy. *Cancer* **41**(5):1949-58.

Donati, C., *et al*. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* **11**(10):R107.

Dopazo, J., *et al*. (2001). Annotated draft genomic sequence from a *Streptococcus pneumoniae* type 19F clinical isolate. *Microb Drug Resist* **7**(2):99-125.

Douglas, R. M., *et al*. (1983). Antibody response to pneumococcal vaccination in children younger than five years of age. *J Infect Dis* **148**(1):131-7.

Dowson, C. G., *et al*. (1993). Evolution of penicillin resistance in *Streptococcus pneumoniae*; the role of *Streptococcus mitis* in the formation of a low affinity PBP2B in *S. pneumoniae*. *Mol Microbiol* **9**(3):635-43.

Dowson, C. G., *et al*. (1989). Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A* **86**(22):8842-6.

Dowson, C. G., *et al*. (1990). Penicillin-resistant viridans streptococci have obtained altered penicillin-binding protein genes from penicillin-resistant strains of *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A* **87**(15):5858-62.

Dowson, C. G., *et al*. (1994). Genetics of oxacillin resistance in clinical isolates of *Streptococcus pneumoniae* that are oxacillin resistant and penicillin susceptible. *Antimicrob Agents Chemother* **38**(1):49-53.

Drummond, A. J., *et al*. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**(5):e88.

Drummond, A. J. and A. Rambaut (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**:214.

Drummond, A. J., *et al*. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**(5):1185-92.

Dublanchet, A. and R. Durieux (1979). Isolement d'une souche de *Streptococcus pneumoniae* multirésistante aux antibiotiques. *Nouv Presse Med* **8**(11):872.

Dubnau, D. (1999). DNA uptake in bacteria. *Annu Rev Microbiol* **53**:217-44.

Dubos, R. and O. T. Avery (1931). Decomposition of the capsular polysaccharide of pneumococcus type III by a bacterial enzyme. *J Exp Med* **54**(1):51-71.

Dunais, B., *et al*. (2003). Influence of child care on nasopharyngeal carriage of *Streptococcus pneumoniae* and *Haemophilus influenzae*. *Pediatr Infect Dis J* **22**(7):589-92.

Eastham, K. M., *et al*. (2004). Clinical features, aetiology and outcome of empyema in children in the north east of England. *Thorax* **59**(6):522-5.

Echaniz-Aviles, G., *et al*. (1998). Predominance of the multiresistant 23F international clone of *Streptococcus pneumoniae* among isolates from Mexico. *Microb Drug Resist* **4**(3):241-6.

Eddy, S. R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol* **4**(5):e1000069.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5):1792-7.

Ekwurzel, G., *et al*. (1938). Studies on immunizing substances in pneumococci. VIII. Report on field tests to determine the prophylactic value of a pneumococcus antigen. *Public Health Rep* **53**:1877-1893.

El Garch, F., *et al*. (2010). Fluoroquinolones induce the expression of *patA* and *patB*, which encode ABC efflux pumps in *Streptococcus pneumoniae*. *J Antimicrob Chemother* **65**(10):2076-82.

Eltringham, G., *et al*. (2003). Culture-negative childhood empyema is usually due to penicillin-sensitive *Streptococcus pneumoniae* capsular serotype 1. *J Clin Microbiol* **41**(1):521-2.

Enright, A. J., *et al*. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**(7):1575-84.

Enright, M., *et al*. (1998). Molecular evolution of rifampicin resistance in *Streptococcus pneumoniae*. *Microb Drug Resist* **4**(1):65-70.

Enright, M. C., *et al*. (1999). The three major Spanish clones of penicillin-resistant *Streptococcus pneumoniae* are the most common clones recovered in recent cases of meningitis in Spain. *J Clin Microbiol* **37**(10):3210-6.

Enright, M. C. and B. G. Spratt (1998). A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology* **144 (Pt 11)**:3049-60.

Enright, M. C. and B. G. Spratt (1999). Extensive variation in the *ddl* gene of penicillin-resistant *Streptococcus pneumoniae* results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. *Mol Biol Evol* **16**(12):1687-95.

Ephrussi-Taylor, H., *et al*. (1965). Genetic recombination in DNA-induced transformation of pneumococcus. I. the problem of relative efficiency of transforming factors. *Genetics* **51**(3):455-75.

Erill, I., *et al*. (2007). Aeons of distress: an evolutionary perspective on the bacterial SOS response. *FEMS Microbiol Rev* **31**(6):637-56.

Ernst, P., *et al*. (2007). Inhaled corticosteroid use in chronic obstructive pulmonary disease and the risk of hospitalization for pneumonia. *Am J Respir Crit Care Med* **176**(2):162-6.

Ertugrul, N., *et al*. (1997). BOX-polymerase chain reaction-based DNA analysis of nonserotypeable *Streptococcus pneumoniae* implicated in outbreaks of conjunctivitis. *J Infect Dis* **176**(5):1401-5.

Espeli, O., *et al*. (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *J Mol Biol* **314**(3):375-86.

Evans, G. and W. Gaisford (1938). Treatment of pneumonia with 2-(*p*-aminobenzenesulphonamido)pyridine. *Lancet* **2**(22):14-19.

Evans, W. and D. Hansman (1963). Tetracycline-resistant pneumococcus. *Lancet* **1**:451.

Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. *Clin Microbiol Rev* **15**(4):613-30.

Fang, G. D., *et al*. (1990). New and emerging etiologies for community-acquired pneumonia with implications for therapy. A prospective multicenter study of 359 cases. *Medicine (Baltimore)* **69**(5):307-16.

Feikin, D. R., *et al*. (2000). Mortality from invasive pneumococcal pneumonia in the era of antibiotic resistance, 1995-1997. *Am J Public Health* **90**(2):223-9.

Feil, E. J., *et al*. (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A* **98**(1):182-7.

Feil, E. J., *et al*. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154**(4):1439-50.

Feinstein, S. I. and K. B. Low (1986). Hyper-recombining recipient strains in bacterial conjugation. *Genetics* **113**(1):13-33.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* **78**(2):737-56.

Felsenstein, J. (1989). PHYLIP - phylogeny inference package. *Cladistics* **5**:164-166.

Felsenstein, J. and S. Yokoyama (1976). The evolutionary advantage of recombination. II Individual selection for recombination. *Genetics* **83**:845-859.

Feng, J., *et al*. (2009). Genome sequencing of linezolid-resistant *Streptococcus pneumoniae* mutants reveals novel mechanisms of resistance. *Genome Res* **19**(7):1214-23.

Fenoll, A., *et al*. (1991). Serotype distribution and antimicrobial resistance of *Streptococcus pneumoniae* isolates causing systemic infections in Spain, 1979-1989. *Rev Infect Dis* **13**(1):56-60.

Fenoll, A., *et al*. (1994). Molecular basis of the optochin-sensitive phenotype of pneumococcus: characterization of the genes encoding the F0 complex of the *Streptococcus pneumoniae* and *Streptococcus oralis* H(+)-ATPases. *Mol Microbiol* **12**(4):587-98.

Ferguson, A. D., *et al*. (1996). The clinical course and management of thoracic empyema. *QJM* **89**(4):285-9.

Ferrandiz, M. J., *et al*. (2005). New mutations and horizontal transfer of *rpoB* among rifampin-resistant *Streptococcus pneumoniae* from four Spanish hospitals. *Antimicrob Agents Chemother* **49**(6):2237-45.

Fiers, W., *et al*. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**(5551):500-7.

Figueiredo, A. M., *et al*. (1995). Novel penicillin-resistant clones of *Streptococcus pneumoniae* in the Czech Republic and in Slovakia. *Microb Drug Resist* **1**(1):71-8.

Filipe, S. R., *et al*. (2000). Distribution of the mosaic structured *murM* genes among natural populations of *Streptococcus pneumoniae*. *J Bacteriol* **182**(23):6798-805.

Filipe, S. R. and A. Tomasz (2000). Inhibition of the expression of penicillin resistance in Streptococcus pneumoniae by inactivation of cell wall muropeptide branching genes. *Proc Natl Acad Sci U S A* **97**(9):4891-6.

Finland, M. and M. W. Barnes (1977). Changes in occurrence of capsular serotypes of Streptococcus pneumoniae at Boston City Hospital during selected years between 1935 and 1974. *J Clin Microbiol* **5**(2):154-66.

Finland, M., *et al*. (1976). Susceptibility of pneumococci and *Haemophilus influenzae* to antibacterial agents. *Antimicrob Agents Chemother* **9**(2):274-87.

Fisher, R. (1930). *The genetical theory of natural selection*. Oxford, Oxford University Press.

Fitch, D. H. and M. Goodman (1991). Phylogenetic scanning: a computer-assisted algorithm for mapping gene conversions and other recombinational events. *Comput Appl Biosci* **7**(2):207-15.

Fitzmaurice, W. P., *et al*. (1984). Characterization of recognition sites on bacteriophage HP1c1 DNA which interact with the DNA uptake system of *Haemophilus influenzae* Rd. *Gene* **31**(1-3):187-96.

Flamaing, J., *et al*. (2010). Pneumococcal colonization in older persons in a nonoutbreak setting. *J Am Geriatr Soc* **58**(2):396-8.

Flannery, B., *et al*. (2006). Changes in invasive Pneumococcal disease among HIV-infected adults living in the era of childhood pneumococcal immunization. *Ann Intern Med* **144**(1):1-9.

Fleischmann, R. D., *et al*. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**(5223):496-512.

Foa, P. and G. Bordoni-Uffreduzzi (1888). Über die Aetiologie der "Meningitis cerebrospinalis epidemica". *Ztschr f Hyg u Infektionskr* **4**(1):67.

Foss Abrahamsen, A., *et al*. (1997). Systemic pneumococcal disease after staging splenectomy for Hodgkin's disease 1969-1980 without pneumococcal vaccine protection: a follow-up study 1994. *Eur J Haematol* **58**(2):73-7.

Fox, M. S. and M. K. Allen (1964). On the mechanism of deoxyribonucleate integration in pneumococcal transformation. *Proc Natl Acad Sci U S A* **52**:412-9.

Francis, T. and W. S. Tillett (1930). Cutaneous reactions in pneumonia. the development of antibodies following the intradermal injection of type-specific polysaccharide. *J Exp Med* **52**(4):573-85.

Fränkel, A. (1884). Über die genuine Pneumonie. *Verhandlungen des Congresses für Innere Medizin, Dritter Congress* **3**:17-31.

Fränkel, A. (1886). Weitere Beiträge zur Lehre von den Mikrococcen der genuinen fibrinosen Pneumonic. *Ztschr f klin Med* **11**:437.

Frankel, R. E., *et al*. (1996). Invasive pneumococcal disease: clinical features, serotypes, and antimicrobial resistance patterns in cases involving patients with and without human immunodeficiency virus infection. *Clin Infect Dis* **23**(3):577-84.

Fraser, C., *et al*. (2007). Recombination and the nature of bacterial speciation. *Science* **315**(5811):476-80.

Friedländer, C. (1882). Über die Schizomyceten bei der acuten fibrinosen Pneumonie. *Virchows Arch f Path Anat* **87**(2):319.

Friedländer, C. (1883). Die Mikrokokken der Pneumonie. *Fortschr Med* **1**(22):715-733.

Friedländer, C. (1886). Weitere Arbeiten über die Schizomyceten der Pneumonie und der Meningitis. *Fortschr d Med* **4**(21):702.

Gamaléia, N. (1888). Sur l'étiologie de la pneumonie fibrineuse chez l'homme. *Ann Inst Pasteur* **2**(8):440-459.

Garau, J., *et al*. (1981). Chloramphenicol-resistant pneumococci. *Lancet* **2**(8238):147-8.

Garcia, E., *et al*. (1988). Molecular evolution of lytic enzymes of *Streptococcus pneumoniae* and its bacteriophages. *Proc Natl Acad Sci U S A* **85**(3):914-8.

Garcia-Bustos, J. and A. Tomasz (1990). A biological price of antibiotic resistance: major changes in the peptidoglycan structure of penicillin-resistant pneumococci. *Proc Natl Acad Sci U S A* **87**(14):5415-9.

Garcia-Bustos, J. F., *et al*. (1988). Altered peptidoglycan structure in a pneumococcal transformant resistant to penicillin. *J Bacteriol* **170**(5):2143-7.

Garvey, M. I., *et al*. (2010). Overexpression of *patA* and *patB*, which encode ABC transporters, is associated with fluoroquinolone resistance in clinical isolates of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **55**(1):190-6.

Garvey, M. I. and L. J. Piddock (2008). The efflux pump inhibitor reserpine selects multidrug-resistant *Streptococcus pneumoniae* strains that overexpress the

ABC transporters PatA and PatB. *Antimicrob Agents Chemother* **52**(5):1677-85.

Gasc, A. M., *et al*. (1987). Mismatch repair during pneumococcal transformation of small deletions produced by site-directed mutagenesis. *Mol Gen Genet* **210**(2):369-72.

Gasc, A. M. and A. M. Sicard (1986). Frame-shift mutants induced by quinacrine are recognized by the mismatch repair system in *Streptococcus pneumoniae*. *Mol Gen Genet* **203**(2):269-73.

Gasc, A. M., *et al*. (1989). Repair of single- and multiple-substitution mismatches during recombination in *Streptococcus pneumoniae*. *Genetics* **121**(1):29-36.

Gay, K. and D. S. Stephens (2001). Structure and dissemination of a chromosomal insertion element encoding macrolide efflux in *Streptococcus pneumoniae*. *J Infect Dis* **184**(1):56-65.

George, R. H., *et al*. (1981). Multiresistant pneumococci. *Lancet* **2**(8249):751-2.

Gianfaldoni, C., *et al*. (2007). *Streptococcus pneumoniae* pilus subunits protect mice against lethal challenge. *Infect Immun* **75**(2):1059-62.

Gill, M. J., *et al*. (1999). Identification of an efflux pump gene, *pmrA*, associated with fluoroquinolone resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **43**(1):187-9.

Gilson, E., *et al*. (1991). Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res* **19**(7):1375-83.

Gindreau, E., *et al*. (2000). MM1, a temperate bacteriophage of the type 23F Spanish/USA multiresistant epidemic clone of *Streptococcus pneumoniae*: structural analysis of the site-specific integration system. *J Virol* **74**(17):7803-13.

Gladstone, R. A., *et al*. (2011). Continued control of pneumococcal disease in the UK - the impact of vaccination. *J Med Microbiol* **60**(Pt 1):1-8.

Goebel, W. F. and O. T. Avery (1931). Chemo-immunological studies on conjugated carbohydrate-proteins : iv. the synthesis of thep-aminobenzyl ether of the soluble specific substance of type III pneumococcus and its coupling with protein. *J Exp Med* **54**(3):431-6.

Goldsmith, C. E., *et al*. (1998). Increased incidence of ciprofloxacin resistance in penicillin-resistant pneumococci in Northern Ireland. *J Antimicrob Chemother* **41**(3):420-1.

Goldstein, F. W. and J. F. Acar (1996). Antimicrobial resistance among lower respiratory tract isolates of *Streptococcus pneumoniae*: results of a 1992-93 western Europe and USA collaborative surveillance study. The Alexander Project Collaborative Group. *J Antimicrob Chemother* **38 Suppl A**:71-84.

Goodman, S. D. and J. J. Scocca (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A* **85**(18):6982-6.

Gordon, M. A., *et al*. (2001). Bacteraemia and mortality among adult medical admissions in Malawi--predominance of non-typhi salmonellae and *Streptococcus pneumoniae*. *J Infect* **42**(1):44-9.

Gram, C. (1884). Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. *Fortschr d Med* **2**:185-189.

Gransden, W. R., *et al*. (1985). Pneumococcal bacteraemia: 325 episodes diagnosed at St Thomas's Hospital. *Br Med J (Clin Res Ed)* **290**(6467):505-8.

Grassly, N. C. and E. C. Holmes (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol* **14**(3):239-47.

Gratten, M., *et al*. (1986). Colonisation of *Haemophilus influenzae* and *Streptococcus pneumoniae* in the upper respiratory tract of neonates in Papua New Guinea: primary acquisition, duration of carriage, and relationship to carriage in mothers. *Biol Neonate* **50**(2):114-20.

Gratten, M., *et al*. (1989). Multiple colonization of the upper respiratory tract of Papua New Guinea children with *Haemophilus influenzae* and *Streptococcus pneumoniae*. *Southeast Asian J Trop Med Public Health* **20**(4):501-9.

Gratten, M., *et al*. (1993). An outbreak of serotype 1 *Streptococcus pneumoniae* infection in central Australia. *Med J Aust* **158**(5):340-2.

Gratten, M., *et al*. (1980). High prevalence of penicillin-insensitive pneumococci in Port Moresby, Papua New Guinea. *Lancet* **2**(8187):192-5.

Gray, B. M., *et al*. (1980). Epidemiologic studies of *Streptococcus pneumoniae* in infants: acquisition, carriage, and infection during the first 24 months of life. *J Infect Dis* **142**(6):923-33.

Gray, B. M., *et al*. (1982). Epidemiologic studies of *Streptococcus pneumoniae* in infants. The effects of season and age on pneumococcal acquisition and carriage in the first 24 months of life. *Am J Epidemiol* **116**(4):692-703.

Greenberg, D., *et al*. (2006). The contribution of smoking and exposure to tobacco smoke to *Streptococcus pneumoniae* and *Haemophilus influenzae* carriage in children and their mothers. *Clin Infect Dis* **42**(7):897-903.

Griffith, F. (1928). The significance of pneumococcal types. *J Hygiene* **27**:113-159.

Guell, M., *et al*. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science* **326**(5957):1268-71.

Guenzi, E., *et al*. (1994). A two-component signal-transducing system is involved in competence and penicillin susceptibility in laboratory mutants of *Streptococcus pneumoniae*. *Mol Microbiol* **12**(3):505-15.

Guevara, M., *et al*. (2009). Changing epidemiology of invasive pneumococcal disease following increased coverage with the heptavalent conjugate vaccine in Navarre, Spain. *Clin Microbiol Infect* **15**(11):1013-9.

Guiral, S., *et al*. (2005). Competence-programmed predation of noncompetent cells in the human pathogen *Streptococcus pneumoniae*: genetic requirements. *Proc Natl Acad Sci U S A* **102**(24):8710-5.

Gundel, M. and F. Schwarz (1932). Studien über die Bakterien flora der Oberen Atmungswege Neugeborener unter besonderer Berucksichtigun ihrer Bedeutung fur das Pneumoniaeproblem. *Z Hyg Infekt* **1932**(113):411-436.

Gurney, T., Jr. and M. S. Fox (1968). Physical and genetic hybrids formed in bacterial transformation. *J Mol Biol* **32**(1):83-100.

Gutierrez-Preciado, A., *et al*. (2007). Comparison of tryptophan biosynthetic operon regulation in different Gram-positive bacterial species. *Trends Genet* **23**(9):422-6.

Haber, L. T., *et al*. (1988). Nucleotide sequence of the *Salmonella typhimurium mutS* gene required for mismatch repair: homology of MutS and HexA of *Streptococcus pneumoniae*. *J Bacteriol* **170**(1):197-202.

Hakenbeck, R., *et al*. (1991a). Antigenic variation of penicillin-binding proteins from penicillin-resistant clinical strains of *Streptococcus pneumoniae*. *J Infect Dis* **164**(2):313-9.

Hakenbeck, R., *et al*. (1991b). Variability of penicillin-binding proteins from penicillin-sensitive *Streptococcus pneumoniae*. *J Infect Dis* **164**(2):307-12.

Hakenbeck, R., *et al*. (1986). Antibodies against the benzylpenicilloyl moiety as a probe for penicillin-binding proteins. *Eur J Biochem* **157**(1):101-6.

Hakenbeck, R., *et al*. (1980). Multiple changes of penicillin-binding proteins in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **17**(3):364-71.

Hamburger, M., *et al*. (1943). The occurrence of sulfonamide-resistant pneumococci in clinical practice. *J Infect Dis* **73**(1):12-30.

Hanage, W. P., *et al*. (2006). Sequences, sequence clusters and bacterial species. *Philos Trans R Soc Lond B Biol Sci* **361**(1475):1917-27.

Hanna, J. N., *et al*. (2008). Invasive pneumococcal disease in Indigenous people in north Queensland: an update, 2005-2007. *Med J Aust* **189**(1):43-6.

Hansman, D. and M. Bullen (1967). A resistant pneumococcus. *Lancet* **2**:264-265.

Hansman, D., *et al*. (1971). Increased resistance to penicillin of pneumococci isolated from man. *N Engl J Med* **284**(4):175-7.

Harboe, Z. B., *et al*. (2009). Pneumococcal serotypes and mortality following invasive pneumococcal disease: a population-based cohort study. *PLoS Med* **6**(5):e1000081.

Hardie, W., *et al*. (1996). Pneumococcal pleural empyemas in children. *Clin Infect Dis* **22**(6):1057-63.

Hare, K. M., *et al*. (2008). Random colony selection versus colony morphology for detection of multiple pneumococcal serotypes in nasopharyngeal swabs. *Pediatr Infect Dis J* **27**(2):178-80.

Harris, H. (1966). Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* **164**(995):298-310.

Harris, S. R., *et al*. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**(5964):469-74.

Hava, D. L. and A. Camilli (2002). Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. *Mol Microbiol* **45**(5):1389-406.

Havarstein, L. S. (2010). Increasing competence in the genus Streptococcus. *Mol Microbiol* **78**(3):541-4.

Havarstein, L. S., *et al*. (1995). An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A* **92**(24):11140-4.

Havarstein, L. S., *et al*. (2006). New insights into the pneumococcal fratricide: relationship to clumping and identification of a novel immunity factor. *Mol Microbiol* **59**(4):1297-307.

He, M., *et al*. (2010). Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A* **107**(16):7527-32.

He, Y., *et al*. (2008). The antisense transcriptomes of human cells. *Science* **322**(5909):1855-7.

Heidelberger, M. and O. T. Avery (1923). The soluble specific substance of pneumococcus. *J Exp Med* **38**(1):73-9.

Heidelberger, M. and W. Goebel (1927). The soluble specific substance of pneumococcus. *J Biol Chem* **74**:613-618.

Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol* **36**(4):396-405.

Hendrickson, D. J., *et al*. (2008). Five-fold increase in pediatric parapneumonic empyema since introduction of pneumococcal conjugate vaccine. *Pediatr Infect Dis J* **27**(11):1030-2.

Henriques, B., *et al*. (2000). Molecular epidemiology of *Streptococcus pneumoniae* causing invasive disease in 5 countries. *J Infect Dis* **182**(3):833-9.

Henriques Normark, B., *et al*. (2001). Clinical isolates of *Streptococcus pneumoniae* that exhibit tolerance of vancomycin. *Clin Infect Dis* **32**(4):552-8.

Hermans, P. W., *et al*. (1997). Penicillin-resistant *Streptococcus pneumoniae* in the Netherlands: results of a 1-year molecular epidemiologic survey. *J Infect Dis* **175**(6):1413-22.

Hernandez, D., *et al*. (2008). *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* **18**(5):802-9.

Hicks, L. A., *et al*. (2007). Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998-2004. *J Infect Dis* **196**(9):1346-54.

Hill, P. C., *et al*. (2008). Nasopharyngeal carriage of *Streptococcus pneumoniae* in Gambian infants: a longitudinal study. *Clin Infect Dis* **46**(6):807-14.

Hill, W. and A. Robertson (1966). The effect of linkage on limits to artificial selection. *Genetical Res* **8**:269-294.

Hiller, N. L., *et al*. (2007). Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol* **189**(22):8186-95.

Hinds, J., *et al*. (2002a). Glass slide microarrays for bacterial genomes. *Methods Microbiol* **33**:83-99.

Hinds, J., *et al*. (2002b). Microarray design for bacterial genomes. *Methods Microbiol* **33**:67-82.

Hirst, R. A., *et al*. (2004). The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clin Exp Immunol* **138**(2):195-201.

Hoelzer, M. A. and R. E. Michod (1991). DNA repair and the evolution of transformation in Bacillus subtilis. III. Sex with damaged DNA. *Genetics* **128**(2):215-23.

Hofacker, I. L. (2009). RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* **Chapter 12**:Unit12 2.

Hoge, C. W., *et al*. (1994). An epidemic of pneumococcal disease in an overcrowded, inadequately ventilated jail. *N Engl J Med* **331**(10):643-8.

Hogg, J. S., *et al*. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* **8**(6):R103.

Holden, M. T., *et al*. (2009a). Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. *PLoS One* **4**(7):e6072.

Holden, M. T., *et al*. (2009b). Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog* **5**(3):e1000346.

Holt, K. E., *et al*. (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* **40**(8):987-93.

Hoskins, J., *et al*. (2001). Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J Bacteriol* **183**(19):5709-17.

Howard, L. V. and H. Gooder (1974). Specificity of the autolysin of *Streptococcus* (*Diplococcus*) *pneumoniae*. *J Bacteriol* **117**(2):796-804.

Howe, J. G. and T. S. Wilson (1972). Co-trimoxazole-resistant pneumococci. *Lancet* **2**(7769):184-5.

Hsieh, Y. C., *et al*. (2006). Serotype competence and penicillin resistance in *Streptococcus pneumoniae*. *Emerg Infect Dis* **12**(11):1709-14.

Hsu, K., *et al*. (2005). Population-based surveillance for childhood invasive pneumococcal disease in the era of conjugate vaccine. *Pediatr Infect Dis J* **24**(1):17-23.

Huang, S. S., *et al*. (2004). Community-level predictors of pneumococcal carriage and resistance in young children. *Am J Epidemiol* **159**(7):645-54.

Huang, S. S., *et al*. (2005). Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004. *Pediatrics* **116**(3):e408-13.

Huebner, R. E., *et al*. (2000). Lack of utility of serotyping multiple colonies for detection of simultaneous nasopharyngeal carriage of different pneumococcal serotypes. *Pediatr Infect Dis J* **19**(10):1017-20.

Hughes, D. T. (1969). Single-blind comparative trial of trimethoprim-sulphamethoxazole and ampicillin in the treatment of exacerbations of chronic bronchitis. *Br Med J* **4**(5681):470-3.

Hulton, C. S., *et al*. (1991). ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. *Mol Microbiol* **5**(4):825-34.

Humbert, O., *et al*. (1995). Homeologous recombination and mismatch repair during transformation in *Streptococcus pneumoniae*: saturation of the Hex mismatch repair system. *Proc Natl Acad Sci U S A* **92**(20):9052-6.

Husmeier, D. (2005). Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics* **21 Suppl 2**:ii166-72.

Hyams, C., *et al*. (2010a). The *Streptococcus pneumoniae* capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. *Infect Immun* **78**(2):704-15.

Hyams, C., *et al*. (2010b). *Streptococcus pneumoniae* resistance to complement-mediated immunity is dependent on the capsular serotype. *Infect Immun* **78**(2):716-25.

Inostroza, J., *et al*. (2001). Influence of patient age on *Streptococcus pneumoniae* serotypes causing invasive disease. *Clin Diagn Lab Immunol* **8**(3):556-9.

Inverarity, D. (2009). Genomic diversity in naturally transformable *Streptococcus pneumoniae*. Division of Infection and Immunity. Glasgow, University of Glasgow. **PhD**.

Isozumi, R., *et al*. (2008). Molecular characteristics of serotype 3 *Streptococcus pneumoniae* isolates among community-acquired pneumonia patients in Japan. *J Infect Chemother* **14**(3):258-61.

Iyer, V. N. and A. W. Ravin (1962). Integration and expression of different lengths of DNA during the transformation of pneumococcus to erythromycin resistance. *Genetics* **47**(10):1355-68.

Jabes, D., *et al*. (1989). Penicillin-binding protein families: evidence for the clonal nature of penicillin resistance in clinical isolates of pneumococci. *J Infect Dis* **159**(1):16-25.

Jacobs, M. R., *et al*. (1978). Emergence of multiply resistant pneumococci. *N Engl J Med* **299**(14):735-40.

Jacobs, N. M. (1991). Pneumococcal osteomyelitis and arthritis in children. A hospital series and literature review. *Am J Dis Child* **145**(1):70-4.

Jacoby, P., *et al*. (2007). Modelling the co-occurrence of *Streptococcus pneumoniae* with other bacterial and viral pathogens in the upper respiratory tract. *Vaccine* **25**(13):2458-64.

Janoff, E. N., *et al*. (1993). *Streptococcus pneumoniae* colonization, bacteremia, and immune response among persons with human immunodeficiency virus infection. *J Infect Dis* **167**(1):49-56.

Janoir, C., *et al*. (1996). High-level fluoroquinolone resistance in *Streptococcus pneumoniae* requires mutations in *parC* and *gyrA*. *Antimicrob Agents Chemother* **40**(12):2760-4.

Janulczyk, R., *et al*. (2000). Hic, a novel surface protein of *Streptococcus pneumoniae* that interferes with complement function. *J Biol Chem* **275**(47):37257-63.

Jennings, H. J., *et al*. (1980). Structure of the complex polysaccharide C-substance from *Streptococcus pneumoniae* type 1. *Biochemistry* **19**(20):4712-9.

Johnsborg, O., *et al*. (2007). Natural genetic transformation: prevalence, mechanisms and function. *Res Microbiol* **158**(10):767-78.

Johnson, A. P., *et al*. (2007). Morbidity and mortality of pneumococcal meningitis and serotypes of causative strains prior to introduction of the 7-valent conjugant pneumococcal vaccine in England. *J Infect* **55**(5):394-9.

Juhn, Y. J., *et al*. (2008). Increased risk of serious pneumococcal disease in patients with asthma. *J Allergy Clin Immunol* **122**(4):719-23.

Kadioglu, A., *et al*. (2008). The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat Rev Microbiol* **6**(4):288-301.

Kalin, M., *et al*. (2000). Prospective study of prognostic factors in community-acquired bacteremic pneumococcal disease in 5 countries. *J Infect Dis* **182**(3):840-7.

Kaltoft, M. S., *et al*. (2008). An easy method for detection of nasopharyngeal carriage of multiple *Streptococcus pneumoniae* serotypes. *J Microbiol Methods* **75**(3):540-4.

Kane, M. D., *et al*. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28**(22):4552-7.

Kaplan, S. L., *et al*. (2004). Decrease of invasive pneumococcal infections in children among 8 children's hospitals in the United States after the introduction of the 7-valent pneumococcal conjugate vaccine. *Pediatrics* **113**(3 Pt 1):443-9.

Karlsson, C., *et al*. (1999). The pneumococcal common antigen C-polysaccharide occurs in different forms. Mono-substituted or di-substituted with phosphocholine. *Eur J Biochem* **265**(3):1091-7.

Kashket, E. (1987). Bioenergetics of lactic acid bacteria: cytoplasmic pH and osmotolerance. *FEMS Microbiol Rev* **46**:233-244.

Kastenbauer, S. and H. W. Pfister (2003). Pneumococcal meningitis in adults: spectrum of complications and prognostic factors in a series of 87 cases. *Brain* **126**(Pt 5):1015-25.

Kawamura, Y., *et al*. (1995). Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus Streptococcus. *Int J Syst Bacteriol* **45**(2):406-8.

Keefer, C., *et al*. (1943). Penicillin in the treatment of infections. *JAMA* **122**(18):1217-1224.

Keith, E. R., *et al*. (2006). Characteristics of *Streptococcus pseudopneumoniae* isolated from purulent sputum samples. *J Clin Microbiol* **44**(3):923-7.

Kellner, J. D., *et al*. (1998). The use of *Streptococcus pneumoniae* nasopharyngeal isolates from healthy children to predict features of invasive disease. *Pediatr Infect Dis J* **17**(4):279-86.

Kelly, T., *et al*. (1994). Effect of genetic switching of capsular type on virulence of *Streptococcus pneumoniae*. *Infect Immun* **62**(5):1813-9.

Kemp, K. (1979). A review of selected federal vaccine and immunization policies, based on case studies of pneumococcal vaccine, Office of Technology Assessment.

Kidgell, C., *et al*. (2002). *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect Genet Evol* **2**(1):39-45.

Kilian, M., *et al*. (2008). Evolution of Streptococcus pneumoniae and its close commensal relatives. *PLoS One* **3**(7):e2683.

Kim, J. O. and J. N. Weiser (1998). Association of intrastrain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of *Streptococcus pneumoniae*. *J Infect Dis* **177**(2):368-77.

Kislak, J. W. (1967). Type 6 pneumococcus resistant to erythromycin and lincomycin. *N Engl J Med* **276**(15):852.

Kislak, J. W., *et al*. (1965). Susceptibility of pneumococci to nine antibiotics. *Am J Med Sci* **250**(3):261-8.

Klein, J. O. (1994). Otitis media. *Clin Infect Dis* **19**(5):823-33.

Klein, J. O. (2000). The burden of otitis media. *Vaccine* **19 Suppl 1**:S2-8.

Klugman, K. (1998). Pneumococcal molecular epidemiology network. *ASM News* **64**(7):371.

Klugman, K. P. (2002). The successful clone: the vector of dissemination of resistance in *Streptococcus pneumoniae*. *J Antimicrob Chemother* **50 Suppl S2**:1-5.

Klugman, K. P. and H. Koornhof (1988a). Bacteremic pneumonia caused by penicillin-resistant pneumococci. *N Engl J Med* **318**(2):123-4.

Klugman, K. P. and H. J. Koornhof (1988b). Drug resistance patterns and serogroups or serotypes of pneumococcal isolates from cerebrospinal fluid or blood, 1979-1986. *J Infect Dis* **158**(5):956-64.

Knutsen, E., *et al*. (2006). BOX elements modulate gene expression in *Streptococcus pneumoniae*: impact on the fine-tuning of competence development. *J Bacteriol* **188**(23):8307-12.

Koch, R. (1893). Über den augenblicklichen Stand der bakteriologischen Choleradiagnose. *Z Hyg Infekt* **14**:319-333.

Koedel, U., *et al*. (2002). Pathogenesis and pathophysiology of pneumococcal meningitis. *Lancet Infect Dis* **2**(12):721-36.

Koeuth, T., *et al*. (1995). Differential subsequence conservation of interspersed repetitive *Streptococcus pneumoniae* BOX elements in diverse bacteria. *Genome Res* **5**(4):408-18.

Koivula, I., *et al*. (1994). Risk factors for pneumonia in the elderly. *Am J Med* **96**(4):313-20.

Kreutzer, D. A. and J. M. Essigmann (1998). Oxidized, deaminated cytosines are a source of C --> T transitions in vivo. *Proc Natl Acad Sci U S A* **95**(7):3578-82.

Kuhn, H. and M. D. Frank-Kamenetskii (2005). Template-independent ligation of single-stranded DNA by T4 DNA ligase. *FEBS J* **272**(23):5991-6000.

Kulldorf, M. (1997). A spatial scan statistic. *Comm Stat Theory Meth* **26**(6):1481-1496.

Kyaw, M. H., *et al*. (2006). Effect of introduction of the pneumococcal conjugate vaccine on drug-resistant *Streptococcus pneumoniae*. *N Engl J Med* **354**(14):1455-63.

Kyaw, M. H., *et al*. (2005). The influence of chronic illnesses on the incidence of invasive pneumococcal disease in adults. *J Infect Dis* **192**(3):377-86.

Lacks, S. (1966). Integration efficiency and genetic recombination in pneumococcal transformation. *Genetics* **53**(1):207-35.

Lacks, S. (1970). Mutants of *Diplococcus pneumoniae* that lack deoxyribonucleases and other activities possibly pertinent to genetic transformation. *J Bacteriol* **101**(2):373-83.

Lacks, S. and M. Neuberger (1975). Membrane location of a deoxyribonuclease implicated in the genetic transformation of *Diplococcus pneumoniae*. *J Bacteriol* **124**(3):1321-9.

Lacks, S. A., *et al*. (1982). Identification of base mismatches recognized by the heteroduplex-DNA-repair system of *Streptococcus pneumoniae*. *Cell* **31**(2 Pt 1):327-36.

Laible, G. and R. Hakenbeck (1991). Five independent combinations of mutations can result in low-affinity penicillin-binding protein 2x of *Streptococcus pneumoniae*. *J Bacteriol* **173**(21):6986-90.

Laible, G., *et al*. (1989). Nucleotide sequences of the *pbpX* genes encoding the penicillin-binding proteins 2x from *Streptococcus pneumoniae* R6 and a cefotaxime-resistant mutant, C506. *Mol Microbiol* **3**(10):1337-48.

Lancereaux, E. and J. Besançon (1886). Étude sur quelques cas de pneumonie observés à l'hôpital de la Pitié au printemps de l'année 1886. *Arch gen de med* **7**(18):257.

Lanie, J. A., *et al*. (2007). Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. *J Bacteriol* **189**(1):38-51.

Latorre, C., *et al*. (1985). Antibiotic resistance and serotypes of 100 *Streptococcus pneumoniae* strains isolated in a children's hospital in Barcelona, Spain. *Antimicrob Agents Chemother* **28**(2):357-9.

Lau, G. W., *et al*. (2001). A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence. *Mol Microbiol* **40**(3):555-71.

Lawrence, E. M., *et al*. (1983). Pneumococcal vaccine in normal children. Primary and secondary vaccination. *Am J Dis Child* **137**(9):846-50.

Le Hello, S., *et al*. (2010). Invasive serotype 1 *Streptococcus pneumoniae* outbreaks in the South Pacific from 2000 to 2007. *J Clin Microbiol* **48**(8):2968-71.

Lefevre, J. C., *et al*. (1989). Conversion of deletions during recombination in pneumococcal transformation. *Genetics* **123**(3):455-64.

Leimkugel, J., *et al*. (2005). An outbreak of serotype 1 *Streptococcus pneumoniae* meningitis in northern Ghana with features that are characteristic of *Neisseria meningitidis* meningitis epidemics. *J Infect Dis* **192**(2):192-9.

Lepoutre, A., *et al*. (2008). Impact of infant pneumococcal vaccination on invasive pneumococcal diseases in France, 2001-2006. *Euro Surveill* **13**(35).

Lexau, C. A., *et al*. (2005). Changing epidemiology of invasive pneumococcal disease among older adults in the era of pediatric pneumococcal conjugate vaccine. *JAMA* **294**(16):2043-51.

Leyden, H. v. (1882). Über infectiöse Pneumonie. *Deutsch Med Wochenschr* **9**:52-54.

Li, H. and R. Durbin (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5):589-95.

Li, H., *et al*. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16):2078-9.

Li, L., *et al*. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**(9):2178-89.

Lin, P. L., *et al*. (2003). Incidence of invasive pneumococcal disease in children 3 to 36 months of age at a tertiary care pediatric center 2 years after licensure of the pneumococcal conjugate vaccine. *Pediatrics* **111**(4 Pt 1):896-9.

Linares, J., *et al*. (1983). Antibiotic resistance and serotypes of *Streptococcus pneumoniae* from patients with community-acquired pneumococcal disease. *Antimicrob Agents Chemother* **23**(4):545-7.

Lipsitch, M. (1997). Vaccination against colonizing bacteria with multiple serotypes. *Proc Natl Acad Sci U S A* **94**(12):6571-6.

Lipsitch, M. (1999). Bacterial vaccines and serotype replacement: lessons from *Haemophilus influenzae* and prospects for *Streptococcus pneumoniae*. *Emerg Infect Dis* **5**(3):336-45.

Lipsky, B. A., *et al*. (1986). Risk factors for acquiring pneumococcal infections. *Arch Intern Med* **146**(11):2179-85.

Lister, R., *et al*. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**(3):523-36.

Llull, D., *et al*. (1999). A single gene (*tts*) located outside the cap locus directs the formation of *Streptococcus pneumoniae* type 37 capsular polysaccharide. Type 37 pneumococci are natural, genetically binary strains. *J Exp Med* **190**(2):241-51.

Loeffler, J. M. and V. A. Fischetti (2006). Lysogeny of *Streptococcus pneumoniae* with MM1 phage: improved adherence and other phenotypic changes. *Infect Immun* **74**(8):4486-95.

Long, P. and E. Bliss (1937). The use of para amino benzene sulphonamide (Sulphanilamide) or its derivatives in the treatment of infections due To beta hemolytic streptococci, pneumococci and meningococci. *South Med J* **30**(5):479-487.

Lopez, R. and E. Garcia (2004). Recent trends on the molecular biology of pneumococcal capsules, lytic enzymes, and bacteriophage. *FEMS Microbiol Rev* **28**(5):553-80.

Lopez, R., *et al*. (1982). Choline-containing bacteriophage receptors in *Streptococcus pneumoniae*. *J Bacteriol* **151**(3):1581-90.

Lopez, R., *et al*. (1992). Structural analysis and biological significance of the cell wall lytic enzymes of *Streptococcus pneumoniae* and its bacteriophage. *FEMS Microbiol Lett* **79**(1-3):439-47.

Lowell, F., *et al*. (1940). Observations on the susceptibility of pneumococci to sulfapyridine, sulfathiazole and sulfamethylthiazole. *Ann Intern Med* **14**(6):1001-1023.

Lowenburg, H. (1929). Pneumococcic empyema. *JAMA* **93**(2):106-107.

Lujan, M., *et al*. (2004). Prospective observational study of bacteremic pneumococcal pneumonia: Effect of discordant therapy on mortality. *Crit Care Med* **32**(3):625-31.

Lund, E. (1960). Laboratory diagnosis of pneumococcus infections. *Bull Wld Hlth Org* **23**:5-13.

Luotonen, J. (1982). *Streptococcus pneumoniae* and *Haemophilus influenzae* in nasal cultures during acute otitis media. *Acta Otolaryngol* **93**(3-4):295-9.

Lux, T., *et al*. (2007). Diversity of bacteriocins and activity spectrum in *Streptococcus pneumoniae*. *J Bacteriol* **189**(21):7741-51.

Lysenko, E. S., *et al*. (2005). The role of innate immune responses in the outcome of interspecies competition for colonization of mucosal surfaces. *PLoS Pathog* **1**(1):e1.

MacFadyen, L. P., *et al*. (2001). Competence development by *Haemophilus influenzae* is regulated by the availability of nucleic acid precursors. *Mol Microbiol* **40**(3):700-7.

Macfarlane, J. (1994). An overview of community acquired pneumonia with lessons learned from the British Thoracic Society Study. *Semin Respir Infect* **9**(3):153-65.

Macleod, C. M., *et al*. (1945). Prevention of pneumococcal pneumonia by immunization with specific capsular polysaccharides. *J Exp Med* **82**(6):445-65.

Madhi, S. A., *et al*. (2007). Long-term effect of pneumococcal conjugate vaccine on nasopharyngeal colonization by *Streptococcus pneumoniae*--and associated interactions with *Staphylococcus aureus* and *Haemophilus influenzae* colonization--in HIV-Infected and HIV-uninfected children. *J Infect Dis* **196**(11):1662-6.

Mahjoub-Messai, F., *et al*. (2009). Population snapshot of *Streptococcus pneumoniae* serotype 19A isolates before and after introduction of seven-valent pneumococcal vaccination for French children. *J Clin Microbiol* **47**(3):837-40.

Maiden, M. C., *et al*. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**(6):3140-5.

Majewski, J. and F. M. Cohan (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* **148**(1):13-8.

Majewski, J., *et al*. (2000). Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J Bacteriol* **182**(4):1016-23.

Makela, P. H., *et al*. (1981). A study of the pneumococcal vaccine in prevention of clinically acute atttacks of recurrent otitis media. *Rev Infect Dis* **3 Suppl**:S124-32.

Maki, D. G., *et al*. (1980). Penicillin susceptibility of *Streptococcus pneumoniae* in 1978. Screening for resistance by disk testing. *Am J Clin Pathol* **73**(2):177-82.

Mangalam, H. (2002). The Bio* toolkits--a brief overview. *Brief Bioinform* **3**(3):296-302.

Mankovich, J. A., *et al*. (1989). Nucleotide sequence of the *Salmonella typhimurium mutL* gene required for mismatch repair: homology of MutL to HexB of *Streptococcus pneumoniae* and to PMS1 of the yeast *Saccharomyces cerevisiae*. *J Bacteriol* **171**(10):5325-31.

Margulies, M., *et al*. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057):376-80.

Marioni, J. C., *et al*. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**(9):1509-17.

Markiewicz, Z. and A. Tomasz (1989). Variation in penicillin-binding protein patterns of penicillin-resistant clinical isolates of pneumococci. *J Clin Microbiol* **27**(3):405-10.

Marrer, E., *et al*. (2006). Involvement of the putative ATP-dependent efflux proteins PatA and PatB in fluoroquinolone resistance of a multidrug-resistant mutant of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **50**(2):685-93.

Marrie, T. J. (1992). Bacteraemic pneumococcal pneumonia: a continuously evolving disease. *J Infect* **24**(3):247-55.

Martens, P., *et al*. (2004). Serotype-specific mortality from invasive *Streptococcus pneumoniae* disease revisited. *BMC Infect Dis* **4**:21.

Martin, A. C., *et al*. (1996). Analysis of the complete nucleotide sequence and functional organization of the genome of *Streptococcus pneumoniae* bacteriophage Cp-1. *J Virol* **70**(6):3678-87.

Martin, B., *et al*. (1992). A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. *Nucleic Acids Res* **20**(13):3479-83.

Martin, M., *et al*. (2003). An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*. *N Engl J Med* **348**(12):1112-21.

Marton, A., *et al*. (1991). Extremely high incidence of antibiotic resistance in clinical isolates of *Streptococcus pneumoniae* in Hungary. *J Infect Dis* **163**(3):542-8.

Matthews, L. W., *et al*. (1963). Studies on pulmonary secretions. I. The over-all chemical composition of pulmonary secretions from patients with cystic fibrosis, bronchiectasis, and laryngectomy. *Am Rev Respir Dis* **88**:199-204.

Maynard, C. (1915). Pneumonia inoculation experiment no. III. *Med J S Afr* **11**:36-38.

Maynard, G. (1913). An enquiry into the etiology, manifestations and prevention of pneumonia amongst natives on the Rand recruited from tropical areas. *Public South Afr Inst Med Res* **1**:1-101.

Maynard Smith, J. (1964). Group selection and kin selection. *Nature* **201**(4924):1145-1147.

Maynard Smith, J. (1978). *The evolution of sex*. Cambridge, Cambridge University Press.

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**(2):126-9.

Maynard Smith, J. (1999). The detection and measurement of recombination from sequence data. *Genetics* **153**(2):1021-7.

Maynard Smith, J. and N. H. Smith (1998). Detecting recombination from gene trees. *Mol Biol Evol* **15**(5):590-9.

Mazzone, M., *et al*. (2001). Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic *Neisseriae*. *Gene* **278**(1-2):211-22.

McClelland, M., *et al*. (1987). Restriction endonucleases for pulsed field mapping of bacterial genomes. *Nucleic Acids Res* **15**(15):5985-6005.

McCool, T. L., *et al*. (2002). The immune response to pneumococcal proteins during experimental human carriage. *J Exp Med* **195**(3):359-65.

McCullers, J. A., *et al*. (2000). Isolation and characterization of vancomycin-tolerant *Streptococcus pneumoniae* from the cerebrospinal fluid of a patient who developed recrudescent meningitis. *J Infect Dis* **181**(1):369-73.

McDonnell, M., *et al*. (1975). "Diplophage": a bacteriophage of *Diplococcus pneumoniae*. *Virology* **63**(2):577-82.

McDougal, L. K., *et al*. (1998). Detection of Tn*917*-like sequences within a Tn*916*-like conjugative transposon (Tn*3872*) in erythromycin-resistant isolates of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **42**(9):2312-8.

McEllistrem, M. C., *et al*. (2003). Epidemiology of acute otitis media caused by *Streptococcus pneumoniae* before and after licensure of the 7-valent pneumococcal protein conjugate vaccine. *J Infect Dis* **188**(11):1679-84.

McGee, L. and K. Klugman. (2011). "Pneumococcal Molecular Epidemiology Network." 2011, from http://www.sph.emory.edu/PMEN/.
McGee, L., *et al*. (2001). Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. *J Clin Microbiol* **39**(7):2565-71.
McGuire, G. and F. Wright (2000). TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* **16**(2):130-4.
McGuire, G., *et al*. (1997). A graphical method for detecting recombination in phylogenetic data sets. *Mol Biol Evol* **14**(11):1125-31.
McGuire, G., *et al*. (2000). A Bayesian model for detecting past recombination events in DNA multiple alignments. *J Comput Biol* **7**(1-2):159-70.
McKee, C. and C. Houck (1943). Induced resistance to penicillin of cultures of staphylococci, pneumococci and streptococci. *Proc Soc Exp Biol Med* **53**:33-34.
McNally, L. M., *et al*. (2006). Lack of association between the nasopharyngeal carriage of *Streptococcus pneumoniae* and *Staphylococcus aureus* in HIV-1-infected South African children. *J Infect Dis* **194**(3):385-90.
Mehtar, S., *et al*. (1990). Clinical evaluation of oral ciprofloxacin in serious infection: an open study. *Eur J Int Med* **1**(383-390):383.
Mejean, V. and J. P. Claverys (1984). Use of a cloned DNA fragment to analyze the fate of donor DNA in transformation of *Streptococcus pneumoniae*. *J Bacteriol* **158**(3):1175-8.
Mejean, V. and J. P. Claverys (1988). Polarity of DNA entry in transformation of *Streptococcus pneumoniae*. *Mol Gen Genet* **213**(2-3):444-8.
Mercat, A., *et al*. (1991). An outbreak of pneumococcal pneumonia in two men's shelters. *Chest* **99**(1):147-51.
Messina, A. F., *et al*. (2007). Impact of the pneumococcal conjugate vaccine on serotype distribution and antimicrobial resistance of invasive *Streptococcus pneumoniae* isolates in Dallas, TX, children from 1999 through 2005. *Pediatr Infect Dis J* **26**(6):461-7.
Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Res* **15**(12):1767-76.
Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet* **11**(1):31-46.
Michel, J. B., *et al*. (2010). Quantitative analysis of culture using millions of digitized books. *Science* **331**(6014):176-82.
Michod, R. E., *et al*. (2008). Adaptive value of sex in microbial pathogens. *Infect Genet Evol* **8**(3):267-85.
Milkman, R. and M. M. Bridges (1993). Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. *Genetics* **133**(3):455-68.
Millar, E. V., *et al*. (2006). Effect of community-wide conjugate pneumococcal vaccine use in infancy on nasopharyngeal carriage through 3 years of age: a cross-sectional study in a high-risk population. *Clin Infect Dis* **43**(1):8-15.
Mingoia, M., *et al*. (2007). Composite structure of *Streptococcus pneumoniae* containing the erythromycin efflux resistance gene mefI and the chloramphenicol resistance gene *catQ*. *Antimicrob Agents Chemother* **51**(11):3983-7.
Mitchell, T. J., *et al*. (1991). Complement activation and antibody binding by pneumolysin via a region of the toxin homologous to a human acute-phase protein. *Mol Microbiol* **5**(8):1883-8.

Moller, P. and H. Wallin (1998). Adduct formation, mutagenesis and nucleotide excision repair of DNA damage produced by reactive oxygen species and lipid peroxidation product. *Mutat Res* **410**(3):271-90.

Moore, H. and A. Chesney (1917). A study of ethyhydrocuprein (optochin) in the treatment of acute lobar pneumonia. *Arch Intern Med* **19**(4):611-682.

Moore, M. R., *et al*. (2008). Population snapshot of emergent *Streptococcus pneumoniae* serotype 19A in the United States, 2005. *J Infect Dis* **197**(7):1016-27.

Moore, M. R., *et al*. (2004). Impact of a conjugate vaccine on community-wide carriage of nonsusceptible *Streptococcus pneumoniae* in Alaska. *J Infect Dis* **190**(11):2031-8.

Morel, P., *et al*. (1993). Antipairing and strand transferase activities of *E. coli* helicase II (UvrD). *Nucleic Acids Res* **21**(14):3205-9.

Moreno, F., *et al*. (1995). The clinical and molecular epidemiology of bacteremias at a university hospital caused by pneumococci not susceptible to penicillin. *J Infect Dis* **172**(2):427-32.

Morgenroth, J. and M. Kaufmann (1912). Arzneifestigkeit bei Bakterien (Pneumokokken). *Z. Immunitaetsforsch* **15**:610-624.

Morrison, D. A. and W. R. Guild (1972). Transformation and deoxyribonucleic acid size: extent of degradation on entry varies with size of donor. *J Bacteriol* **112**(3):1157-68.

Morse, H. G. and L. S. Lerman (1969). A genetic analysis by transformation of a group of uracil-requiring mutants of *Diplococcus pneumoniae*. *Genetics* **61**(1):41-60.

Mortazavi, A., *et al*. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7):621-8.

Mortier-Barriere, I., *et al*. (2007). A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA. *Cell* **130**(5):824-36.

Moscoso, M., *et al*. (2010). Vancomycin tolerance in clinical and laboratory *Streptococcus pneumoniae* isolates depends on reduced enzyme activity of the major LytA autolysin or cooperation between CiaH histidine kinase and capsular polysaccharide. *Mol Microbiol*.

Mosser, J. L. and A. Tomasz (1970). Choline-containing teichoic acid as a structural component of pneumococcal cell wall and its role in sensitivity to lysis by an autolytic enzyme. *J Biol Chem* **245**(2):287-98.

Moxon, E. R. and K. A. Vaughn (1981). The type b capsular polysaccharide as a virulence determinant of *Haemophilus influenzae*: studies using clinical isolates and laboratory transformants. *J Infect Dis* **143**(4):517-24.

Muller, H. (1932). Some genetic aspects of sex. *Am Nat* **66**(703):118-138.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat Res* **106**:2-9.

Munoz, R., *et al*. (1991). Intercontinental spread of a multiresistant clone of serotype 23F *Streptococcus pneumoniae*. *J Infect Dis* **164**(2):302-6.

Munoz, R. and A. G. De La Campa (1996). ParC subunit of DNA topoisomerase IV of *Streptococcus pneumoniae* is a primary target of fluoroquinolones and cooperates with DNA gyrase A subunit in forming resistance phenotype. *Antimicrob Agents Chemother* **40**(10):2252-7.

Munoz, R., *et al*. (1992). Genetics of resistance to third-generation cephalosporins in clinical isolates of *Streptococcus pneumoniae*. *Mol Microbiol* **6**(17):2461-5.

Munoz, R., *et al*. (1999). Construction of a new *Streptococcus pneumoniae-Escherichia coli* shuttle vector based on the replicon of an indigenous pneumococcal cryptic plasmid. *Int Microbiol* **2**(1):23-8.

Munoz-Almagro, C., *et al*. (2008). Emergence of invasive pneumococcal disease caused by nonvaccine serotypes in the era of 7-valent conjugate vaccine. *Clin Infect Dis* **46**(2):174-82.

Munoz-Najar, U. and M. N. Vijayakumar (1999). An operon that confers UV resistance by evoking the SOS mutagenic response in streptococcal conjugative transposon Tn*5252*. *J Bacteriol* **181**(9):2782-8.

Murdoch, J. M., *et al*. (1964). Clinical trial of cephaloridine (ceporin), a new broad-spectrum antibiotic derived from cephalosporin C. *Br Med J* **2**(5419):1238-40.

Nagalakshmi, U., *et al*. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881):1344-9.

Nandoskar, M., *et al*. (1986). Inhibition of human monocyte respiratory burst, degranulation, phospholipid methylation and bactericidal activity by pneumolysin. *Immunology* **59**(4):515-20.

Netter (1886). De l'endocardite vegetante-ulcereuse d'origine pneumonique. *Arch de physiol norm et path* **8**(2):106-161.

Netter (1887). De la méningite due au pneumocoque (avec ou sans pneumonie). *Arch gen de med* **7**(19):257-77.

Neufeld, F. (1900). Über eine spezifische bakteriolytische Wirkung der Galle. *Z Hyg Infekt* **34**:454-464.

Neufeld, F. (1902). Über die Agglutination der Pneumokokken und über die Theorieen der Agglutination *Z Hyg Infekt* **40**:54-72.

Neufeld, F. and L. Händel (1910). Weitere Untersuchungen über Pneumokokken-Heilsera. III. Mitteilung. Über Vorkommen and Bedeutung atypischer Varietäten des Pneumokokkes. *Arbeiten aus dem kaiserlichen Gesundheitsamte* **34**:293-304.

Neuhaus, F. C. and J. Baddiley (2003). A continuum of anionic charge: structures and functions of D-alanyl-teichoic acids in gram-positive bacteria. *Microbiol Mol Biol Rev* **67**(4):686-723.

Niederman, M. S., *et al*. (2001). Guidelines for the management of adults with community-acquired pneumonia. Diagnosis, assessment of severity, antimicrobial therapy, and prevention. *Am J Respir Crit Care Med* **163**(7):1730-54.

Ning, Z., *et al*. (2001). SSAHA: a fast search method for large DNA databases. *Genome Res* **11**(10):1725-9.

Nizet, V. (2002). Streptococcal beta-hemolysins: genetics and role in disease pathogenesis. *Trends Microbiol* **10**(12):575-80.

Novak, R., *et al*. (1999). Emergence of vancomycin tolerance in *Streptococcus pneumoniae*. *Nature* **399**(6736):590-3.

Nuorti, J. P., *et al*. (1998). An outbreak of multidrug-resistant pneumococcal pneumonia and bacteremia among unvaccinated nursing home residents. *N Engl J Med* **338**(26):1861-8.

Nuorti, J. P., *et al*. (2000a). Cigarette smoking and invasive pneumococcal disease. Active Bacterial Core Surveillance Team. *N Engl J Med* **342**(10):681-9.

Nuorti, J. P., *et al*. (2000b). Epidemiologic relation between HIV and invasive pneumococcal disease in San Francisco County, California. *Ann Intern Med* **132**(3):182-90.

O'Brien, K. L., *et al*. (2009). Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet* **374**(9693):893-902.

Obert, C., *et al*. (2006). Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect Immun* **74**(8):4766-77.

Obregon, V., *et al*. (2003). Genome organization and molecular analysis of the temperate bacteriophage MM1 of *Streptococcus pneumoniae*. *J Bacteriol* **185**(7):2362-8.

Obregon, V., *et al*. (2002). Molecular peculiarities of the *lytA* gene isolated from clinical pneumococcal strains that are bile insoluble. *J Clin Microbiol* **40**(7):2545-54.

Ochman, H., *et al*. (1999). Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* **96**(22):12638-43.

Ogata, H., *et al*. (2002). Protein coding palindromes are a unique but recurrent feature in *Rickettsia*. *Genome Res* **12**(5):808-16.

Ogata, H., *et al*. (2000). Selfish DNA in protein-coding genes of Rickettsia. *Science* **290**(5490):347-50.

Oggioni, M. R. and J. P. Claverys (1999). Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology* **145 (Pt 10)**:2647-53.

Oggioni, M. R., *et al*. (1999). Characterization of cryptic plasmids pDP1 and pSMB1 of *Streptococcus pneumoniae*. *Plasmid* **41**(1):70-2.

Oggioni, M. R., *et al*. (2003). Pneumococcal zinc metalloproteinase ZmpC cleaves human matrix metalloproteinase 9 and is a virulence factor in experimental pneumonia. *Mol Microbiol* **49**(3):795-805.

Ojo, K. K., *et al*. (2006). The presence of a conjugative Gram-positive Tn*2009* in Gram-negative commensal bacteria. *J Antimicrob Chemother* **57**(6):1065-9.

Oliver, H. F., *et al*. (2009). Deep RNA sequencing of *L. monocytogenes* reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics* **10**:641.

Orihuela, C. J., *et al*. (2004). Microarray analysis of pneumococcal gene expression during invasive disease. *Infect Immun* **72**(10):5582-96.

Ortiz, P. J. (1970). Dihydrofolate and dihydropteroate synthesis by partially purified enzynes from wild-type and sulfonamide-resistant pneumonococcus. *Biochemistry* **9**(2):355-61.

Otto, T. D., *et al*. (2010). Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**(14):1704-7.

Pai, R., *et al*. (2006). Sequential multiplex PCR approach for determining capsular serotypes of *Streptococcus pneumoniae* isolates. *J Clin Microbiol* **44**(1):124-31.

Pai, R., *et al*. (2005). Postvaccine genetic structure of *Streptococcus pneumoniae* serotype 19A from children in the United States. *J Infect Dis* **192**(11):1988-95.

Pallares, R., *et al*. (1995). Resistance to penicillin and cephalosporin and mortality from severe pneumococcal pneumonia in Barcelona, Spain. *N Engl J Med* **333**(8):474-80.

Pankuch, G. A., *et al*. (1995). Activity of CP99,219 compared with DU-6859a, ciprofloxacin, ofloxacin, levofloxacin, lomefloxacin, tosufloxacin,

sparfloxacin and grepafloxacin against penicillin-susceptible and -resistant pneumococci. *J Antimicrob Chemother* **35**(1):230-2.

Park, I. H., *et al*. (2007). Genetic basis for the new pneumococcal serotype, 6C. *Infect Immun* **75**(9):4482-9.

Park, S. Y., *et al*. (2008). Impact of conjugate vaccine on transmission of antimicrobial-resistant *Streptococcus pneumoniae* among Alaskan children. *Pediatr Infect Dis J* **27**(4):335-40.

Parkhill, J. (2002). Annotation of microbial genomes. *Methods in Microbiology* **33**:3-26.

Parkhill, J., *et al*. (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**(6777):502-6.

Parry, C. M., *et al*. (2000). Nasal carriage in Vietnamese children of *Streptococcus pneumoniae* resistant to multiple antimicrobial agents. *Antimicrob Agents Chemother* **44**(3):484-8.

Passalacqua, K. D., *et al*. (2009). Structure and complexity of a bacterial transcriptome. *J Bacteriol* **191**(10):3203-11.

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

Pasteur, L. (1881). Sur une maladie nouvelle, provoquée par la salive d'un enfant mort de la rage. *C R Acad Sci (Paris)* **92**:159-165.

Pastor, P., *et al*. (1998). Invasive pneumococcal disease in Dallas County, Texas: results from population-based surveillance in 1995. *Clin Infect Dis* **26**(3):590-5.

Paterson, G. K., *et al*. (2008). PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol Lett* **285**(2):170-6.

Paulsen, I. T., *et al*. (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J Mol Biol* **301**(1):75-100.

Pearce, B. J., *et al*. (2002). Construction of new unencapsulated (rough) strains of *Streptococcus pneumoniae*. *Res Microbiol* **153**(4):243-7.

Peltola, H. (2000). Worldwide *Haemophilus influenzae* type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates. *Clin Microbiol Rev* **13**(2):302-17.

Peltola, V. T. and J. A. McCullers (2004). Respiratory viruses predisposing to bacterial infections: role of neuraminidase. *Pediatr Infect Dis J* **23**(1 Suppl):S87-97.

Pelton, S. I., *et al*. (2007). Emergence of 19A as virulent and multidrug resistant Pneumococcus in Massachusetts following universal immunization of infants with pneumococcal conjugate vaccine. *Pediatr Infect Dis J* **26**(6):468-72.

Perez-Trallero, E., *et al*. (1990). Therapeutic failure and selection of resistance to quinolones in a case of pneumococcal pneumonia treated with ciprofloxacin. *Eur J Clin Microbiol Infect Dis* **9**(12):905-6.

Perez-Trallero, E., *et al*. (2003). Fluoroquinolone and macrolide treatment failure in pneumococcal pneumonia and selection of multidrug-resistant isolates. *Emerg Infect Dis* **9**(9):1159-62.

Perichon, B., *et al*. (1997). Characterization of a mutation in the *parE* gene that confers fluoroquinolone resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **41**(5):1166-7.

Pericone, C. D., *et al*. (2000). Inhibitory and bactericidal effects of hydrogen peroxide production by *Streptococcus pneumoniae* on other inhabitants of the upper respiratory tract. *Infect Immun* **68**(7):3990-7.

Perkins, T. T., *et al*. (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet* **5**(7):e1000569.

Pestova, E. V., *et al*. (1996). Regulation of competence for genetic transformation in *Streptococcus pneumoniae* by an auto-induced peptide pheromone and a two-component regulatory system. *Mol Microbiol* **21**(4):853-62.

Pestova, E. V. and D. A. Morrison (1998). Isolation and characterization of three *Streptococcus pneumoniae* transformation-specific loci by use of a *lacZ* reporter insertion vector. *J Bacteriol* **180**(10):2701-10.

Peterson, S., *et al*. (2000). Gene expression analysis of the *Streptococcus pneumoniae* competence regulons by use of DNA microarrays. *J Bacteriol* **182**(21):6192-202.

Peterson, S. N., *et al*. (2004). Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays. *Mol Microbiol* **51**(4):1051-70.

Petranovic, M., *et al*. (2001). Genetic evidence that the elevated levels of *Escherichia coli* helicase II antagonize recombinational DNA repair. *Biochimie* **83**(11-12):1041-7.

Petrosillo, N., *et al*. (2002). Prevalence, determinants, and molecular epidemiology of *Streptococcus pneumoniae* isolates colonizing the nasopharynx of healthy children in Rome. *Eur J Clin Microbiol Infect Dis* **21**(3):181-8.

Pikis, A., *et al*. (2001). Optochin resistance in *Streptococcus pneumoniae*: mechanism, significance, and clinical implications. *J Infect Dis* **184**(5):582-90.

Pikis, A., *et al*. (1998). A conservative amino acid mutation in the chromosome-encoded dihydrofolate reductase confers trimethoprim resistance in *Streptococcus pneumoniae*. *J Infect Dis* **178**(3):700-6.

Pletz, M. W., *et al*. (2004). Levofloxacin-resistant invasive *Streptococcus pneumoniae* in the United States: evidence for clonal spread and the impact of conjugate pneumococcal vaccine. *Antimicrob Agents Chemother* **48**(9):3491-7.

Poehling, K. A., *et al*. (2006). Invasive pneumococcal disease among infants before and after introduction of pneumococcal conjugate vaccine. *JAMA* **295**(14):1668-74.

Polack, F. P., *et al*. (2000). Colonization by *Streptococcus pneumoniae* in human immunodeficiency virus-infected children. *Pediatr Infect Dis J* **19**(7):608-12.

Polissi, A., *et al*. (1998). Large-scale identification of virulence genes from *Streptococcus pneumoniae*. *Infect Immun* **66**(12):5620-9.

Ponstingl, H. (2011). "SMALT." from www.sanger.ac.uk/resources/software/smalt/.

Poolman, B. (1993). Energy transduction in lactic acid bacteria. *FEMS Microbiol Rev* **12**(1-3):125-47.

Posada, D., *et al*. (2002). Recombination in evolutionary genomics. *Annu Rev Genet* **36**:75-97.

Powars, D., *et al*. (1981). Pneumococcal septicemia in children with sickle cell anemia. Changing trend of survival. *JAMA* **245**(18):1839-42.

Powel, K. (2004). Changing interest among physicians toward pneumococcal vaccination throughout the twentieth century. *J Hist Med Allied Sci* **59**(4):555-587.

Poyart-Salmeron, C., *et al*. (1991). Nucleotide sequences specific for Tn*1545*-like conjugative transposons in pneumococci and staphylococci resistant to tetracycline. *Antimicrob Agents Chemother* **35**(8):1657-60.

Pozzi, G., *et al*. (1996). Competence for genetic transformation in encapsulated strains of *Streptococcus pneumoniae*: two allelic variants of the peptide pheromone. *J Bacteriol* **178**(20):6087-90.

Prats, H., *et al*. (1985). The *hexB* mismatch repair gene of *Streptococcus pneumoniae*: characterisation, cloning and identification of the product. *Mol Gen Genet* **200**(3):482-9.

Price, A. L., *et al*. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**:i351-8.

Priebe, S. D., *et al*. (1988). Nucleotide sequence of the *hexA* gene for DNA mismatch repair in *Streptococcus pneumoniae* and homology of *hexA* to *mutS* of *Escherichia coli* and *Salmonella typhimurium*. *J Bacteriol* **170**(1):190-6.

Principi, N., *et al*. (1999). Risk factors for carriage of respiratory pathogens in the nasopharynx of healthy children. Ascanius Project Collaborative Group. *Pediatr Infect Dis J* **18**(6):517-23.

Provvedi, R., *et al*. (1996). Characterization of conjugative transposon Tn*5251* of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **135**(2-3):231-6.

Prudhomme, M., *et al*. (2006). Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*. *Science* **313**(5783):89-92.

Prudhomme, M., *et al*. (1989). Nucleotide sequence of the *Streptococcus pneumoniae hexB* mismatch repair gene: homology of HexB to MutL of *Salmonella typhimurium* and to PMS1 of *Saccharomyces cerevisiae*. *J Bacteriol* **171**(10):5332-8.

Quick, R. E., *et al*. (1993). Underutilization of pneumococcal vaccine in nursing home in Washington State: report of a serotype-specific outbreak and a survey. *Am J Med* **94**(2):149-52.

Quin, L. R., *et al*. (2005). *In vivo* binding of complement regulator factor H by *Streptococcus pneumoniae*. *J Infect Dis* **192**(11):1996-2003.

R Development Core Team (2011). *R: a language and environment for statistical computing*. Vienna, R Foundation for Statistical Computing.

Radetsky, M. S., *et al*. (1981). Multiply resistant pneumococcus causing meningitis: its epidemiology within a day-care centre. *Lancet* **2**(8250):771-3.

Ramirez, M., *et al*. (1999). A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*. *J Bacteriol* **181**(12):3618-25.

Rane, L. and Y. Subbarow (1940). Nutritional Requirements of the Pneumococcus: I. Growth Factors for Types I, II, V, VII, VIII. *J Bacteriol* **40**(5):695-704.

Rauhut, R. and G. Klug (1999). mRNA degradation in bacteria. *FEMS Microbiol Rev* **23**(3):353-70.

Ravitch, M. M. and R. Fein (1961). The changing picture of pneumonia and empyema in infants and children. A review of the experience at the Harriet Lane Home from 1934 through 1958. *JAMA* **175**:1039-44.

Raymond, J., *et al*. (2000). Sequential colonization by *Streptococcus pneumoniae* of healthy children living in an orphanage. *J Infect Dis* **181**(6):1983-8.

Reber, W. (1917). Some phases of modern ocular therapeutics. *Br J Ophthalmol* **1**(5):294-309.

Redfield, R. J. (1988). Evolution of bacterial transformation: is sex with dead cells ever better than no sex at all? *Genetics* **119**(1):213-21.

Redfield, R. J. (1993a). Evolution of natural transformation: testing the DNA repair hypothesis in *Bacillus subtilis* and *Haemophilus influenzae*. *Genetics* **133**(4):755-61.

Redfield, R. J. (1993b). Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. *J Hered* **84**(5):400-4.

Regev-Yochay, G., *et al*. (2004a). Association between carriage of *Streptococcus pneumoniae* and *Staphylococcus aureus* in Children. *JAMA* **292**(6):716-20.

Regev-Yochay, G., *et al*. (2009). The pneumococcal pilus predicts the absence of *Staphylococcus aureus* co-colonization in pneumococcal carriers. *Clin Infect Dis* **48**(6):760-3.

Regev-Yochay, G., *et al*. (2004b). Nasopharyngeal carriage of *Streptococcus pneumoniae* by adults and children in community and family settings. *Clin Infect Dis* **38**(5):632-9.

Regev-Yochay, G., *et al*. (2006). Interference between *Streptococcus pneumoniae* and *Staphylococcus aureus*: *In vitro* hydrogen peroxide-mediated killing by *Streptococcus pneumoniae*. *J Bacteriol* **188**(13):4996-5001.

Rehrauer, W. M., *et al*. (1998). Modulation of RecA nucleoprotein function by the mutagenic UmuD'C protein complex. *J Biol Chem* **273**(49):32384-7.

Reichmann, P. and R. Hakenbeck (2000). Allelic variation in a peptide-inducible two-component system of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **190**(2):231-6.

Reinert, R. R., *et al*. (2005a). Antimicrobial susceptibility of *Streptococcus pneumoniae* in eight European countries from 2001 to 2003. *Antimicrob Agents Chemother* **49**(7):2903-13.

Reinert, R. R., *et al*. (2005b). Molecular epidemiology of macrolide-resistant *Streptococcus pneumoniae* isolates in Europe. *J Clin Microbiol* **43**(3):1294-300.

Reis, J. N., *et al*. (2008). Transmission of *Streptococcus pneumoniae* in an urban slum community. *J Infect* **57**(3):204-13.

Rennels, M. B., *et al*. (1998). Safety and immunogenicity of heptavalent pneumococcal vaccine conjugated to CRM197 in United States infants. *Pediatrics* **101**(4 Pt 1):604-11.

Rice, P., *et al*. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**(6):276-7.

Ridda, I., *et al*. (2010). Lack of pneumococcal carriage in the hospitalised elderly. *Vaccine* **28**(23):3902-4.

Riley, H. D., Jr. (1950). Pneumococcal meningitis with hyperglycemia and development of subdural effusion; successful treatment with chloramphenicol. *J Pediatr* **37**(6):909-16.

Ring, A., *et al*. (1998). Pneumococcal trafficking across the blood-brain barrier. Molecular analysis of a novel bidirectional pathway. *J Clin Invest* **102**(2):347-60.

Roberts, A. P. and P. Mullany (2009). A modular master on the move: the Tn*916* family of mobile genetic elements. *Trends Microbiol* **17**(6):251-8.

Roberts, M. S. and F. M. Cohan (1993). The effect of DNA sequence divergence on sexual isolation in *Bacillus*. *Genetics* **134**(2):401-8.

Robertson, G. T., *et al*. (2005). Use of an efflux-deficient *Streptococcus pneumoniae* strain panel to identify ABC-class multidrug transporters involved in intrinsic resistance to antimicrobial agents. *Antimicrob Agents Chemother* **49**(11):4781-3.

Robinson, K. A., *et al*. (2001). Epidemiology of invasive *Streptococcus pneumoniae* infections in the United States, 1995-1998: Opportunities for prevention in the conjugate vaccine era. *JAMA* **285**(13):1729-35.

Rocha, E. P. (2008). The organization of the bacterial genome. *Annu Rev Genet* **42**:211-33.

Rodenburg, G. D., *et al*. (2010). Effects of pneumococcal conjugate vaccine 2 years after its introduction, the Netherlands. *Emerg Infect Dis* **16**(5):816-23.

Rodriguez-Barradas, M. C., *et al*. (1997). Colonization by *Streptococcus pneumoniae* among human immunodeficiency virus-infected adults: prevalence of antibiotic resistance, impact of immunization, and characterization by polymerase chain reaction with BOX primers of isolates from persistent *S. pneumoniae* carriers. *J Infect Dis* **175**(3):590-7.

Romero, P., *et al*. (2009). Development of a prophage typing system and analysis of prophage carriage in *Streptococcus pneumoniae*. *Appl Environ Microbiol* **75**(6):1642-9.

Romero, P., *et al*. (2007). Isolation and characterization of a new plasmid p*Spn*P1 from a multidrug-resistant clone of *Streptococcus pneumoniae*. *Plasmid* **58**(1):51-60.

Ross, R. (1939). Acquired tolerance of pneumococcus to M. + B. *Lancet* **1**:1207-1208.

Ruckinger, S., *et al*. (2009a). Reduction in the incidence of invasive pneumococcal disease after general vaccination with 7-valent pneumococcal conjugate vaccine in Germany. *Vaccine* **27**(31):4136-41.

Ruckinger, S., *et al*. (2009b). Association of serotype of *Streptococcus pneumoniae* with risk of severe and fatal outcome. *Pediatr Infect Dis J* **28**(2):118-22.

Rudolph, K. M., *et al*. (2000). Serotype distribution and antimicrobial resistance patterns of invasive isolates of *Streptococcus pneumoniae*: Alaska, 1991-1998. *J Infect Dis* **182**(2):490-6.

Ruiz, M., *et al*. (1999). Etiology of community-acquired pneumonia: impact of age, comorbidity, and severity. *Am J Respir Crit Care Med* **160**(2):397-405.

Sa-Leao, R., *et al*. (2009). Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine. *Clin Microbiol Infect* **15**(11):1002-7.

Saah, A. J., *et al*. (1980). Relative resistance to penicillin in the pneumococcus. A prevalence and case-control study. *JAMA* **243**(18):1924-7.

Sabri, M., *et al*. (2011). Genome annotation and intraviral interactome for the *Streptococcus pneumoniae* virulent phage Dp-1. *J Bacteriol* **193**(2):551-62.

Saha, S. K., *et al*. (2009). Surveillance for invasive *Streptococcus pneumoniae* disease among hospitalized children in Bangladesh: antimicrobial susceptibility and serotype distribution. *Clin Infect Dis* **48 Suppl 2**:S75-81.

Salmon, D. E. (1884). Discrediting American science. *Science* **4**(86):303-5.

Saluja, S. K. and J. N. Weiser (1995). The genetic basis of colony opacity in *Streptococcus pneumoniae*: evidence for the effect of box elements on the frequency of phenotypic variation. *Mol Microbiol* **16**(2):215-27.

Sanchez, C. J., *et al*. (2010). The pneumococcal serine-rich repeat protein is an intra-species bacterial adhesin that promotes bacterial aggregation *in vivo* and in biofilms. *PLoS Pathog* **6**(8).

Sanchez-Pescador, R., *et al*. (1988). Homology of the TetM with translational elongation factors: implications for potential modes of *tetM*-conferred tetracycline resistance. *Nucleic Acids Res* **16**(3):1218.

Sandgren, A., *et al*. (2004). Effect of clonal and serotype-specific properties on the invasive capacity of *Streptococcus pneumoniae*. *J Infect Dis* **189**(5):785-96.

Sanger, F., *et al*. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12):5463-7.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol* **6**(5):526-38.

Schmidt, L. and C. Sesler (1943). Development of resistance to penicillin by pneumococci. *Proc Soc Exp Biol Med* **52**:353-357.

Schuchat, A., *et al*. (1997). Bacterial meningitis in the United States in 1995. Active Surveillance Team. *N Engl J Med* **337**(14):970-6.

Schuster, C., *et al*. (1998). Small cryptic plasmids of *Streptococcus pneumoniae* belong to the pC194/pUB110 family of rolling circle plasmids. *FEMS Microbiol Lett* **164**(2):427-31.

Schuster-Bockler, B., *et al*. (2004). HMM Logos for visualization of protein families. *BMC Bioinformatics* **5**:7.

Schwartz, D. C. and C. R. Cantor (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**(1):67-75.

Scott, J. A., *et al*. (1996). Serogroup-specific epidemiology of *Streptococcus pneumoniae*: associations with age, sex, and geography in 7,000 episodes of invasive disease. *Clin Infect Dis* **22**(6):973-81.

Sebert, M. E., *et al*. (2005). Pneumococcal HtrA protease mediates inhibition of competence by the CiaRH two-component signaling system. *J Bacteriol* **187**(12):3969-79.

Selander, R. K., *et al*. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* **51**(5):873-84.

Sell, S. H., *et al*. (1981). Clinical studies of pneumococcal vaccines in infants. I. Reactogenicity and immunogenicity of two polyvalent polysaccharide vaccines. *Rev Infect Dis* **3 Suppl**:S97-107.

Senger, E. (1886). Bakteriologische Untersuchungen über die Pneumonie und pneumonische Metastasen. *Arch f exper Path u Pharakol* **20**:389.

Seo, H. S., *et al*. (2008). A new model of pneumococcal lipoteichoic acid structure resolves biochemical, biosynthetic, and serologic inconsistencies of the current model. *J Bacteriol* **190**(7):2379-87.

Seral, C., *et al*. (2001). Distribution of resistance genes *tet*(M), *aph*3'-III, *cat*pC194 and the integrase gene of Tn*1545* in clinical *Streptococcus pneumoniae* harbouring *erm*(B) and *mef*(A) genes in Spain. *J Antimicrob Chemother* **47**(6):863-6.

Serino, L. and M. Virji (2000). Phosphorylcholine decoration of lipopolysaccharide differentiates commensal *Neisseriae* from pathogenic strains: identification of *licA*-type genes in commensal *Neisseriae*. *Mol Microbiol* **35**(6):1550-9.

Shakhnovich, E. A., *et al*. (2002). Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of *Neisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract. *Infect Immun* **70**(12):7161-4.

Shaper, M., *et al*. (2004). PspA protects *Streptococcus pneumoniae* from killing by apolactoferrin, and antibody to PspA enhances killing of pneumococci by apolactoferrin. *Infect Immun* **72**(9):5031-40.

Shapiro, E. D., *et al*. (1991). The protective efficacy of polyvalent pneumococcal polysaccharide vaccine. *N Engl J Med* **325**(21):1453-60.

Shapiro, E. D. and J. D. Clemens (1984). A controlled evaluation of the protective efficacy of pneumococcal vaccine for patients at high risk of serious pneumococcal infections. *Ann Intern Med* **101**(3):325-30.

Sharma, C. M., *et al*. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**(7286):250-5.

Shaw, J. H. and D. B. Clewell (1985). Complete nucleotide sequence of macrolide-lincosamide-streptogramin B-resistance transposon Tn*917* in *Streptococcus faecalis*. *J Bacteriol* **164**(2):782-96.

Sheehan, M. M., *et al*. (1997). The lytic enzyme of the pneumococcal phage Dp-1: a chimeric lysin of intergeneric origin. *Mol Microbiol* **25**(4):717-25.

Shi, Z. Y., *et al*. (1998). Identification of three major clones of multiply antibiotic-resistant *Streptococcus pneumoniae* in Taiwanese hospitals by multilocus sequence typing. *J Clin Microbiol* **36**(12):3514-9.

Shivshankar, P., *et al*. (2009). The *Streptococcus pneumoniae* adhesin PsrP binds to Keratin 10 on lung cells. *Mol Microbiol* **73**(4):663-79.

Shoemaker, N. B., *et al*. (1979). Organization and transfer of heterologous chloramphenicol and tetracycline resistance genes in pneumococcus. *J Bacteriol* **139**(2):432-41.

Shoemaker, N. B., *et al*. (1980). DNase-resistant transfer of chromosomal *cat* and *tet* insertions by filter mating in Pneumococcus. *Plasmid* **3**(1):80-7.

Sibold, C., *et al*. (1994). Mosaic *pbpX* genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from *pbpX* genes of a penicillin-sensitive *Streptococcus oralis*. *Mol Microbiol* **12**(6):1013-23.

Sibold, C., *et al*. (1991). Novel plasmids in clinical strains of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **61**(1):91-5.

Sibold, C., *et al*. (1992). Genetic relationships of penicillin-susceptible and -resistant *Streptococcus pneumoniae* strains isolated on different continents. *Infect Immun* **60**(10):4119-26.

Siegman-Igra, Y., *et al*. (2002). Reappraisal of community-acquired bacteremia: a proposal of a new classification for the spectrum of acquisition of bacteremia. *Clin Infect Dis* **34**(11):1431-9.

Siguier, P., *et al*. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**(Database issue):D32-6.

Sims, R. V., *et al*. (1988). The clinical effectiveness of pneumococcal vaccine in the elderly. *Ann Intern Med* **108**(5):653-7.

Singleton, R. J., *et al*. (2007). Invasive pneumococcal disease caused by nonvaccine serotypes among Alaska native children with high levels of 7-valent pneumococcal conjugate vaccine coverage. *JAMA* **297**(16):1784-92.

Sirotnak, F. M., *et al*. (1969). Increased dihydrofolate reductase synthesis in *Diplococcus pneumoniae* following translatable alteration of the structural gene. II. Individual and dual effects on the properties and rate of synthesis of the enzyme. *Genetics* **61**(2):313-26.

Slamti, L. and D. Lereclus (2002). A cell-cell signaling peptide activates the PlcR virulence regulon in bacteria of the *Bacillus cereus* group. *EMBO J* **21**(17):4550-9.

Sleeman, K. L., *et al*. (2006). Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. *J Infect Dis* **194**(5):682-8.

Sloyer, J. L., Jr., *et al*. (1981). Efficacy of pneumococcal polysaccharide vaccine in preventing acute otitis media in infants in Huntsville, Alabama. *Rev Infect Dis* **3 Suppl**:S119-23.

Smith, A. M. and K. P. Klugman (2001). Alterations in MurM, a cell wall muropeptide branching enzyme, increase high-level penicillin and cephalosporin resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **45**(8):2393-6.

Smith, C. B., *et al*. (1976). Interactions between viruses and bacteria in patients with chronic bronchitis. *J Infect Dis* **134**(6):552-61.

Smith, H. O., *et al*. (1999). DNA uptake signal sequences in naturally transformable bacteria. *Res Microbiol* **150**(9-10):603-16.

Smith, M. D. and W. R. Guild (1979). A plasmid in *Streptococcus pneumoniae*. *J Bacteriol* **137**(2):735-9.

Smith, T., *et al*. (1993). Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children. *Epidemiol Infect* **111**(1):27-39.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications Genet Mol Biol* **3**(1):3.

Sneath, P., *et al*. (1975). Detecting evolutionary incompatibilities from protein sequences. *Syst Zool* **24**(3):311-322.

Snyder, L. A., *et al*. (2009). Comparative analysis of two *Neisseria gonorrhoeae* genome sequences reveals evidence of mobilization of Correia Repeat Enclosed Elements and their role in regulation. *BMC Genomics* **10**:70.

Sommer, S., *et al*. (1993). The appearance of the UmuD'C protein complex in *Escherichia coli* switches repair from homologous recombination to SOS mutagenesis. *Mol Microbiol* **10**(5):963-71.

Sorek, R., *et al*. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**(5855):1449-52.

Spratt, B. G. (1975). Distinct penicillin binding proteins involved in the division, elongation, and shape of *Escherichia coli* K12. *Proc Natl Acad Sci U S A* **72**(8):2999-3003.

Spratt, B. G. and B. M. Greenwood (2000). Prevention of pneumococcal disease by vaccination: does serotype replacement matter? *Lancet* **356**(9237):1210-1.

Spratt, B. G. and A. B. Pardee (1975). Penicillin-binding proteins and cell shape in *E. coli*. *Nature* **254**(5500):516-7.

Spratt, B. G., *et al*. (1989). Recruitment of a penicillin-binding protein gene from *Neisseria flavescens* during the emergence of penicillin resistance in *Neisseria meningitidis*. *Proc Natl Acad Sci U S A* **86**(22):8988-92.

Stamatakis, A., *et al*. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**(4):456-63.

Stein, K. E. (1992). Thymus-independent and thymus-dependent responses to polysaccharide antigens. *J Infect Dis* **165 Suppl 1**:S49-52.

Stephens, D. S., *et al*. (2005). Incidence of macrolide resistance in *Streptococcus pneumoniae* after introduction of the pneumococcal conjugate vaccine: population-based assessment. *Lancet* **365**(9462):855-63.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol Biol Evol* **2**(6):539-56.

Sternberg, G. M. (1881). A fatal form of septicemia in the rabbit produced by subcutaneous injection of human saliva. *Natl Board of Health Bull* **3**(87):108.

Sternberg, G. M. (1882). Induced septicaemia in the rabbit. *Am J Med Sci* **84**(167):69-76.

Sternberg, G. M. (1885). The pneumonia-coccus of Friedlander (Micrococcus Pasteuri, Sternberg). *Am J Med Sci* **179**:106-122.

Stool, S. E. and M. J. Field (1989). The impact of otitis media. *Pediatr Infect Dis J* **8**(1 Suppl):S11-4.

Sutcliffe, J., *et al*. (1996). *Streptococcus pneumoniae* and *Streptococcus pyogenes* resistant to macrolides but sensitive to clindamycin: a common resistance pattern mediated by an efflux system. *Antimicrob Agents Chemother* **40**(8):1817-24.

Syrjanen, R. K., *et al*. (2001). Nasopharyngeal carriage of *Streptococcus pneumoniae* in Finnish children younger than 2 years old. *J Infect Dis* **184**(4):451-9.

Szklarczyk, D., *et al*. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**(Database issue):D561-8.

Takala, A. K., *et al*. (1995). Risk factors for primary invasive pneumococcal disease among children in Finland. *JAMA* **273**(11):859-64.

Talamon, C. (1883). Coccus de la pneumonie. *Bull Soc Anat Paris* **58**:475-481.

Talbot, T. R., *et al*. (2005). Asthma as a risk factor for invasive pneumococcal disease. *N Engl J Med* **352**(20):2082-90.

Talbot, T. R., *et al*. (2004). Reduction in high rates of antibiotic-nonsusceptible invasive pneumococcal disease in Tennessee after introduction of the pneumococcal conjugate vaccine. *Clin Infect Dis* **39**(5):641-8.

Taniai, H., *et al*. (2008). Concerted action of lactate oxidase and pyruvate oxidase in aerobic growth of *Streptococcus pneumoniae*: role of lactate as an energy source. *J Bacteriol* **190**(10):3572-9.

Tankovic, J., *et al*. (1996). Contribution of mutations in *gyrA* and *parC* genes to fluoroquinolone resistance of mutants of *Streptococcus pneumoniae* obtained *in vivo* and *in vitro*. *Antimicrob Agents Chemother* **40**(11):2505-10.

Tarasi, A., *et al*. (1997). Spread of the serotype 23F multidrug-resistant *Streptococcus pneumoniae* clone to South Korea. *Microb Drug Resist* **3**(1):105-9.

Tarasi, A., *et al*. (1995). Penicillin-resistant and multidrug-resistant *Streptococcus pneumoniae* in a pediatric hospital in Zagreb, Croatia. *Microb Drug Resist* **1**(2):169-76.

Taylor, S. N. and C. V. Sanders (1999). Unusual manifestations of invasive pneumococcal infection. *Am J Med* **107**(1A):12S-27S.

Teele, D. W., *et al*. (1981). Use of pneumococcal vaccine for prevention of recurrent acute otitis media in infants in Boston. The Greater Boston Collaborative Otitis Media Study Group. *Rev Infect Dis* **3 Suppl**:S113-8.

Tempest, B., *et al*. (1974). Distribution of the sensitivities to penicillin of types of *Diplococcus pneumoniae* in an American Indian population. *J Infect Dis* **130**(1):67-9.

Tettelin, H., *et al*. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**(39):13950-5.

Tettelin, H., *et al*. (2001). Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**(5529):498-506.

Thornsberry, C., *et al*. (1999). Survey of susceptibilities of *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis* isolates to 26 antimicrobial agents: a prospective U.S. study. *Antimicrob Agents Chemother* **43**(11):2612-23.

Thornton, G. F. and V. T. Andriole (1966). Laboratory and clinical studies of a new antibiotic, cephaloridine, in the treatment of gram-positive infections. *Yale J Biol Med* **39**(1):9-20.

Thys, J. P., *et al*. (1989). Quinolones in the treatment of lower respiratory tract infections. *Rev Infect Dis* **11 Suppl 5**:S1212-9.

Tilley, S. J., *et al*. (2005). Structural basis of pore formation by the bacterial toxin pneumolysin. *Cell* **121**(2):247-56.

Tiraby, J. G. and M. S. Fox (1973). Marker discrimination in transformation and mutation of pneumococcus. *Proc Natl Acad Sci U S A* **70**(12):3541-5.

Tiraby, J. G., *et al*. (1975). Pneumococcal bacteriophages. *Virology* **68**(2):566-9.

Tomasz, A. (1966). Model for the mechanism controlling the expression of competent state in Pneumococcus cultures. *J Bacteriol* **91**(3):1050-61.

Tomasz, A., *et al*. (1998). Molecular epidemiologic characterization of penicillin-resistant *Streptococcus pneumoniae* invasive pediatric isolates recovered in six Latin-American countries: an overview. PAHO/Rockefeller University Workshop. Pan American Health Organization. *Microb Drug Resist* **4**(3):195-207.

Tomasz, A. and J. L. Mosser (1966). On the nature of the pneumococcal activator substance. *Proc Natl Acad Sci U S A* **55**(1):58-66.

Torzillo, P. J., *et al*. (1995). Invasive pneumococcal disease in central Australia. *Med J Aust* **162**(4):182-6.

Trotman, J., *et al*. (1995). Invasive pneumococcal disease in central Australia. *Clin Infect Dis* **20**(6):1553-6.

Trzcinski, K., *et al*. (2004). Single-step capsular transformation and acquisition of penicillin resistance in *Streptococcus pneumoniae*. *J Bacteriol* **186**(11):3447-52.

Tsai, I. J., *et al*. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11**(4):R41.

Tu, A. H., *et al*. (1999). Pneumococcal surface protein A inhibits complement activation by *Streptococcus pneumoniae*. *Infect Immun* **67**(9):4720-4.

Tuomanen, E., *et al.*, Eds. (2004). The pneumococcus. Washington, DC, ASM Press.

Turner, P., *et al*. (2011). Improved detection of nasopharyngeal co-colonization by multiple pneumococcal serotypes using latex agglutination or molecular serotyping by microarray. *J Clin Microbiol*.

Tzianabos, A. O. (2000). Polysaccharide immunomodulators as therapeutic agents: structural aspects and biologic function. *Clin Microbiol Rev* **13**(4):523-33.

Valles, J., *et al*. (2003). Community-acquired bloodstream infection in critically ill adult patients: impact of shock and inappropriate antibiotic therapy on survival. *Chest* **123**(5):1615-24.

Valouev, A., *et al*. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**(7):1051-63.

van Belkum, A., *et al*. (1996). Novel BOX repeat PCR assay for high-resolution typing of *Streptococcus pneumoniae* strains. *J Clin Microbiol* **34**(5):1176-9.

van der Meer, J. T., *et al*. (1991). Distribution, antibiotic susceptibility and tolerance of bacterial isolates in culture-positive cases of endocarditis in The Netherlands. *Eur J Clin Microbiol Infect Dis* **10**(9):728-34.

Veaute, X., *et al*. (2005). UvrD helicase, unlike Rep helicase, dismantles RecA nucleoprotein filaments in *Escherichia coli*. *EMBO J* **24**(1):180-9.

Vernikos, G. S. and J. Parkhill (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**(18):2196-203.

Vijayakumar, M. N., *et al*. (1986). Structure of a conjugative element in *Streptococcus pneumoniae*. *J Bacteriol* **166**(3):978-84.

Vos, M. (2009). Why do bacteria engage in homologous recombination? *Trends Microbiol* **17**(6):226-32.

Vos, M. and X. Didelot (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**(2):199-208.

Vulic, M., *et al*. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A* **94**(18):9763-7.

Waite, R. D., *et al*. (2003). Spontaneous sequence duplications within capsule genes *cap8E* and *tts* control phase variation in *Streptococcus pneumoniae* serotypes 8 and 37. *Microbiology* **149**(Pt 2):497-504.

Waite, R. D., *et al*. (2001). Spontaneous sequence duplication within an open reading frame of the pneumococcal type 3 capsule locus causes high-frequency phase variation. *Mol Microbiol* **42**(5):1223-32.

Walker, J. A., *et al*. (1987). Molecular cloning, characterization, and complete nucleotide sequence of the gene for pneumolysin, the sulfhydryl-activated toxin of *Streptococcus pneumoniae*. *Infect Immun* **55**(5):1184-9.

Wani, J. H., *et al*. (1996). Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of *Streptococcus pneumoniae*. *Infect Immun* **64**(10):3967-74.

Wannamaker, L. and J. Matsen, Eds. (1972). Streptococci and streptococcal diseases. New York, Academic Press Inc.

Ward, J. (1981). Antibiotic-resistant Streptococcus pneumoniae: clinical and epidemiologic aspects. *Rev Infect Dis* **3**(2):254-66.

Warren, M. J. and M. P. Jennings (2003). Identification and characterization of *pptA*: a gene involved in the phase-variable expression of phosphorylcholine on pili of *Neisseria meningitidis*. *Infect Immun* **71**(12):6892-8.

Watson, D. A., *et al*. (1993). A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery. *Clin Infect Dis* **17**(5):913-24.

Watson, K., *et al*. (2006). Upper respiratory tract bacterial carriage in Aboriginal and non-Aboriginal children in a semi-arid area of Western Australia. *Pediatr Infect Dis J* **25**(9):782-90.

Weichselbaum, A. (1886). Über die Aetiolgie der acuten Lungen- und Rippenfellentzündungen. *Med Jhrbchr* **1**:483-554.

Weiller, G. F. (1998). Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol Biol Evol* **15**(3):326-35.

Weinberger, D. M., *et al*. (2011). Serotype replacement in disease after pneumococcal vaccination. *Lancet*.

Weisblum, B. (1967). Pneumococcus resistant to erythromycin and lincomycin. *Lancet* **1**(7494):843-4.

Weiser, J. N., *et al*. (1994). Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonization. *Infect Immun* **62**(6):2582-9.

Weiser, J. N., *et al*. (1998). The phosphorylcholine epitope undergoes phase variation on a 43-kilodalton protein in *Pseudomonas aeruginosa* and on pili of *Neisseria meningitidis* and *Neisseria gonorrhoeae*. *Infect Immun* **66**(9):4263-7.

Weiser, J. N., *et al*. (1997). Decoration of lipopolysaccharide with phosphorylcholine: a phase-variable characteristic of *Haemophilus influenzae*. *Infect Immun* **65**(3):943-50.

Weisfelt, M., *et al*. (2006). Clinical features, complications, and outcome in adults with pneumococcal meningitis: a prospective case series. *Lancet Neurol* **5**(2):123-9.

Whatmore, A. M. and C. G. Dowson (1999). The autolysin-encoding gene (*lytA*) of *Streptococcus pneumoniae* displays restricted allelic variation despite localized recombination events with genes of pneumococcal bacteriophage encoding cell wall lytic enzymes. *Infect Immun* **67**(9):4551-6.

Whitby, L. (1938). Chemotherapy of pneumococcal and other infections with 2-(*p*-aminobenzenesulphonamido) pyridine. *Lancet* **1**:1210-1212.

White, B., *et al*. (1938). *The biology of pneumococcus*. New York, The Commonwealth Fund.

Whiting, G. C. and S. H. Gillespie (1996). Incorporation of choline into *Streptococcus pneumoniae* cell wall antigens: evidence for choline kinase activity. *FEMS Microbiol Lett* **138**(2-3):141-5.

Whitney, C. G., *et al*. (2003). Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine. *N Engl J Med* **348**(18):1737-46.

WHO (2003). Pneumococcal vaccines. *Wkly Epidemiol Rec* **78**:100-19.

Widdowson, C. A., *et al*. (2000). Acquisition of chloramphenicol resistance by the linearization and integration of the entire staphylococcal plasmid pC194 into the chromosome of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **44**(2):393-5.

Wijnands, W. J., *et al*. (1986). Enoxacin in lower respiratory tract infections. *J Antimicrob Chemother* **18**(6):719-27.

Wilhelm, B. T., *et al*. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**(7199):1239-43.

Willey, J. M. and W. A. van der Donk (2007). Lantibiotics: peptides of diverse structure and function. *Annu Rev Microbiol* **61**:477-501.

Williams, E. W., *et al*. (1981). *Streptococcus pneumoniae* resistant to penicillin and chloramphenicol in the U.K. *Lancet* **2**(8248):699.

Williams, G. (1966). *Adaptation and natural selection*. Princeton, Princeton University Press.

Williamson, R., *et al*. (1980). *In vivo* interaction of beta-lactam antibiotics with the penicillin-binding proteins of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **18**(4):629-37.

Wolf, B. and R. D. Hotchkiss (1963). Genetically modified folic acid synthesizing enzymes of pneumococcus. *Biochemistry* **2**:145-50.

Wood, B. and W. Holzapfel, Eds. (1995). The Genera of Lactic Acid Bacteria. Glasgow, Blackie Academic and Professional.

Wright, A., *et al*. (1914). Observations on prophylactic inoculation against pneumococcus infections, and on the results which have have been achieved by it. *Lancet* **183**(4714):1-10.

Xu, P., *et al*. (2007). Genome of the opportunistic pathogen *Streptococcus sanguinis*. *J Bacteriol* **189**(8):3166-75.

Yamada, T. and J. Carlsson (1973). Phosphoenolpyruvate carboxylase and ammonium metabolism in oral streptococci. *Arch Oral Biol* **18**(7):799-812.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8):1586-91.

Yesilkaya, H., *et al*. (2000). Role of manganese-containing superoxide dismutase in oxidative stress and virulence of *Streptococcus pneumoniae*. *Infect Immun* **68**(5):2819-26.

Yildirim, I., *et al*. (2010). Serotype specific invasive capacity and persistent reduction in invasive pneumococcal disease. *Vaccine* **29**(2):283-8.

Yoder-Himes, D. R., *et al*. (2009). Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc Natl Acad Sci U S A* **106**(10):3976-81.

Zaufal, E. (1887). Mikrooganismen im Secrete der Otitis media acuta. *Prag Med Wchnschr* **12**(27):225-227.

Zerbino, D. R. and E. Birney (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**(5):821-9.

Zhang, J. R., *et al*. (2000). The polymeric immunoglobulin receptor translocates pneumococci across human nasopharyngeal epithelial cells. *Cell* **102**(6):827-37.

Zhou, F., *et al*. (2008). Nezha, a novel active miniature inverted-repeat transposable element in cyanobacteria. *Biochem Biophys Res Commun* **365**(4):790-4.

Zighelboim, S. and A. Tomasz (1980). Penicillin-binding proteins of multiply antibiotic-resistant South African strains of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **17**(3):434-42.

Appendix I: Primer sequences

Primer Name	Sequence	Use
NTPase_P_L	TTTGGAAGGAATGGTTCAGG	Check for presence of SPN23F00710
NTPase_P_R	CCCTTCTTGAGCAATTTTGA	Check for presence of SPN23F00710
NTPase_N_L	CCCCACGACCACATAACTCT	Check for absence of SPN23F00710
NTPase_N_R	TTAGGGATTTCCCCCAACTC	Check for absence of SPN23F00710
Lanti_L	AGGGCATTCGTACGGTAGTG	Check for presence of lantibiotic synthesis operon
Lanti_R	AAAGGAGACATGCCACCAAG	Check for presence of lantibiotic synthesis operon
pSpn1_L	CGCCCAAGCATAGACATCTT	Check circularization of plasmid p*Spn*1
pSpn1_R	ACCAGAGCAGATCGAAGCAT	Check circularization of plasmid p*Spn*1
ICE_L_L	TCTGGTGAACGCTTGTACCTT	Check insertion site of ICE*Spn*8140
ICE_L_R	CGCCCAAGCATAGACATCTT	Check insertion site of ICE*Spn*8140
ICE_R_L	ACCAGAGCAGATCGAAGCAT	Check insertion site of ICE*Spn*8140
ICE_R_R	AGATGGTGACGATGTCTTGG	Check insertion site of ICE*Spn*8140
comYC_L	TACGATTTGCCCCTCCATT	Check insertion site of group 2b prophage
Phage_L	TCAAACCAGACGGAACACTG	Check insertion site of group 2b prophage
comYC_R	GGTTTTTATCTTTGTGGCACTG	Check insertion site of group 2b prophage
Phage_R	GCAACTTGGCAAGCTAGGAC	Check insertion site of group 2b prophage
13150_L_L	TCTAACGCTTGGCAGGAAAC	Check partial deletion of ICE*Spn*23FST81
13150_L_R	TGTCCTAATATAGATTCAACGCACT	Check partial deletion of ICE*Spn*23FST81

13150_R_L	GTGCGTTGAATCTATATTAGGACA	Check partial deletion of ICE*Spn*23FST81
13150_R_R	CACCACAGAGTGTCCTTTCTTAAC	Check partial deletion of ICE*Spn*23FST81
Tn916_att1L	GACAGACCGACACAAGCAGA	Check for insertion of Omega element
Tn916_att1R	TCACGCTTGAAGCCAAGATA	Check for insertion of Omega element
Tn916_att2L	CTACCGGTGAACCTGTTTGC	Check for insertion of Mega or Tn*917* element
Tn916_att2R	GTGAACAAGTGGGAGCATTG	Check for insertion of Mega or Tn*917* element
hexBL	GCAGCTGCATCGTGAAATAC	Amplify region upstream of *hexB*
hexBR	TAGAGGTAGCCTGGGTTCCA	Amplify region downstream of *hexB*
hexBboxL	GCTGACGTGGTTTGAAGAGA	Amplify region upstream of *hexB*
hexBboxR	CCACGTCAGTTTTATCAGTAATCTC	Amplify region downstream of *hexB*
ermBL	GTCATGGATATCTGGAAATAAGACTTAGAAGCAAACTT	Amplify *ermB* gene from Tn*917*
ermBR	GATATCTCTCCATTCCCTTTAGTAACGTGT	Amplify *ermB* gene from Tn*917*
T7	TAATACGACTCACTATAGGG	Amplify MCS of pGEM-T Easy
SM_L	CCAGCTTTGAACGTTGGTTA	Check for DNA contamination of RNA samples
SM_R	AACACCTGCAGTATCAAGTGC	Check for DNA contamination of RNA samples
05060_PL	TCAAACCACGTCAGCTTCAC	Check for expression of SPN23F05060 locus
05060_PR	CTTGTCCACCACTTCTGCAA	Check for expression of SPN23F05060 locus
05060_PS	TAAAATCATTTTATACTCTTCGAAAATCTC	Check for expression of SPN23F05060 locus
16220_PL	ACCTGATAAAATTTCAGTAAAATGC	Check for expression of SPN23F16220
16220_PR	TGTAAACGAGTAAAAGCGAAT	Check for expression of SPN23F16220
17630_PL	GAAGCCAAAACCTTCATCCA	Check for

		expression of SPN23F17630 locus
17630_PR	CAGGCTGCTCAAAACACG	Check for expression of SPN23F17630 locus
17630_PS	TTCACCTTTTGTGGATTGGTC	Check for expression of SPN23F17630 locus
21390_PL	TGTTCATATTCTAGGAAGATTTGTTGA	Check for expression of SPN23F21390 locus
21390_PR	GCAGGTTGCTCAAAACACTG	Check for expression of SPN23F21390 locus
21390_PS	TCCATAATATCTATAGTGGATTTACCC	Check for expression of SPN23F21390 locus
Tbox_PL	AATCTGCAATCGCAGCTAGG	Check for expression of *trp* operon leader sequence
Tbox_PR	CGTTACCAACGCCCTCAC	Check for expression of *trp* operon leader sequence
IntGL	GCCTGCCACTTGTAGGTTTT	Clone sequence upstream of *patAB*
InGR	GATAGGGCAGAAGAGCATCC	Clone sequence upstream of *patAB*

Appendix II: PMEN1 strains

Strain	Country	Year	Serotype	Inferred Serotype	ST	Inferred ST	MIC (Pen)	Source
ARG 740	Argentina	1995	23F	23F		81	1	Alexander Tomasz
3122	Canada	1994	19A	19A	81	81		Dylan Pillai
9295	Canada	1997	19F	19F	81	81	0.1	Dylan Pillai
9426	Canada	1997	19F	19F	81	81	4	Dylan Pillai
11126	Canada	1998	19F	19F	81	81	4	Dylan Pillai
34117	Canada	2007	23F	23F	81	81		Dylan Pillai
36148	Canada	2008	23F	23F	81	81		Dylan Pillai
HK P1	China	2000	23F	23F	81	81	2	Jae Hoon Song
HK P38	China	2000	23F	23F	81	81	2	Jae Hoon Song
TW M30	China	2000	19F	19F	81	81	2	Jae Hoon Song
HK P57	China	2001	23F	23F	81	81	2	Jae Hoon Song
HK P58	China	2001	19F	19F	81	81	1	Jae Hoon Song
HK P65	China	2001	23F	23F	81	81	2	Jae Hoon Song
S-030	China	2006	15B	15C	83	83		Lesley McGee
S-054	China	2006	15B	15C	83	83		Lesley McGee
S-055	China	2006	15B	15C	83	83		Lesley McGee
S-153	China	2006	23F	23F	81	81		Lesley McGee
COL 252	Colombia	1995	23F	23F		81	2	Alexander Tomasz
385	Croatia	2001	23F	23F	285	81		Lesley McGee
412	Croatia	2001	23F	23F	285	81		Lesley McGee
118-87	Denmark	1987	23F	23F		81	2	Lotte Lambertsen
484-93	Denmark	1993	23F	23F		DLV81	6	Lotte Lambertsen
88-93	Denmark	1993	23F	23F		81	3	Lotte Lambertsen
135-94	Denmark	1994	23F	23F		81	3	Lotte Lambertsen
848-95	Denmark	1995	23F	23F		81	2	Lotte Lambertsen
1392-96	Denmark	1996	23F	23F		81	4	Lotte Lambertsen
1437-96	Denmark	1996	23F	23F		81	8	Lotte Lambertsen
933-98	Denmark	1998	23F	23F		81	4	Lotte Lambertsen
BM4200	France	1978	23F	23F		1010		Sylvain Brisse
9300	France	2002	23F	23F	1598	1598	0.5	Mark van der Linden
9409	France	2002	23F	23F	81	81	2	Mark van der Linden
11865	France		23F	23F	81	81	1	Mark van der Linden
11867	France		23F	23F	81	81	2	Mark van der Linden
11868	France		23F	23F	81	81	1	Mark van der Linden
11875	France		23F	23F	81	81	1	Mark van der Linden
11876	France		23F	23F	81	81	1	Mark van der Linden
11881	France		23F	23F	81	81	2	Mark van der Linden
11883	France		23F	23F	81	81	1	Mark van der Linden
2859	Germany	1992	23F	23F	81	81	0.5	Mark van der Linden
2905	Germany	1992	23F	23F	81	81	2	Mark van der Linden
2934	Germany	1992	23F	23F	81	81	1	Mark van der Linden
3027	Germany	1992	23F	23F	81	81	0.5	Mark van der Linden
3296	Germany	1992	23F	23F	81	81	0.5	Mark van der Linden
3413	Germany	1992	23F	23F	81	81	0.5	Mark van der Linden
3696	Germany	1993	23F	23F	81	81	0.5	Mark van der Linden
3842	Germany	1993	23F	23F	81	81	0.5	Mark van der Linden

3848	Germany	1993	23F	23F	81	81	0.25	Mark van der Linden
3901	Germany	1993	23F	23F	81	81	0.5	Mark van der Linden
4114	Germany	1994	23F	23F	81	81	0.5	Mark van der Linden
4545	Germany	1995	23F	23F	81	81	0.5	Mark van der Linden
PMEN1_4605	Germany	1995	23F	23F	81	81	0.5	Mark van der Linden
5195	Germany	1996	23F	23F	81	81	0.25	Mark van der Linden
5204	Germany	1996	23F	23F	81	81	0.25	Mark van der Linden
24478	Germany	1997		23F	81	81	1	Mark van der Linden
13150	Germany	1998	23F	23F	81	81	2	Mark van der Linden
277	Germany	1998	23F	23F	81	81	0.5	Mark van der Linden
13804	Germany	1999	23F	23F	81	81	2	Mark van der Linden
6025	Germany	2000	23F	23F	1595	1595	1	Mark van der Linden
6039	Germany	2000	23F	23F	81	81	1	Mark van der Linden
902	Germany	2001	23F	23F	81	81	0.5	Mark van der Linden
4826	Germany	1995	19F	19F	1591	1591	1	Mark van der Linden
Kor 136	Korea	1998	23F	23F	81	81	2	Jae Hoon Song
Kor 138	Korea	1998	23F	23F	81	81	2	Jae Hoon Song
Kor 146	Korea	1998	6A	06A	81	81	2	Jae Hoon Song
Kor 148	Korea	1998	UT	14	81	81	1	Jae Hoon Song
Kor 28	Korea	1999	19F	19F	81	81	2	Jae Hoon Song
Kor 82	Korea	1999	23F	23F	81	81	2	Jae Hoon Song
Kor 906	Korea	1999	19F	19F	81	81	1	Jae Hoon Song
Kor 909	Korea	1999	23B	19F	81	81	1	Jae Hoon Song
Kor 910	Korea	1999	23F	23F	81	81	1	Jae Hoon Song
Kor 14	Korea	2000	6A	06A	81	81	4	Jae Hoon Song
Kor 16	Korea	2000	23F	23F	81	81	2	Jae Hoon Song
04-004	Korea	2004		19F	81	81	2	Jae Hoon Song
04-036	Korea	2004		06A	81	81	2	Jae Hoon Song
04-124	Korea	2004		23F	81	81	2	Jae Hoon Song
07-050	Korea	2007	6A	06A	81	81	2	Jae Hoon Song
07-117	Korea	2007	6A	06A	81	81	2	Jae Hoon Song
08-B-120	Korea	2007	23F	23F	81	81	2	Jae Hoon Song
Mal M12	Malaysia	1999	19F	19F	81	81	2	Jae Hoon Song
Mal M6	Malaysia	1999	19F	19F	81	81	2	Jae Hoon Song
Mal P4	Malaysia	1999	19F	19F	81	81	2	Jae Hoon Song
13 HIM	Mexico	1994	23F	23F		DLV81	4	Alexander Tomasz
35 HIM	Mexico	1994	23F	23F		81	8	Alexander Tomasz
46 HIM	Mexico	1994	23F	23F		81	4	Alexander Tomasz
9 HIM	Mexico	1994	23F	23F		81	4	Alexander Tomasz
109 HIM	Mexico	1995	23F	23F		81	4	Alexander Tomasz
11 CMN	Mexico	1995		23F		81	8	Alexander Tomasz
23 CMN	Mexico	1995	23F	23F		81	4	Alexander Tomasz
35 CMN	Mexico	1995	23F	23F		81	4	Alexander Tomasz
36 CMN	Mexico	1995	23F	23F		81	4	Alexander Tomasz
76 HIM	Mexico	1995	23F	23F		81	2	Alexander Tomasz
132 HIM	Mexico	1996	23F	23F		81	4	Alexander Tomasz
61 CMN	Mexico	1996	23F	23F		SLV81	4	Alexander Tomasz
8454	Portugal	2002	23F	23F	81	81	1	Mark van der Linden
23782	Russia	2003	23F	23F	81	81	2	Sergei Sidorenko
23784	Russia	2003	23F	23F	81	81	0.06	Sergei Sidorenko
23748	Russia	2004	23F	23F	81	81	2	Sergei Sidorenko

23771	Russia	2004	23F	23F	81	81	2	Sergei Sidorenko
23805	Russia	2004	23F	23F	81	81	2	Sergei Sidorenko
23841	Russia	2004		23F	81	81	2	Sergei Sidorenko
23809	Russia	2005	23F	23F	SLV81 ?	81	4	Sergei Sidorenko
23920	Russia	2005		23F	81	81	4	Sergei Sidorenko
SI M1	Singapore	2000	23F	23F	81	81	4	Jae Hoon Song
SI O1	Singapore	2000	19F	19F	81	81	2	Jae Hoon Song
SI O2	Singapore	2000	23F	23F	81	81	1	Jae Hoon Song
SI P12	Singapore	2000	19F	19F	81	81	1	Jae Hoon Song
SA151	South Africa	1989	23F	23F		SLV81	2	Anne von Gottberg
SA142	South Africa	1990	23F	23F		81	2	Anne von Gottberg
SA148	South Africa	1990	23F	23F		81	4	Anne von Gottberg
SA150	South Africa	1990	23F	23F		81	2	Anne von Gottberg
SA152	South Africa	1991	23F	23F		81	2	Anne von Gottberg
SA141	South Africa	1995	23F	23F		81	2	Anne von Gottberg
SA149	South Africa	1995	23F	23F		81	2	Anne von Gottberg
SA140	South Africa	1996	23F	23F		81	4	Anne von Gottberg
SA146	South Africa	1996	23F	23F		81	2	Anne von Gottberg
SA147	South Africa	1996	23F	23F		81	4	Anne von Gottberg
SA153	South Africa	1996	23F	23F		81	4	Anne von Gottberg
B743	South Africa	2001	23F	23F	81	81		Lesley McGee
SA1	South Africa	2001	23F	23F		81	0.75	Anne von Gottberg
SA2	South Africa	2001	23F	23F		SLV81	1	Anne von Gottberg
SA3	South Africa	2001	23F	23F		81	0.5	Anne von Gottberg
SA4	South Africa	2001	23F	23F		SLV81	1	Anne von Gottberg
SA5	South Africa	2001	23F	23F		DLV81	0.5	Anne von Gottberg
SA6	South Africa	2001	23F	23F		81	0.25	Anne von Gottberg
SA10	South Africa	2002	23F	23F		81	1	Anne von Gottberg
SA12	South Africa	2002	23F	23F		81	1	Anne von Gottberg
SA7	South Africa	2002	23F	23F		SLV81	1	Anne von Gottberg
SA8	South Africa	2002	23F	23F		81	0.25	Anne von Gottberg
SA9	South Africa	2002	23F	23F		81	0.25	Anne von Gottberg
SA17	South Africa	2003	23F	23F		SLV81	1	Anne von Gottberg

SA20	South Africa	2003	23F	23F		81	1	Anne von Gottberg
SA21	South Africa	2003	23F	23F		81	1	Anne von Gottberg
SA23	South Africa	2003	23F	23F		81	0.5	Anne von Gottberg
SA25	South Africa	2003	23F	23F		81	1	Anne von Gottberg
SA27	South Africa	2003	23F	23F		SLV81	1	Anne von Gottberg
SA29	South Africa	2004	23F	23F		81	1	Anne von Gottberg
SA34	South Africa	2004	23F	23F		81	1	Anne von Gottberg
SA36	South Africa	2004	23F	23F		81	1	Anne von Gottberg
SA57	South Africa	2005	23F	23F		81	1	Anne von Gottberg
SA61	South Africa	2005	23F	23F		2395	1	Anne von Gottberg
SA101	South Africa	2006	23F	23F		81	1	Anne von Gottberg
SA106	South Africa	2006	23F	23F		81	1	Anne von Gottberg
SA78	South Africa	2006	23F	23F		SLV81	1	Anne von Gottberg
ATCC 700669	Spain	1984	23F	23F	81	81		Tim Mitchell
8140	Spain	2001	19A	19A	81	81	2	Mark van der Linden
8143	Spain	2001	19A	19A	81	81	2	Mark van der Linden
B532	Spain	2001	23F	23F	81	81		Lesley McGee
9029	Spain	2002	19F	19F	81	81	4	Mark van der Linden
257	Spain	1989	23F	23F		81		Alexander Tomasz
622	Spain	1989	23F	23F		81		Alexander Tomasz
11919	Spain		23F	23F	81	81	1	Mark van der Linden
11923	Spain		23F	23F	81	81	1	Mark van der Linden
11924	Spain		23F	23F	81	81	1	Mark van der Linden
11925	Spain		23F	23F	81	81	0.5	Mark van der Linden
11928	Spain		23F	23F	81	81	1	Mark van der Linden
11930	Spain		23F	23F	81	81	1	Mark van der Linden
11932	Spain		23F	23F	81	81	1	Mark van der Linden
11933	Spain		23F	23F	81	81	1	Mark van der Linden
11934	Spain		23F	23F	81	81	1	Mark van der Linden
11935	Spain		23F	23F	81	81	1	Mark van der Linden
11936	Spain		23F	23F	81	81	1	Mark van der Linden
11937	Spain		23F	23F	81	81	1	Mark van der Linden
Th 13	Thailand	2000	19F	19F	81	81	1	Jae Hoon Song
Th 32	Thailand	2000	23F	23F	81	81	4	Jae Hoon Song
Th 5	Thailand	2000	19F	19F	81	81	2	Jae Hoon Song
Th 104	Thailand	2001	23F	23F	81	81	2	Jae Hoon Song
BS1	Turkey	2005	23F	23F	81	81		Lesley McGee
BS26	Turkey	2005	6A	06A	81	81		Lesley McGee
DG104	Turkey	2005	23F	23F	81	SLV81		Lesley McGee
DG116	Turkey	2005	23F	23F	81	81		Lesley McGee
DG15	Turkey	2005	19F	19F	81	81		Lesley McGee

DG18	Turkey	2005	6A	06A	81	81		Lesley McGee
DG29	Turkey	2005	6A	06A	81	81		Lesley McGee
DG80	Turkey	2005	6A	06A	81	81		Lesley McGee
DG99	Turkey	2005	23F	23F	81	81		Lesley McGee
GH10	Turkey	2005	23F	23F	81	81		Lesley McGee
MA28	Turkey	2005	23F	23F	81	81		Lesley McGee
LIV 5	UK	1990	23	23F		81	2	Bruno Pichon
PLY 7	UK	1990	23	23F		81	0.5	Bruno Pichon
STO 19	UK	1990	23	23F		81	2	Bruno Pichon
STO 20	UK	1990	23	23F		81	2	Bruno Pichon
2PN05348	UK	2002	23F	23F	81	81		Bruno Pichon
3PN00020	UK	2002	23F	23F	2532	2532		Bruno Pichon
3PN00133	UK	2003	23F	23F	81	81		Bruno Pichon
3PN00734	UK	2003	23F	23F	81	81		Bruno Pichon
3PN01287	UK	2003	23F	23F	81	81		Bruno Pichon
H034800032	UK	2003	23F	23F	81	81	1	Bruno Pichon
H034960030	UK	2003	23F	23F	1623	81	1	Bruno Pichon
H035220198	UK	2003	23F	23F	81	81	1	Bruno Pichon
04-2922	UK	2004	23F	23F	81	81		Tim Mitchell
04-2937	UK	2004	23F	23F	81	SLV81		Tim Mitchell
H050340158	UK	2004	23F	23F	2508	2508		Bruno Pichon
H050300113	UK	2005	23F	23F	81	81	2	Bruno Pichon
H071440111	UK	2007	23F	23F	3456	3456	0.5	Bruno Pichon
H0752805310 1	UK	2007	19F	19F	81	81	1	Bruno Pichon
H0801601790 1	UK	2007	3	3	81	81	1	Bruno Pichon
UR 112	Uruguay	1993	23F	23F		81	2	Alexander Tomasz
UR 208	Uruguay	1994	23F	23F		81	2	Alexander Tomasz
UR 428	Uruguay	1996	23F	23F		81	2	Alexander Tomasz
SVMC 10	USA	1995	23F	23F		81	4	Alexander Tomasz
SVMC 6	USA	1995	23F	23F		81	2	Alexander Tomasz
2231	USA	1998	23F	23F	81	81	4	Lesley McGee
2311	USA	1998	23F	19F	81	81	2	Lesley McGee
2321	USA	1998	23F	23F	81	81	0.5	Lesley McGee
228	USA	1999	23F	23F	81	81		Lesley McGee
92	USA	2001	19F	19F	81	81	2	Lesley McGee
111	USA	2001	23F	23F	81	81	2	Lesley McGee
2167	USA	2002	23F	23F	81	81	2	Lesley McGee
1590	USA	2005	23F	23F	81	81	2	Lesley McGee
1731	USA	2005	19A	19A	81	2346	4	Lesley McGee
1788	USA	2005	19A	19A	2346	2346	4	Lesley McGee
1789	USA	2005	19A	19A	2346	2346	4	Lesley McGee
1842	USA	2005	19A	19A	2346	2346	4	Lesley McGee
2097	USA	2005	19A	19A	2346	2346	4	Lesley McGee
2098	USA	2005	19A	19A	2346	2346	3	Lesley McGee
2615	USA	2006	19A	19F	81	81	4	Lesley McGee
2682	USA	2006	19A	19A	81	SLV81	4	Lesley McGee
2699	USA	2006	19A	19A	81	2346	4	Lesley McGee
3267	USA	2007	19A	19A	81	2346	4	Lesley McGee
3350	USA	2007	19A	19A	81	2346	4	Lesley McGee
3355	USA	2007	19A	19A	81	2346	4	Lesley McGee

Clev 2	USA	1989	23F	23F		81		Alexander Tomasz
VII-17	USA	1998	23F	23F		81	2	Alexander Tomasz
VII-22	USA	1998	23F	23F		81	2	Alexander Tomasz
VII-26	USA	1998	23F	23F		81	2	Alexander Tomasz
VII-28	USA	1998	23F	23F		81	3	Alexander Tomasz
VIII-1	USA	1998	23F	23F		81	3	Alexander Tomasz
VIII-2	USA	1998	23F	23F		81	3	Alexander Tomasz
VIII-6	USA	1998	23F	23F		81	2	Alexander Tomasz
X-8	USA	1998	23F	23F		81	2	Alexander Tomasz
SB12	Vietnam	1997	23F	23F	81	81	0.75	Christopher Parry
SB33	Vietnam	1997	23F	23F	81	81	1	Christopher Parry
SB35	Vietnam	1997	23F	23F	81	81	0.25	Christopher Parry
SB36	Vietnam	1997	23F	23F	81	81	1	Christopher Parry
SB6	Vietnam	1997	23F	23F	81	81	2	Christopher Parry
SB2	Vietnam	1998	23F	23F	81	81	0.125	Christopher Parry
SB3	Vietnam	1999	23F	23F	81	SLV81	1	Christopher Parry
SB31	Vietnam	1999	23F	23F	81	SLV81	1.5	Christopher Parry
SB50	Vietnam	1999	23F	23F	81	81	1.5	Christopher Parry
SB18	Vietnam	2000	23F	23F	81	SLV81	1.5	Christopher Parry
VN B15	Vietnam	2000	NT	23F	81	81	2	Christopher Parry

Appendix III: EMBL accession codes

Sequence	EMBL Accession	Host strain	Description
S. pneumoniae ATCC 700669	FM211187	-	Complete genome of *S. pneumoniae* ATCC 700669
S. pneumoniae OXC141	FQ312027	-	Complete genome of *S. pneumoniae* OXC141
ICE*Spn*11930	FR671403	11930	Tn*5252*/Tn*916*-type composite ICE
ICE*Spn*11876	FR671404	11876	Tn*5252*/Tn*916*-type composite ICE
ICE*Spn*8140	FR671412	8140	ICE*St*1-type ICE
Tn*916*-type ICE	FR671413	Mal M6	Tn*916*-type ICE
Tn*916*-type ICE	FR671414	H034800032	Tn*916*-type ICE
Tn*916*-type ICE	FR671415	23771	Tn*916*-type ICE
Tn*916*-type ICE	FR671416	11930	Tn*916*-type ICE
Tn*916*-type ICE	FR671417	11928	Tn*916*-type ICE
Tn*916*-type ICE	FR671418	9409	Tn*916*-type ICE
Prophage ϕV22	FR671405	V22	Group 2b prophage
Prophage ϕ040922	FR671406	04-0922	Group 2a prophage
Prophage ϕ34117	FR671407	34117	Group 2b prophage
Prophage ϕ23782	FR671408	23782	Group 1 prophage
Prophage ϕ11865	FR671409	11865	Group 1 prophage
Prophage ϕ8140	FR671410	8140	Group 2b prophage
Prophage ϕ2167	FR671411	2167	Group 2b prophage

Appendix IV: Serotype 3 strains

Strain	Country	Year	Serotype	Inferred serotype	CC	Inferred ST	Source
07-2838	Bolivia	2007	3	3	180	180	Tim Mitchell
BHN604	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark
BHN605	Czech Rep	2000	3	3	180	180	Birgitta Henriques Nomark
BHN610	Czech Rep	2000	3	3	180	SLV180	Birgitta Henriques Nomark
BHN611	Czech Rep	2002	3	3	180	505	Birgitta Henriques Nomark
BHN596	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark
BHN602	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark
BHN594	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark
BHN593	Czech Rep	2001	3	3	180	180	Birgitta Henriques Nomark
BHN597	Czech Rep	2000	3	3	180	180	Birgitta Henriques Nomark
BHN588	Czech Rep	1998	3	3	180	180	Birgitta Henriques Nomark
BHN590	Czech Rep	2002	3	3	180	180	Birgitta Henriques Nomark
BHN595	Czech Rep	2001	3	3	180	180	Birgitta Henriques Nomark
BHN598	Czech Rep	2000	3	3	180	180	Birgitta Henriques Nomark
BHN612	Czech Rep	1999	3	3	180	180	Birgitta Henriques Nomark
BHN586	Czech Rep	1995	3	3	180	505	Birgitta Henriques Nomark
BHN607	Czech Rep	1999	3	3	180	180	Birgitta Henriques Nomark
BHN589	Czech Rep	1996	3	3	180	180	Birgitta Henriques Nomark
BHN599	Czech Rep	2001	3	3	180	180	Birgitta Henriques Nomark
BHN606	Czech Rep	2001	3	3	180	180	Birgitta Henriques Nomark
BHN591	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark
BHN587	Czech Rep	2001	3	3	180	180	Birgitta Henriques Nomark
BHN585	Czech Rep	1997	3	3	180	180	Birgitta Henriques Nomark
BHN603	Czech Rep	2002	3	3	180	SLV180	Birgitta Henriques Nomark
BHN601	Czech Rep	1995	3	3	180	180	Birgitta Henriques Nomark
BHN600	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark

BHN608	Czech Rep	2003	3	3	180	180	Birgitta Henriques Nomark
BHN592	Czech Rep	2000	3	3	180	180	Birgitta Henriques Nomark
BHN609	Czech Rep	2004	3	3	180	180	Birgitta Henriques Nomark
BHN582	England		3	3	180	180	Birgitta Henriques Nomark
OXC141	England		3	3	180	180	Tim Mitchell
03-4183	Netherlands	2003	3	3	180	180	Tim Mitchell
03-4156	Netherlands	2003	3	3	180	180	Tim Mitchell
02-1198	Scotland	2002	3	3	180	180	Tim Mitchell
99-4038	Scotland	1999	3	3	180	180	Tim Mitchell
99-4039	Scotland	1999	3	3	180	180	Tim Mitchell
BHN584	Spain		3	3	180	180	Birgitta Henriques Nomark
BHN035	Spain		3	3	180	180	Birgitta Henriques Nomark
BHN630	Sweden	2000	3	3	180	180	Birgitta Henriques Nomark
BHN626	Sweden	1999	3	3	180	180	Birgitta Henriques Nomark
BHN654	Sweden	2003	3	3	180	180	Birgitta Henriques Nomark
BHN646	Sweden	2001	3	3	180	180	Birgitta Henriques Nomark
BHN383	Sweden	2001	3	3	180	180	Birgitta Henriques Nomark
BHN576	Sweden	1994	3	3	180	180	Birgitta Henriques Nomark
BHN625	Sweden	1999	3	3	180	180	Birgitta Henriques Nomark
BHN621	Sweden	1998	3	3	180	180	Birgitta Henriques Nomark
BHN622	Sweden	1999	3	3	180	180	Birgitta Henriques Nomark
BHN652	Sweden	2002	3	3	180	180	Birgitta Henriques Nomark
BHN624	Sweden	1999	3	3	180	180	Birgitta Henriques Nomark
BHN583	Sweden	1995	3	3	180	180	Birgitta Henriques Nomark
BHN577	Sweden	1995	3	3	180	180	Birgitta Henriques Nomark
BHN578	Sweden		3	3	180	180	Birgitta Henriques Nomark
BHN037	Sweden		3	3	180	1826	Birgitta Henriques Nomark
BHN579	Sweden		3	3	180	180	Birgitta Henriques Nomark
BHN638	Sweden	1998	3	3	180	180	Birgitta Henriques Nomark
BHN628	Sweden	2000	3	3	180	180	Birgitta Henriques Nomark
BHN167	Sweden		3	3	180	180	Birgitta Henriques

BHN167	Sweden		3	3	180	180	Birgitta Henriques Nomark
BHN623	Sweden	1999	3	3	180	180	Birgitta Henriques Nomark
BHN632	Sweden	1998	3	3	180	180	Birgitta Henriques Nomark
BHN633	Sweden	2000	3	3	180	SLV180	Birgitta Henriques Nomark
BHN629	Sweden	2000	3	3	180	180	Birgitta Henriques Nomark
BHN341	Sweden	2001	3	3	180	180	Birgitta Henriques Nomark
BHN580	Sweden		3	3	180	180	Birgitta Henriques Nomark
BHN616	Sweden		3	3	180	180	Birgitta Henriques Nomark
BHN643	Sweden	1998	3	3	180	180	Birgitta Henriques Nomark
BHN657	Sweden	2006	3	3	180	180	Birgitta Henriques Nomark
BHN575	Sweden	1994	3	3	180	180	Birgitta Henriques Nomark
BHN369	Sweden	2000	3	3	180	180	Birgitta Henriques Nomark
BHN645	Sweden	2001	3	3	180	180	Birgitta Henriques Nomark
BHN642	Sweden	1998	3	3	180	180	Birgitta Henriques Nomark
BHN636	Sweden	2001	3	3	180	180	Birgitta Henriques Nomark
BHN651	Sweden	1997	3	3	180	180	Birgitta Henriques Nomark
BHN647	Sweden	2007	3	3	180	180	Birgitta Henriques Nomark
BHN644	Sweden	2000	3	3	180	180	Birgitta Henriques Nomark
BHN637	Sweden	2001	3	3	180	180	Birgitta Henriques Nomark
BHN640	Sweden	1998	3	3	180	180	Birgitta Henriques Nomark
BHN631	Sweden	2000	3	3	180	180	Birgitta Henriques Nomark
BHN650	Sweden	1997	3	3	180	180	Birgitta Henriques Nomark
BHN572	USA	1994	3	3	180	180	Birgitta Henriques Nomark
BHN413	USA	1993	3	3	180	180	Birgitta Henriques Nomark
BHN570	USA	1994	3	3	180	180	Birgitta Henriques Nomark
BHN573	USA	1994	3	3	180	180	Birgitta Henriques Nomark
BHN574	USA	1995	3	3	180	180	Birgitta Henriques Nomark
BHN571	USA	1995	3	3	180	180	Birgitta Henriques