Pneumococcal Genomics and Evolution

Andries J. van Tonder

(Nuffield Department of Medicine and St Catherine's College, University of Oxford)

Thesis submitted for examination for the degree of Doctor of Philosophy in Clinical Medicine (Trinity 2016)

Supervisor: Dr Angela B. Brueggemann

Declaration

This thesis of is submitted for the degree of Doctor of Philosophy in Clinical Medicine in Trinity Term 2016. None of the contents of this thesis have previously been submitted for any other degree.

The data herein described were generated by myself under the supervision of Dr Angela B. Brueggemann and/or by our collaborators as described in Chapter 2, *General Methods*, and summarised at the start of each Results chapter (Chapters 3 – 7). I completed all analyses under the supervision of Dr Brueggemann.

The results described in Chapter 4 were presented in-part at the 12th European Meeting On the Molecular Biology of the Pneumococcus (Europneumo 2015), Oxford 2015 and the results described in Chapter 5 were presented in-part at the 10th International Symposium on Pneumococci and Pneumococcal Diseases (ISPPD 2016), Glasgow 2016.

Chapters 3, 4 and 5 were submitted for publication in peer-reviewed journals; these manuscripts were written by myself and Dr. Brueggemann and were subsequently reviewed by each of the co-authors (as listed at the beginning of these chapters). All other sections of this thesis were written by myself.

Signed,

Andries J. van Tonder

Pneumococcal Genomics and Evolution

Andries J. van Tonder

St Catherine's College

Thesis submitted for examination for the degree of Doctor of Philosophy in Clinical Medicine (Trinity 2016)

Abstract

Streptococcus pneumoniae (the "pneumococcus") is a global pathogen responsible for nearly a million deaths each year in children under the age of five. The polysaccharide capsule or serotype is the primary pneumococcal virulence factor and is encoded by the *cps* locus. The introduction of pneumococcal conjugate vaccines has led to a significant reduction in the prevalence of invasive diseases caused by vaccine serotypes and a concomitant reduction of vaccine serotypes recovered from the nasopharynx of healthy children, although the overall rates of carriage have remained constant due to the replacement of vaccine serotypes with non-vaccine serotypes. The increasing numbers of genomes from large well-characterised pneumococcal datasets allows for key biological questions related to the pneumococcus, such as capsular diversity, to be examined in unprecedented detail. This thesis describes the application of genomics to investigating important aspects of pneumococcal biology including the pan-genome, *cps* locus diversity and genomic regions potentially associated with invasive disease. The first part of the thesis described a Bayesian decision model for estimating bacterial core genomes. This model was applied to genome datasets from five different bacterial species. The same model was later used to estimate core genomes for four different pneumococcal carriage datasets. These estimated core genomes were then compared and revealed a very small shared core genome of 303 genes. The results also highlighted the unexpected diversity of a pneumococcal dataset from Thailand. The second part of the thesis examined the diversity of the *cps* locus starting with serogroup 6 and in particular the recently described novel serotype 6E. This analysis revealed that the majority of what had previously been described as serotype 6B pneumococci were in fact serotype 6E and serological assays demonstrated that vaccine-induced serotype 6B antibodies were able to kill serotype 6E pneumococci. The analysis of *cps* loci was then extended to consider the diversity of 49 serotypes and included 5,405 genomes in the analyses. These analyses revealed several hybrid and putative novel serotypes. The final part of the thesis used a genome-wide association study to identify 89 candidate loci significantly associated with invasive disease. The results in this thesis showed that large numbers of genome sequences could provide a detailed insight into important aspects of pneumococcal biology.

Acknowledgements

The work in this thesis was supported, firstly, by the Wellcome Trust through a Career Development Fellowship awarded to Dr Angela Brueggemann, and secondly by GlaxoSmithKline Biologicals who provided funding for fees and a stipend. To start, I'd like to thank Angela for taking the chance on a young-ish research assistant and making me an offer I couldn't refuse; and then for providing guidance, support and the confidence to pursue my own research interests. Thanks must also go to my other supervisor, Dr Danny Wilson, who provided much needed advice with regards to performing genome wide association studies.

I would like to thank my collaborators in Iceland who provided the majority of the funding for my DPhil as well as arranged for funding for the VICE project; in particular the primary investigators, Prof. Karl Krisstinsson, Prof. Ásgeir Haraldsson and Prof. Helga Erlendsdóttir as well as the other members of the VICE study: Dr Martha Hjálmarsdóttir, Dr Gunnsteinn Haraldsson and Sigríður Júlía Quirk. *Takk.*

Thanks to Anne von Gottberg for providing the South African antibiotic resistant pneumococci included in the analyses in Chapters 3, 4 and 5 as well as for arranging a thoroughly interesting visit to the National Institute for Communicable Diseases (NICD) in Johannesburg, South Africa; Prof. Julian Parkill, Prof. Stephen Bentley and Dr David Harris who shepherded the thousands of DNA extractions through the Wellcome Trust Sanger Institute's sequencing pipeline; and the many staff at the Sanger Institute who made the libraries and completed the whole genome sequencing. Thanks to Prof. David Goldblatt who arranged for the opsonophagocytosis assays described in Chapter 4 to take place in his lab at the Institute of Child Health, University College London where Lucy Roalfe, Rebecca White and Marta Zancolli performed the assays.

Special thanks must go to Prof. Martin Maiden's group in the Department of Zoology who have been extremely supportive and, in particular, Dr Keith Jolley and Dr James Bray, without whom my DPhil would've been far more complicated! Indeed, the majority of the work in this thesis would have been nearly impossible without the infrastructure provided by BIGSdb.

To the former and current members of our merry band of pneumococcal researchers, Dr Kelly Wyres, Carlijn Bogaardt, Caroline "Caz" Harrold, and Reza "Ray" Rezaei Javan, thanks for the formal and informal chats and lunch time getaways to get tasty food. The last four years would have been a lot duller without you all. To Dr Shilan Mistry, our itinerant mathematician and my housemate for over two years, thanks buddy! Who knew where those chats around your blackboard would lead...

On a more personal note I'd like to mention the teachers and mentors who provided learning, guidance, the occasional kick up the bum, and whom, most importantly, saw potential: Mrs Loveridge, the late Dr Gaden Robinson and Dr Geoff Oxford.

Finally, the most important people of all, my parents Gerhard and Tracey, who took the brave decision to move my sister and I to a different country to give us the best possible opportunities, at no small cost to their own comfort and happiness. Whilst, it's not always been a smooth ride, I couldn't have done it without your love and support. Thanks for letting me choose my own path and for letting me make my own mistakes along the way.

Thank you.

Contents

Abbreviations and acronyms

BIGSdb	Bacterial Isolate Genome Sequence database
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
bp	base pair(s)
CC	Clonal Complex
CDS	Coding Sequence(s)
CLSI	Clinical Laboratory Standards Institute
CO_2	Carbon dioxide
COG(s)	Cluster of Orthologous Gene(s)
cps	capsular polysaccharide synthesis
CSF	Cerebrospinal Fluid
CSP	Competence Stimulating Peptide
DLV	Double Locus Variant
DNA	DeoxyriboNucleic Acid
DSK	Disk Streaming of K-mers
eBURST	e-Based Upon Related Sequence Types
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
ENA	European Nucleotide Archive
EUCAST	The European Committee on Antimicrobial Susceptibility Testing
goeBURST	global optimal e-Based Upon Related Sequence Types
GWAS(s)	Genome Wide Association Study(ies)

HGT	Horizontal Gene Transfer
hierBAPS	hierarchical Bayesian Analysis of Population Structure
HIV	Human Immunodeficiency Virus
ICE	Integrative Conjugative Element
IPD	Invasive Pneumococcal Disease
iTOL	interactive Tree of Life
KEGG	Kyoto Encyclopaedia of Genes and Genomes
kb	kilo base
ks density	kernel smoothing density
MA	Massachusetts
MAFFT	Multiple sequence Alignment based on the Fast Fourier Transform
MEGA5	Molecular Evolutionary Genetics Analysis 5
MEGA6	Molecular Evolutionary Genetics Analysis 6
MGE	Mobile Genetic Element
MIC	Minimum Inhibitory Concentration
Ml	millilitres
MLEE	Multi-Locus Enzyme Electrophoresis
MLST	Multi-Locus Sequence Typing
MRF	Meningitis Research Foundation
mRNA(s)	messenger Ribonucleic Acid
MRSA	Methicillin-Resistant *Staphylococcus aureus*
MSSA	Methicillin-Susceptible *Staphylococcus aureus*
MUSCLE	MUltiple Sequence Comparison by Log Expectation

n	number
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
ng	nano-gram(s)
NGS	Next-Generation Sequencing
NS	Non-Susceptible
NT	Non-Typeable
NVT	Non-Vaccine Type
OPA	Opsonophagocytosis Assay
p-distance	pairwise evolutionary distance
p-value	probability-value
PBP	Penicillin Binding Protein
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PCV	Pneumococcal Conjugate Vaccine
PFAM	Protein family
PFGE	Pulsed-Field Gel Electrophoresis
PMEN	Pneumococcal Molecular Epidemiology Network
PPV	Pneumococcal Polysaccharide Vaccine
PTS	PhosphoTransferase System
pubMLST	public Multi-Locus Sequence Type
ref	reference
rpm	revolutions per minute
rMLST	ribosomal Multi-Locus Sequence Type

rRNA	ribosomal Ribonucleic Acid
S	Susceptible
SC	Sequence Cluster
SEER	Sequence Element Enrichment
SLV	Single Locus Variant
SNP(s)	Single Nucleotide Polymorphism(s)
SSPACE	SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension
ST	Sequence Type
TMP/SMX	TriMethoPrim-SulfaMethoXazole
Tn-seq	Transposon sequencing
TSA	Tryptic Soy Agar
UK	United Kingdom
USA	United States of America
UV	UltraViolet
VICE	Vaccine Impact in Iceland
VT	Vaccine Type
WGS	Whole Genome Sequencing
µg	micro-gram(s)
µl	micro-litre(s)
σ^x	sigma factor X

1. Introduction

1.1 The pneumococcus

Streptococcus pneumoniae (the "pneumococcus") is a highly recombinant gram-positive, facultatively anaerobic human commensal and pathogen and was first independently isolated in 1881 by both George Sternberg and Louis Pasteur [1, 2]. *S. pneumoniae* is alpha-hemolytic, which means that when the bacterium is grown on blood agar, the agar underneath the colonies turns dark and greenish due to oxidation of the iron in haemoglobin [3]. In order to differentiate *S. pneumoniae* from other alpha-hemolytic streptococci (viridans streptococci), colonies need to be shown to be optochin-susceptible or bile soluble [4].

S. pneumoniae is a member of the *Streptococcus* (from the Greek *streptos* meaning twisted and *kokkos*, meaning berry or grain) genus, named for the chains formed when grown in liquid media. Originally named *Diplococcus pneumoniae* due to the very short chains generally consisting of only two cells, pneumococcus was renamed in 1974 [5, 6] as it was shown, in liquid media, to grow in the distinctive chains representative of the *Streptococci*. Broadly, the streptococci can be divided into two groups: alpha-hemolytic, beta-hemolytic (the red blood cells are completely ruptured leading to colourless regions forming around the bacterial colonies when grown on blood agar plates) and gamma-hemolytic (no hemolysis occurs). The alpha-hemolytic *Streptococci* can be further classified into five groups: Anginosus, Salivarius, Bovis, Mutans and Mitis. Based on phylogenies generated using 16S rRNA sequences [7], *S. pneumoniae* falls in the Mitis group along with other closely-related oral *Streptococci* like *S. mitis*, *S. oralis*, *S. infantis* and *S. pseudopneumoniae*. There has been some debate as to what the closest relative of *S. pneumoniae* is with studies either suggesting *S. mitis* [8] or *S. pseudopneumoniae* [9]. This is likely a result of the extensive recombination occurring between the different oral *Streptococci* leading to certain genes being more similar between one pair of species whilst other genes share similarity with genes from a different

species. More recent work suggests that *S. mitis*, *S. pneumoniae* and *S. pseudopneumoniae* are descended from a pneumococcus-like pathogen found in an early homonin ancestor [10]. The beta-hemolytic *Streptococci* include important human pathogens such as *S. pyogenes* (causes 'strep throat') and *S. agalactiae* (implicated in neonatal meningitis and sepsis). The majority of the gamma-hemolytic *Streptococcus* species were reclassified as *Enterococcus* in 1984 and are rarely implicated in disease [11].

1.2 Epidemiological characterisation of the pneumococcus

Serotyping

The gold standard for identifying pneumococcal serotypes is the Quellung reaction, first described by Fred Neufeld in 1902 [12]. Quellung is the German word for "swelling" and describes the appearance of the pneumococcal capsule after specific antibodies have combined with their polysaccharide antigens. Briefly, antisera specific to particular groups of serotypes are mixed sequentially until a positive reaction, in the form of capsular swelling, is observed under a microscope. Commercially available antisera are available for over 90 pneumococcal serotypes and the technique is in use in microbiology reference laboratories around the world. Quellung, is however, an expensive technique that requires extensive training and experience to ensure that technicians are able to correctly identify the serotype being tested.

One alternative to the Quellung reaction for serotyping pneumococci is to use the latex agglutination method [13, 14]. Pooled rabbit antisera that recognize multiple serotypes and serogroups are applied to the surface of latex beads and mixed with pneumococcal colonies suspended in phosphate-buffered saline (PBS). A positive reaction between capsule antigens and the serotype/serogroup specific antisera is shown by clumping or agglutination in the mixture of pneumococci and the latex beads. The latex agglutination method is faster than

Quellung at identifying serogroups and serotypes included in the pooled antisera and requires less training to use. However, further testing by Quellung may be required to characterise a pneumococcus to the level of serotype as, for many serogroups, latex agglutination will only identify the serogroup of the pneumococcus being assayed.

For the majority of serotypes, the polysaccharide capsule is encoded by the *cps* locus (see Capsule genetics) and in 2006, the genome sequences for 90 *cps* loci that had been identified at that time were published. This allowed multiplex PCR schemes to be developed to differentiate serotypes [15, 16]. Numerous schemes have been developed by various groups around the world and have been successfully implemented as part of national pneumococcal epidemiology monitoring programmes. More recently with the development of next-generation sequencing technologies and the widespread application of these technologies to sequencing pneumococci, a new method of serotyping based on the whole genome sequence (WGS) is growing in popularity (see Pneumococcal genomics).

Genotyping

Genotypic classification of pneumococcus has traditionally been performed using techniques such as PFGE (Pulsed Field Gel Electrophoresis) [17] and rRNA restriction (ribotyping) analysis [18]. In the past two decades, MLST (MultiLocus Sequence Typing) has become the gold standard method for distinguishing pneumococcal genotypes.

Multilocus Sequence Typing (MLST)

Originally developed and validated using *Neisseria meningitidis* [19], MLST is a genotyping method that uses the nucleotide sequences of ~500 bp regions of seven housekeeping genes to assign a sequence type (ST). Each unique nucleotide sequence identified for each of the seven genes is assigned an allele number and every unique combination of allele numbers is

used to determine a unique ST. The seven housekeeping genes used for MLST in pneumococcus are: *aroE* (shikimate kinase), *gdh* (glucose-6-phosphate), *gki* (glucose kinase), *recP* (transketolase), *spi* (signal peptidase I), *xpt* (xanthine phosphoribosyltransferase) and *ddl* (D-alanine-D-alanine ligase). Databases for a number of different bacterial species, including pneumococcus, have been created and unique allele sequences and STs are deposited following curation and quality control. Currently there are ~32,000 isolates and nearly 12,000 unique STs in the pneumococcal MLST database (August 2016; http://pubmlst.org/spneumoniae/). Before the wider availability of affordable whole genome sequencing, the nucleotide sequences were obtained by using defined primers to amplify the regions using PCR. It is now possible to search for and assign the MLST sequences within whole genome sequences using an interface such as BIGSdb [20] which is directly linked to the MLST database.

STs can be assigned to clonal complexes (CCs) using the related algorithms eBURST and goeBurst. [21, 22]. CCs are groups of related STs (generally those that share at least six of the seven MLST alleles with each other) and are considered to share descent from a common ancestor (known as the group founder). The group founder is defined as the ST with the highest number of single locus variants (SLVs; STs that differ by only one allele). Each ST that is a SLV of the group founder can also be linked to its own SLVs (double locus variants, DLVs, of the group founder) and in this way a network of related isolates can be established with each ST linked to its own SLVs. This can be represented diagrammatically using goeBURST where the STs are shown as nodes (the size of the node is determined by the number of isolates in the MLST database) and the SLVs are shown as links. In theory, if all possible pneumococcal STs were known, the entire pneumococcal population could be linked together. In practice, many of these links are missing, so a number of unrelated CCs have been defined. There is a strong association between ST and serotype and unusual

combinations suggest that genetic changes have occurred [23]. The circulating serotypes and STs of disease and carriage isolates, combined with patient and isolate metadata, can be used to reveal the underlying population structure.

1.3 Interactions between pneumococcus and its host

Carriage

Pneumococcus is a human commensal and its primary ecological niche is the nasopharynx where it is typically carried asymptomatically (a state known as carriage). Pneumococcal carriage rates are highest in young children. In developing countries 50-90% of children are colonized within the first few months of their lives [24-27] and the cumulative frequency of carriage amongst children in the first year of their lives is close to 100% meaning that nearly every child will have experienced at least one colonization event by the time they turn one [28]. Carriage rates begin to decline after the age of four and reach approximately 10% by the time children turn ten and remains at this level into adulthood. The rate of pneumococcal carriage in children is lower in the developed world and peaks at 40-50% around the age of two before declining to approximately 10% in adults [29-33].

There are three distinct phases involved in pneumococcus gaining access to and colonising the nasopharynx: acquisition, carriage and termination of carriage. Acquisition, where pneumococcus enters and establishes itself in the nasopharynx, is followed by a period of colonization or carriage which may vary in duration and is dependent on a variety of factors related to both the host and the pneumococcus. Finally, termination of carriage results when the pneumococcus is lost from the host or is otherwise undetectable [10]. The most important risk factors related to pneumococcal carriage are age of the host at the time of acquisition (see above) and whether or not a child has a sibling or else comes into close, regular contact with other children in a setting such as a day-care centre. A number of studies have shown that

children, aged between 2 and 5, that attend day care centres have an increased level of pneumococcal carriage compared to older children and adults suggesting that children of this age are a primary source of pneumococcal transmission [34-36]. The role of host genetics may also play an important part in pneumococcal carriage: Native American populations have been shown to have very high rates of carriage [37] and, in Australia, Aboriginal children have higher rates of carriage than non-Aboriginal children [38]. However, these findings may also be related to socio-economic factors as both these populations are generally poorer and more deprived and there is some evidence for socio-economic factors playing a role in the rates of carriage [39]. The evidence for other risk factors related to pneumococcal carriage is mixed but studies have shown evidence for pneumococcal colonization being promoted by viral respiratory infections [40-44] or exposure to cigarette smoke [45, 46].

The length of duration of carriage may vary from a few days to several months and duration of carriage is most associated with the serotype of the colonising strain [31, 47, 48]. Sleeman and colleagues showed, as part of a longitudinal study where children were swabbed weekly and then monthly, that the median duration of carriage for each serotype varied from 5.9 weeks for serotype 15C to nearly 20 weeks for serotype 6B [48]. Other serotypes such as 1, 12, 20, 21 and 33 have been shown to persist for shorter periods of time (up to 2 weeks) [47]. Colonisation of the nasopharynx by more than strain or serotype occurs in up to 30% of individuals [25, 49-51]. Generally, there is a single, predominant serotype and the other types are only present as minority strains [31, 52, 53]. However, estimates of multiple serotype carriage are likely to be lower than the actual prevalence as sampling techniques, such as those based on serotyping individual colonies, may miss additional serotypes especially if the pneumococci have a similar colony morphology. As techniques such as sequence-based serotyping improve (see Pneumococcal genomics) and WGS becomes

standard in routine epidemiological monitoring programmes, estimates of the prevalence of multiple carriage are likely to increase as multiple serotypes can be deduced computationally without relying on having to visually separate different serotypes based on their colony morphology.

Molecular epidemiology of carriage

The distribution of serotypes identified amongst pneumococci isolated in the nasopharynx varies depending on the country of isolation. In developed countries such as the Netherlands, Greece and France the post-PCV implementation serotype distribution is broadly similar and the same serotypes (19A, 15B/C, 11A, 15A and 6C) are frequently found [54-56] In other locations such as countries in Africa and Asia, many of which have not yet introduced any PCV, the distribution of the most frequently carried serotypes is slightly different; for instance in Nepal the most commonly carried serotypes after those from serogroup 6 were non-typeable (NT; NT pneumococci include those that fail to react with currently available typing antisera and those that fail to express a polysaccharide capsule because the capsular locus genes are disrupted or missing) , 15B/C, 34 and 21 [57] and in Gambia serotype 3 is often the most commonly carried serotype [26, 58].

Disease

The pneumococcus can cause infections in sites outside its normal nasopharyngeal niche including the inner ear (otitis media), sinuses (sinusitis), conjunctiva (conjunctivitis), alveoli (pneumonia) and pleural fluid (empyema). Most of these diseases are not usually life-threatening but have significant socio-economic costs such as treatment and time away from work due to illness. The most severe infections, which often result in death if untreated, occur when the pneumococcus breaches the blood-brain barrier and enters the cerebrospinal

fluid (CSF) which leads to meningitis or when it penetrates the alveolar wall allowing the pneumococcus to enter the bloodstream resulting in bacteraemia [59].

S. pneumoniae is responsible for approximately 14.5 million cases of disease and 825,000 deaths annually in children under the age of 5 [60]. Of these cases pneumococcal pneumoniae causes an estimated 2.5 million severe episodes and 400,000 deaths globally each year [61] with pneumococcal meningitis responsible for a further 118,000 deaths [62]. As well as being associated with invasive diseases such as bacteraemia and meningitis, *S. pneumoniae* is a leading cause of childhood otitis media [63]. Though the majority of deaths caused by pneumococcus are amongst children, the pneumococcus is also responsible for a significant disease burden in the elderly and amongst immunocompromised patients.

Several risk factors have been identified that predispose individuals to pneumococcal disease. As well as the same risk factors that increase the risk of carriage such as young age, attendance at day care centres or socio-economic status [10, 64-66], other risk factors include compromised immune status (the elderly and individuals with HIV or sickle cell anaemia) [67-71], reduced clearance of pneumococci from the airways caused by smoking and/or chronic obstructive pulmonary disease , and heavy alcohol consumption [66, 70, 72]. In particular, the high burden of pneumococcal disease in the elderly, where carriage rates are much lower than in children, is predominately due to an increased chance of progression to pneumococcal disease due to a lowered immune response [67].

Molecular epidemiology of pneumococcal disease

Many of the same serotypes that are commonly found in carriage are also commonly found in invasive disease; a pre-vaccine meta-analysis of serotype prevalence from cases of IPD in children under the age of five showed that the most frequently occurring IPD serotypes were 14, 6B, 1, 23F, 5, 19F and 6A [73]. Geographical differences in the serotypes that cause IPD

were also noted in this study with serotypes 1 and 5 being more likely to cause disease in developing countries when compared to developed countries. Certain serotypes have been shown to be associated with elevated risk in serious clinical outcomes for certain types of disease: empyema (serotypes 1, 3, 5, 7F, 8, 19A), necrotizing pneumonia (serotype 3), septic shock (serotypes 3, 19A) and meningitis (serotypes 10A, 15B, 19F, 23F) [23, 74].

Different pneumococcal serotypes can be assessed for their invasive disease potential by calculating their odds ratios (defined as the odds of a serotype causing disease compared to the odds of that serotype being found in carriage) [23]. Based on odds ratios, a study conducted in Oxford in 2003 showed that serotypes 1, 4, 14 and 18C had a significantly increased invasive disease potential [23]. A follow-up study conducted by the same group, which included a meta-analyses of seven different pneumococcal datasets, showed that the odds ratios for the most common serotypes were consistent across all the datasets with serotypes 1, 5 and 7F having the highest odds ratios [75] and thus the highest invasive disease potential. This study also found that different lineages with the same serotype had similar odds ratios indicating that the serotype is more important for determining invasiveness than the genotype (ST) [75].

Progression from carriage to disease is usually opportunistic on the part of the pneumococcus but a number of examples of disease outbreaks have been described. These are often found in locations with a higher density of human populations such as prisons or else in sites where hosts are likely to be immunocompromised such as nursing homes or hospitals [76-78]. In Africa and Asia, serotype 1 is responsible for large numbers of cases of pneumococcal meningitis [79, 80] and unencapsulated (NT) pneumococci have been shown to be associated with outbreaks of conjunctivitis [81].

1.4 Horizontal Gene Transfer (HGT)

There are three main mechanisms in bacteria through which DNA is able to be horizontally transferred: transformation (the uptake and incorporation of environmental DNA), conjugation (the transfer of DNA *via* mobile genetic elements or a pilus system) and transduction (DNA transfer is mediated by bacterial viruses referred to as bacteriophages) [82]. Of the three mechanisms, there is no evidence for conjugation in pneumococcus and the evidence for transduction is limited although, given the relative ubiquity of bacteriophages and putative conjugative elements in the pneumococcal population, conjugation and transduction may yet be shown to have an important role in the transfer of genetic material between different pneumococci [83]. Thus, the most important type of horizontal gene transfer in pneumococcus is transformation.

Transformation

The pneumococcus is competent and naturally transformable which means that it readily incorporates exogenous DNA through homologous recombination. This ability in pneumococcus was first identified by Frederick Griffith in 1928 [84] and his work would be used as the basis for the experiments performed by Avery, Macleod and McCarty which proved that DNA was the basis of heredity [85]. The competent state in pneumococci, when they are able to undergo transformation, is encoded by the *comABCDE* operon and occurs early in the mid-logarithmic growth phase lasting for approximately 40 minutes [10]. Competence in the pneumococcus is regulated by an extracellular signalling mechanism [86]; the signal is generated by the ComAB transporter which cleaves the ComC peptide as it is exported to form the competence stimulating protein (CSP), an extracellular pheromone [87, 88]. ComD and ComE, which function as a two-component signal transduction system, detect the CSP and initiate the upregulation of the early competence genes as well as

stimulating the transcription of *comX* which encodes the sigma factor X (σ^X). σ^X, is responsible for the transcription of approximately 60 further competence genes known as late genes which encode the proteins required for genetic transformation and other functions as yet poorly understood [89]. The transcription of these genes is tightly regulated and they are upregulated at the same time before transcription of these genes is terminated a few minutes after induction [89].

Included amongst the late genes are the genes that encode for a competence pseudopilus (cglC and cglD), the DNA uptake pore and a ssDNA binding protein (RecA). Exogenous DNA binds to the competence pseudopilus which then retracts pulling the bound DNA across the cell wall where the DNA interacts with the uptake pore which contains nucleases that degrade one strand of the DNA resulting in a double-stranded break [90-92]. RecA is then loaded onto the nicked DNA strand in the cytosol with the help of DprA and the resulting nucleoprotein filament is capable of invading the cell's dsDNA though ~75% of these molecules will be degraded without being incorporated into the host genome [93-95].

A study that examined six pneumococci isolated from the same patient over a period of seven months showed that multiple recombination events can occur quickly and frequently with 23 recombination events being observed between the collection of the first pneumococcus and the last [96]. The regions involved in these recombination events made up approximately 7.8% of the genome. The average size of DNA incorporated into the pneumococcal genome through recombination is 6.3 Kb with various studies having shown that these regions may vary from 3 bp up to as much as 235,000 bp [96, 97].

Although all pneumococci recombine, the frequency of recombination has been shown to vary considerably; an experimental study showed, that in carriage pneumococci, the frequency of recombination varied by up to four orders of magnitude and was not correlated

with the genetic relatedness of the pneumococci [98]. Similar variation in the rates of recombination relative to mutation (r/m) have been shown in different lineages in the same pneumococcal population [99]. There is also evidence from a large longitudinal carriage study that some pneumococci are better at donating or receiving exogenous DNA than others, with an unencapsulated (NT) lineage being shown to be more likely to both donate and gain DNA when compared to the other lineages in the population [100].

Transduction

Bacteriophages can be divided into two categories: lytic bacteriophages and lysogenic bacteriophages. Lytic bacteriophages enter a bacterial cell and immediately use the cell's machinery to replicate before lysing the bacterial cell to release infective phage particles that then go on to infect other bacterial cells. Upon entering a bacterial cell, lysogenic bacteriophages do not enter a replication cycle or lyse the bacterial cell but instead integrate themselves into the bacterial chromosome where they remain dormant until an external stimulus such as UV radiation induces the bacteriophage to enter the lytic cycle [82]. Occasionally, upon entering the lytic cycle, host DNA can be incorporated into the phage particles which leads to the transfer of bacterial DNA when the bacteriophages infect another cell and again incorporate themselves into a new bacterial chromosome. This process of transferring bacterial DNA via bacteriophages is known as transduction.

The first pneumococcal bacteriophages were identified in 1975 by two groups independently in samples isolated from the throat swabs of healthy children [101, 102]. One earlier study estimated that up to 76% of pneumococci isolated in a clinical setting contained bacteriophages or bacteriophage remnants in their genomes [103]. Pneumococcal bacteriophages are a relatively understudied area of pneumococcal biology and it is largely unknown what role bacteriophage sequences might play in pneumococcal virulence or

pathogenesis but the presence of the MM-1 phage in a pneumococcal chromosome has been shown to increase adherence to pharyngeal surfaces [104]. The high frequency of bacteriophage sequences in pneumococcal genomes is, however, suggestive of a possible role in enhancing genetic diversity as well as improving the ability of pneumococcus to evade selection pressures [83].

1.5 The pneumococcal polysaccharide capsule

Pneumococcus, like many bacterial species, produces an extracellular polysaccharide capsule which is covalently attached to the peptidoglycan cell wall [105]. The polysaccharide capsule is the primary pneumococcal virulence factor and is believed to aid bacterial survival in blood [106] as encapsulated pneumococci were shown to have increased bacterial survival during IPD when compared to unencapsulated pneumococci. In order to evade the host immune system, antigenic diversity has been selected for leading to the observed diversity of capsular types. Nearly 100 different serotypes are now known and new types are still being identified [107-110]. The classification of pneumococcal capsules into distinct serotypes is based on the reaction of the capsule to serotype-specific antisera (see The epidemiological typing of pneumococci section) and the majority of serotypes were first defined in the 1930s [111]. The current system of nomenclature for serotypes was developed in the 1940s in Denmark [112] and the nearly 100 different serotypes are grouped into 46 serogroups based on the cross-reactivity of antisera to the respective polysaccharide capsules [113].

Capsule synthesis

With the exception of serotypes 3 and 37, all capsule types are synthesised through the Wzy-dependent pathway [15]. The monosaccharide repeat unit is synthesized through sequential addition of nucleotide-charged sugars to an undecaprenyl phosphate acceptor on the inner leaflet of the cell membrane. The complete repeat unit is then transferred to the extracellular

face of the membrane by a flippase (encoded by the *wzx* gene) where the Wzy polymerase adds the growing polymer chain to the new repeat unit before attachment to the peptidoglycan cell wall [114]. Capsular synthesis in serotypes 3 and 37 requires the synthase-dependent pathway. The synthase gene (*wchE*) in serotype 3 is located in a truncated cps locus that consists of only six genes, five complete and one truncated [115-117]. The *cps* locus in serotype 37 is defective and closely resembles the cps locus of serotype 33F. Instead the serotype 37 capsule is produced by the synthase gene *tts*, which encodes a protein with sequence motifs typical of cellulose synthases and β-glucosyltransferases, and is located in a different part of the genome [118, 119].

The capsule is a significant metabolic burden to the cell as it may account for more than 50% of the pneumococcal volume [120]. There are two different phases of capsule production in pneumococcus: the transparent and opaque phases. The two phases are characterised by the differences between pneumococcal colonies that are observable when the colonies are viewed on transparent surfaces instead of blood agar [121]. Transparent variants, where expression of the capsule is reduced, have been shown to have increase adherence to human epithelial cells. In animal models of carriage, transparent phase variants were selected for during nasopharyngeal carriage and were responsible for lower mortality rates as compared to opaque variants introduced intraperitoneally in a mouse model [121, 122]. Opaque variants, which express 2.5 to 22 times more polysaccharide capsule when compared to transparent variants of the same strain, are associated with invasive disease and are not generally found in nasopharyngeal colonisation [123]. The increased polysaccharide expression by opaque variants hinders cytoadherance, which may explain their inefficiency at colonising the nasopharynx. Increased polysaccharide expression also increases resistance to opsonophagocytosis and has been implicated in increased virulence in murine models of sepsis [121, 124]. More recent work has shown that phase variation in pneumococcus is

regulated by genetic rearrangements in the Type 1 restriction modification system [125]. These rearrangements resulted in six different subpopulations with different gene expression, variable patterns of methylation and which exhibited phenotypic changes including phase variation [125].

Capsule genetics

The genes encoding the Wzy-dependent pathway are located in a region of the pneumococcal chromosome known as the *cps* locus. The *cps* locus is situated between *dexB* and *aliA* and varies in size from ~10 kb to ~30 kb with the number of genes varying accordingly [15]. In almost all serotypes, the *cps* genes are in the same orientation as *dexB* and *aliA* and are predicted to be transcribed as a single operon [126, 127]. A promoter sequence is located upstream of the first gene, *wzg*, and a stem-loop termination structure is located downstream of the final gene in the locus [126]. Typically, the first four genes in the *cps* locus are conserved and common to nearly all serotypes; these four genes *wzg*, *wzh*, *wzd* and *wze* are involved in the regulation and processing of the capsule. The next gene in the *cps* locus is normally *wchA*, the initial glucose phosphate transferase responsible for linking the activated glucose phosphate to the membrane-bound lipid carrier. The serotype-specific genes are located downstream of the regulatory and processing genes along with the polysaccharide polymerase (*wzy*) and flippase genes (*wzx*). Clustering of the original 90 cps loci sequences into homology groups revealed that there is a good correlation between the serogroups grouped together based on shared reactions to antisera and clustering based on the nucleotide sequence [15]. However, the analysis did reveal shared ancestry between certain serotypes, based on nucleotide sequence, that do not share any serological cross-reactivity: serotypes 44 and 46 were closely related to serogroup 12, and serotype 14 was extremely similar to serogroup 15 [15].

The observed antigenic diversity seen amongst serotypes is due to different mechanisms. For instance, small changes at a genetic level can be responsible for the production of different serotypes within the same serogroup. Within serogroup 6, a single amino acid change in the *wciP* gene changes the serotype from 6A to 6B [128] whilst the insertion or deletion of a single gene in serogroup 18 leads to the different serotypes 18A and 18B [15]. Conversely, >5% divergence at a nucleotide level as in the case of serotype 6B and 6Bii (6E) does not result in a change to the polysaccharide structure [129]. The presence or absence of O-acetylation of the peptidoglycan differentiates serotypes 15B and 15C with the differences caused by variable numbers of TA tandem repeats at the 5' end of *wciZ* which cause a frameshift in 15C but not 15B [130].

1.6 Pneumococcal vaccines

The first attempt at developing and administering a pneumococcal vaccine took place in 1911 in South Africa where it was given to workers on the diamond and gold fields who suffered from a high burden of morbidity and mortality due to pneumococcal pneumonia [131, 132]. The vaccine in this first trial, which consisted of dead pneumococci (a whole cell vaccine), reduced the rate of pneumonia in the miners receiving it in the first four months after administration but this effect declined over time. Development and testing of whole cell vaccines continued sporadically until the 1940s without any great success in the reduction of pneumococcal disease [28].

Pneumococcal Polysaccharide Vaccines (PPVs)

In 1917 it was shown that the pneumococcal capsule triggered an adaptive immune response during human disease which led to the discovery that the injection of a polysaccharide intradermally would also lead to an immune response [133, 134]. This work would eventually lead to the development of the first polysaccharide-based vaccine, containing type

1 and type 2 polysaccharides, in the 1930s in America [135]. The first polysaccharide vaccine shown to have a significant effect on vaccine-type pneumococcal pneumonia was tested on American airmen during the Second World War. This tetravalent vaccine, made up of capsule types 1, 2, 5 and 7, reduced the incidence of pneumococcal pneumonia in the target group by 85% and was also shown to provide protection against nasopharyngeal acquisition of the capsular types included in the vaccine [136]. This successful trial would lead to the commercial release of two hexavalent vaccines, one for adults (serotypes 1, 2, 3, 5, 7 and 8) and one for children (serotypes 1, 4, 6, 14, 18 and 19) in the United States in 1946. However, the accelerated development of penicillin to treat wounded servicemen during the Second World War, as well as its remarkable efficacy at treating bacterial infections, including those caused by pneumococcus, meant that interest in preventing pneumococcal disease by vaccination diminished and the hexavalent vaccines were withdrawn in 1954 [137].

Following a revived interested in pneumococcal vaccination in the late 1960s, new vaccine formulations began to be developed. A trial of a 13-valent vaccine led to a reduction in the burden of vaccine-type pneumococcal pneumonia of almost 80% amongst ~12,000 South African miners [138]. In 1977, a 14-valent vaccine was licenced for use in the USA and this was subsequently replaced by a 23-valent pneumococcal polysaccharide vaccine (PPV-23) containing serotypes 1, 2, 3, 4, 5, 6B, 7F, 8, 9N, 9V, 10A, 11A, 12F, 14, 15B, 17F, 18C, 19F, 19A, 20, 22F, 23F and 33F, which was introduced in 1983 and is still in use today [139, 140]. The introduction of PPV-23 led to a 60%-70% reduction in pneumococcal disease in immunocompetent adults though disease incidence amongst children was largely unaffected as PPV-23 was later shown to be ineffective in children under the age of two as their immature immune system is unable to mount a strong immune response against the polysaccharides included in the vaccine [141, 142]. For this reason, PPV-23 is currently only

recommended for use in adult groups at risk of serious pneumococcal disease such as the elderly, the immunocompromised or else people suffering from chronic cardiac, pulmonary or renal disease [143]. The ineffectiveness of polysaccharide vaccines amongst children under the age of five, the primary reservoir of pneumococcus, would lead to the development of conjugate vaccines which are able to elicit an immune response in children [28].

Pneumococcal Conjugate Vaccines (PCVs)

Whilst the first attempt at conjugating a pneumococcal polysaccharide to a protein (serotype 3 polysaccharide was conjugated to a horse serum globulin and used to immunise rabbits against experimental pneumococcal infection) was made in the 1930s [144, 145], the first application of the technique to a human vaccine would not take place until the end of the 20th century [146]. The first pneumococcal conjugate vaccine, PCV7 (Prevnar), was developed in the 1990's and first licenced for use in infants in America in 2000 [147]. PCV7 is made up of seven capsular polysaccharides (serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F), chosen for their high incidence in invasive disease in the USA, covalently bound to the diphtheria toxoid CRM197, which is highly immunogenic but non-toxic [146]. The combination leads to an anamestic immune response in children under the age of five [147]. More recently two other PCVs have been licenced for use: PCV10 (Synflorix, serotypes 1, 4, 5, 6B, 7F, 9V, 14, 18C, 19F, and 23F) and PCV13 (Prevnar 13, serotypes 1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, 23F) [148, 149] and these began to replace PCV7 in national immunisation programmes in 2009. As of March 2016, pneumococcal conjugate vaccines (PCV10 and PCV13) are now routinely administered in 135 countries around the world and have had a major effect on the burden of morbidity and mortality due to pneumococcal disease [150]. The next PCV, PCV15 which will include serotypes 22F and 33F (these were chosen as they accounted for ~10% of IPD cases in the USA in 2007 [151]) as well as the thirteen serotypes included in

PCV13 is currently undergoing clinical trials and will likely be licenced for use in the near future [152, 153].

The effect of PCVs on the pneumococcal population

The introduction of PCV7 into the United States led to a decline of 60%-70% in the incidence of IPD in young children given the vaccine, along with a small decrease of 6-7% in cases of acute otitis media caused by pneumococcus [154-158]. A decline was also seen in the prevalence of IPD amongst adults and unvaccinated children suggesting that the vaccination of children under the age of five produced a level of herd immunity that provided protection amongst the unvaccinated population [159]. Though the overall rates of pneumococcal carriage did not decrease post-vaccination, there was a significant decline in the carriage of vaccine-type (VT) pneumococci [160-162] as well as a reduction in the carriage of antibiotic-resistant pneumococci (many of the major antibiotic-resistant clones have vaccine serotypes [156]). The sustained rates of carriage were caused by VTs being replaced by non-vaccine types (NVTs) in the population, a phenomenon known as serotype replacement [163, 164]. Following the initial decrease in cases of IPD after the introduction of PCV7, serotype replacement was responsible for a subsequent increase in the cases of invasive disease in the USA and elsewhere caused by NVTs such as serotype 19A [165-167]. The increase in IPD caused by serotype 19A can also partly be explained by a second phenomenon, that of capsular switching [168]. Here, VT pneumococci acquire a *cps* locus that codes for a NVT capsule through recombination. This allows a disease-causing clone to "escape" the vaccine and thus continue to cause disease [168]. In the USA, it has been shown that there were at least five independent capsule switching events that led to the dissemination of serotype 19A variants of the previously serotype 4 ST695 clone that subsequently became the third most common cause of IPD due to serotype 19A [169, 170]. The replacement of PCV7 by PCV10 and PCV13 in vaccination programmes has also had a significant effect on the incidence of

IPD caused by VT pneumococci, mainly due to reductions in serotype 19A and 7F disease [171].

As pneumococcus is highly recombinant, there is a high likelihood of capsular switches occurring in the future which would lead to a potential reduction in vaccine efficacy. Thus, continued epidemiological surveillance of circulating pneumococcal serotypes is important as the long-term effects of the PCVs on the pneumococcal population need to be assessed, especially with regards to changes in the prevalence of NVT pneumococci.

1.7 The mechanisms and emergence of antibiotic resistance in pneumococcus

β-lactam resistance

β-lactams are a broad-spectrum class of antimicrobial agents comprised of different structural classes including the penicillins (*e.g.* penicillin and amoxicillin) and the cephalosporins (*e.g.* cefotaxime and ceftriaxone). β-lactam resistance in pneumococcus is mediated by alterations to the penicillin-binding proteins (PBPs) which reduce the affinity of PBPs for β-lactams and prevent the binding of the β-lactam molecules [172]. PBPs are essential in pneumococcus for cell division, growth and remodelling [173, 174] and six PBPs have been described: PBP1A, PBP1B, PBP2A, PBP2B, PBP2X and PBP3 [175]. Whilst variants of all six PBPs have been shown to be associated with β–lactam resistance in various settings, the primary resistance determinants are PBP2X and PBP2B as changes to just these two genes are capable of conferring resistance [176]. High level resistance in resistant clinical pneumococci is conferred by PBP1A, provided that the pneumococci also have altered PBP2B and/or PBP2X genes [177]. Though PBP1B and PBP3 have been implicated in resistance in laboratory mutants, the evidence for the role these genes may play in resistance in the wider pneumococcal population is limited [178-180]. Alterations to the PBPs are the result of homologous recombination whereby fragments of these genes from other closely-related

streptococci such as *S. mitis* are incorporated into the pneumococcal genome leading to a mosaic structure in the PBPs [181, 182]. The diversity of PBPs between different pneumococcal lineages suggests that there have been multiple independent acquisitions of resistant variants [183, 184]. As well as PBPs, mutations in other protein-encoding genes, including the *MurMN* and *ciaRH* operons and the *clpL* locus have also been shown to be responsible for β –lactam resistance [185-188].

The widespread adoption of penicillin in the late 1940s led to a dramatic improvement in the survival rates of individuals suffering from IPD and the pneumococcus was considered to be fully susceptible to penicillin. The first penicillin non-susceptible pneumococcus was reported in Australia in 1967 by Hansman and colleagues [189]. Further examples of penicillin resistant pneumococci were isolated in Papua New Guinea in 1971 by the same group [190]. In 1978 the first fully penicillin-resistant pneumococci were isolated from children in hospitals in South Africa [191]. These pneumococci were also resistant to tetracycline, chloramphenicol, macrolides and trimethoprim-sulfamethoxazole (TMP-SMX) and thus marked the emergence of multi-drug resistant (MDR) pneumococci. MDR pneumococcal lineages also began to emerge in Spain in the 1980s and would spread around the world such that by the time the first PCV was introduced in 2000, a significant proportion of all IPD pneumococci were resistant to penicillin and other antibiotics.

Resistance to other antibiotics

Resistance to tetracycline in pneumococcus is mediated by the acquisition of *tet(M)* and occasionally its paralogue *tet(O)* which share ~77% amino acid similarity [192-194]. The acquired *tet(M)* and *tet(O)* genes prevent the binding of tetracycline to the 30S ribosomal subunit and are commonly carried on the Tn*916* Integrative Conjugative Element (ICE). The first tetracycline resistant pneumococci were identified in Australia in 1963 [195] and current

rates of tetracycline resistance in pneumococcus vary significantly depending on location, with rates of up to 85% observed in pneumococci carried by healthy children in South Korea [196].

Chloramphenicol works by targeting peptidyl transferase, an enzyme that is responsible for catalysing peptide bond formation during bacterial translation [197, 198]. Chloramphenicol resistance is due to an enzyme encoded by the chloramphenicol acetyltransferase gene, *cat*, which inactivates chloramphenicol upon entry to the cell by catalysing the conversion of chloramphenicol to its various derivatives [199, 200]. The *cat* gene is carried by the Tn5253 and Tn5253-like ICEs [200, 201]. Chloramphenicol resistance in pneumococcus took longer to develop than resistance to other antimicrobials, but the first chloramphenicol resistant pneumococci were collected in France in 1973 [199]. Contemporary rates of chloramphenicol resistance are lower than that for other antimicrobials but may still reach as high as 67% in some populations [202].

Macrolide resistance in pneumococcus is caused by at least two different mechanisms. Erythromycin resistance is mediated by the erythromycin ribosome methylase, erm(B), which is generally acquired through the import of ICEs such as Tn1545 and Tn917 and modifies the rRNA that macrolides target [203-205]. Macrolide resistance is also caused by the mef efflux pump; two variants with >90% amino acid sequence similarity, *mefA* and *mefE* are found in pneumococcus [206-208]. Like the *cat* and *erm(B)* genes *mefA* (Tn1207.1 and Tn1207.2) and *mefE* (mega element) are typically carried on transposon-related elements [207, 209, 210]. The first erythromycin-resistant pneumococci were reported in 1967, the same year the first penicillin non-susceptible pneumococcus was reported [211], and erythromycin-resistant pneumococci rapidly increased in prevalence in the USA, South Africa and parts of Asia from the 1980s onwards [206, 212]. The modification to the 23S rRNA mediated by erm(B) may also lead to reduced binding of the structurally distinct

lincosamide class of antibiotics (*e.g.* clindamycin) which target the same sites as macrolides [213]. Clindamycin resistance in pneumococcus is a particular problem in Asia *e.g.* 92.4% of clinical pneumococci were reported to be resistant to clindamycin in China [218].

Nucleotide changes in dihydrofolate synthase (DHFS) and dihydrofolate reductase (DHFR) cause a reduced affinity for sulphonamides, such as sulfamethoxazole, and trimethoprim respectively [214-216] and mutations in the the rpoB gene, which encodes a subunit of RNA polymerase, lead to rifampicin (rifampin) resistance [217, 218]. Sulfamethoxazole and trimethoprim are usually administered together (TMP/SMX) and the prevalence of pneumococci resistant to this combination has increased from less than 10% in the 1980s to as much as 86.1% in China [212, 219]. Rifampicin, which is only used to treat pneumococcal disease in certain settings, is still an effective antimicrobial as levels of resistance are comparatively low [220, 221].

1.8 Pneumococcal genomics

The first bacterial genome to be sequenced, *Haemophilus influenzae* Rd, was made publicly available in 1995 [222] and the first sequenced pneumococcal genome, TIGR4, was published in 2001 [223]. Since then, with the development of high-throughput sequencing technologies such as Illumina, the number of pneumococcal genomes has increased significantly: over 23,000 WGS now exist in the public domain (August 2016).

The aim of the original pneumococcal sequencing projects was to provide high quality reference genomes through a process known as finishing. Attempts were made to close all the gaps in the original assembly using PCR or transposon insert libraries (TILs) and improve the quality of each base position in the genome to above a certain threshold [224]. This was a manual, time consuming and expensive process that often took years to finish a single genome to the required standard. The exponential increase in the number of pneumococci

being sequenced has meant that the majority of genomes today are draft genomes with gaps in the sequence. Often, the only attempts that have been made to improve the quality of the genomes have been to use computational algorithms to map the original sequence reads back to the initial assembly (see General methods). The new generation of sequencing technologies such as PacBio now offer the ability to sequence pneumococci using a methodology that allows for the creation of gap-free draft genomes.

Sequencing technologies

The first pneumococcal genomes were sequenced using a capillary-based, semi-automatic implementation of the Sanger method (the same chemistry was used to sequence the first DNA genome, ϕX174 in 1977 [225]). In 2005, the first next-generation sequencing (NGS) platform, the 454 system, became commercially available [226]. The first generation of the 454 platform could output data equivalent to that produced by 50 capillary sequencing machines at a fraction of the cost, immediately increasing the numbers of bacterial genomes sequenced [227]. Today, the vast majority of bacterial genome sequencing is done using various forms of the Illumina platform which was first made publicly available in 2008 [228] (see General Methods for a more detailed description of the Illumina chemistry). Multiplexing of samples (Illumina libraries for each DNA extract are tagged with a unique barcode) makes it possible to sequence up to 1,344 bacterial genomes on a single Illumina flowcell in a process that takes around 9 days. This has vastly increased the number of bacterial genomes that have been sequenced in the last decade. New sequencing technologies continue to become available and these include single-molecule long-read platforms such as PacBio and real-time sequencers like the Oxford Nanopore MiniION. The ability of the PacBio chemistry to output sequences with read lengths of up to 50 Kb (average read lengths of 10-15 Kb) has led to gap-free bacterial assemblies, although the overall confidence in the

calling of each base is slightly lower than that obtained using the Illumina platform [229, 230].

The application of genomics to pneumococcal epidemiology

As the sequencing of pneumococci becomes increasingly part of routine epidemiological monitoring, the methods for phenotyping and genotyping pneumococci, that were previously performed experimentally, will begin to be replaced with computational analyses that can be performed at a fraction of the cost and on a far shorter timescale. For example, MLST locus allele numbers can easily be assigned based on the WGS of a pneumococcus (see the Multilocus Sequence Typing (MLST) section above).

In a similar way, serotypes can now be assigned to pneumococci based on the sequence of the cps loci; here, the genomes of sequenced pneumococci are compared against a set of *cps* locus references and in this way the serotype can be assigned based on the best match in the set of references. Beyond the costs associated with sequencing and computational infrastructure, WGS-based serotyping costs very little and is significantly faster than experimental methods. Public Health England and the CDC have begun to move away from traditional methods of serotyping such as Quellung or PCR, and have begun implementing WGS-based serotyping as part of their routine epidemiological monitoring programmes.

There are various experimental protocols and methodologies used for assessing antimicrobial resistance in pneumococcus. Typically, the level of resistance of a pneumococcal strain to a particular antimicrobial agent is based on the minimum inhibitory concentration (MIC) value. The MIC is defined as the minimum concentration of an antimicrobial that will inhibit visible bacterial growth after a period of incubation, usually overnight. Given the role of altered PBPs in β-lactam resistance, there is a correlation between β-lactam MIC and the amino acid sequence of PBP1a, PBP2b and PBP2x [231]. A recent study proposed a classification

system for PBP-types that links the amino acid residues of the transpeptidase domains (responsible for cross-linking peptidoglycan chains to form rigid cell walls) of PBP1a, PBP2b and PBP2x to β-lactam MIC levels identified among invasive pneumococci [232].

One potential application of this typing system would be to predict β-lactam resistance based on pneumococcal WGS: β-lactam resistance could still be discerned for clinical samples which could not be cultured and, given the large volume of whole genome data being generated that may not have metadata such as MIC values, theoretically every pneumococcus that has been sequenced could be assessed for β-lactam resistance. The use of this typing system, once fully validated, also has significant implications for cost, time saved and antibiotic usage. If pneumococci isolated from clinically-relevant sites are being routinely sequenced as part of epidemiological monitoring programmes, the ability to quickly and easily ascertain the presence or absence of β-lactam resistance will guide physicians when deciding which antimicrobials to administer. In theory, it should also be possible to predict resistance profiles for other antimicrobials like chloramphenicol, tetracycline and erythromycin, based on the presence or absence of the genes that have been shown to confer resistance. However, to date no such typing methods have been developed but it is likely that such methods will begin to become available in the near future.

Pneumococcal population genomics

A bacterial species can be defined by its pan-genome, which consists of a core genome conventionally defined as those genes present in all isolates, and an accessory or dispensable genome, which includes the genes absent from one or more genomes or unique to a given genome [233]. The size of the core genome is highly dependent on the number and diversity of genomes included in the analysis; as more genomes are added the core genome will continue to decrease in size. Conversely, as more genomes are added the size of the pan-

genome will continue to increase as rare and unique genes not found in all the genomes increase the total number of genes in a given population. Any collection of genomes is a subset of the total population; if the subset has limited diversity then number of core genes will be higher than in a dataset which is more diverse. Various methodologies have been developed for calculating the core and accessory genomes of bacterial species (see Chapter 3).

The first pan-genome analysis to study pneumococcal diversity was applied to a set of 17 genomes; the study revealed a pan-genome of 3,170 genes with a core genome, defined as the genes found in all 17 genomes, of 1,454 genes [234]. Using their methodology, the authors predicted that the total pneumococcal pan-genome was finite and would contain approximately 5,000 genes. Subsequent studies have shown conclusively that the pneumococcal pan-genome is in point of fact open *i.e.* including new genomes in the analysis will continue to uncover novel genes [99, 235]. Calculating the pneumococcal core genome has become the starting point for many comparative genomics analyses and published estimates for the size of the pneumococcal genome range from 851 to 1,194 depending on the methodology used, and the number and relatedness of the genomes included [99, 236].

Genome Wide Association Studies (GWAS)

Genome-wide association studies (GWAS) were developed to test large numbers of genetic variants, such as single nucleotide polymorphisms (SNPs), k-mers (regular sized genome fragments) or indels, within a population of individuals for statistically significant associations with a given phenotype. The technique was originally developed to search for variants associated with disease in human genomes but is now being extended to consider genome variants found in bacteria. Since the first successful GWAS in humans was published in 2005 [237], the results of hundreds of GWASs have been published. The results

of the first bacterial GWAS were published in 2013 and revealed genetic factors responsible for host adaptation in 192 *Campylobacter jejuni* and *C. coli* WGS [238]. The majority of early bacterial GWASs adapted software designed for eukaryotic GWAS or else used custom tools to perform the analyses. However, tools designed for eukaryotes may not provide sufficient resolution or statistical robustness as bacterial populations tend to display genome-wide linkage disequilibrium and strongly structured genetic lineages (see Chapter 7) [239]. More recently, software specifically designed for bacterial GWAS that explicitly takes bacterial population structure into account has become publically available and the first pneumococcal GWAS, which identified genetic variants associated with resistance to penicillin and other antibiotics, was published in 2014 [240].

1.9 Pneumococcus in Iceland

Iceland is a unique location for pneumococcal research as it is an island with a small, ethnically-homogenous population of ~327,000 (http://www.statice.is/statistics/population/) where patient demographic and microbiological data are comprehensive and have been routinely collected for decades. The capital and largest city, Reykjavik, is home to approximately 120,000 people. If the surrounding suburbs are included the total population of the Capital Region is 200,000 people which accounts for nearly two thirds of the overall Icelandic population. Over 90% of children aged between 18 months and six years of age attend day care centres in Reykjavik and the surrounding area and each year nasopharyngeal swabs are collected from 400-500 children as part of the routine epidemiological monitoring programme. From these samples, the rates of *S. pneumoniae* and *Haemophilus influenzae* carriage are determined. The results of these studies have shown that, in the pre-vaccine era, the carriage rate of the pneumococcus was comparatively high (58% to 72% for the years 2009-2011) when compared to the carriage rates reported in other European countries [241-243]. Antimicrobial susceptibility testing of these samples revealed a high rate of resistance,

especially amongst the youngest children and those currently receiving antimicrobials [243, 244]. As well as monitoring pneumococcal carriage rates amongst children, nearly all pneumococci isolated from invasive infections by the Department of Clinical Microbiology laboratory at the Landspitali University Hospital in Reykjavik have been stored since 1990.

PVC10 was introduced in Iceland in April 2011 and in order to examine the effects of PCV10 on the Icelandic pneumococcal population, the VICE (Vaccine Impact in Iceland) study was started in 2012. The aims of the study were: i) assess the impact of PCV10 on the nasopharyngeal carriage of *S. pneumoniae* and *Haemophilus influenzae* in healthy children attending day care centres; ii) compare the prevalence of serotypes and genotypes from a selection of pneumococci isolated from blood, cerebrospinal fluid, joint fluid, the lower respiratory tract and middle ear fluid pre- and post-vaccine; iii) assess the impact of PCV10 on pneumococci isolated from the middle ear fluid of children with otitis media; and iv) assess changes in antibiotic usage and healthcare costs after the introduction of PCV10. The years included in the study were 2009 to 2014 so as to cover both the pre- and post-vaccine periods. All pneumococci isolated from invasive infections, every other pneumococcus collected from healthy children attending day care centres in Reykjavik and the surrounding areas, as well every other pneumococcus collected from the lower respiratory tract and middle ear fluid during the study period were selected for whole genome sequencing. These pneumococcal samples were sent to Oxford, where I performed the DNA extractions and subsequently sent the DNA extracts to the Wellcome Trust Sanger Institute for sequencing.

The VICE study dataset is the first large well-characterised genome dataset to include both carriage and disease (invasive and non-invasive) pneumococci and includes 1,916 pre- and post-vaccine genomes, in addition to 526 pneumococcal genomes selected to address specific questions. The traditional epidemiology, vaccine impact and molecular epidemiology analyses in the VICE study are being performed by an Icelandic PhD student, in collaboration

with us. I used the 1,916 genomes from the invasive and carriage pneumococci to address specific questions related to capsular sequence diversity, the pneumococcal pan-genome and genes associated with invasive pneumococcal disease.

1.10 Aim of this Thesis

The aim of this thesis was to show how large numbers of well-sampled, well-characterised genome sequences could be used to explore various aspects of pneumococcal biology and evolution. Chapter 3 outlines the development of a Bayesian decision model for estimating the core genome of a set of bacterial genomes. This model was used to calculate the estimated core genomes for five different bacterial species (n=2,096 genomes in total) and case studies demonstrated how the model could be used to identify areas for further study. In Chapter 4, the prevalence and distribution of the recently described serotype 6E as well as the other members of serogroup 6 was assessed using a diverse dataset of ~1000 genomes. Opsonophagocytosis assays were then conducted by collaborators in London to see whether PCV-induced serotype 6A and 6B antibodies inhibited serotype 6E pneumococci. The finding that there was significant diversity amongst serogroup 6 pneumococci in Chapter 4 led to the analyses conducted in Chapter 5 where capsular diversity amongst 49 serotypes from 14 different serogroups was assessed. A number of hybrid and putative serotypes were identified and characterised. The core genome model described in Chapter 3 was then used in Chapter 6 to calculate the estimated core genomes for four distinct pneumococcal carriage datasets. The core genomes were compared to each other which revealed a small shared core genome of 303 genes. The comparison also revealed that one of the datasets from a refugee camp in Thailand was very different to the other three datasets and possible explanations for these differences were explored. In Chapter 7 a genome-wide association study was conducted to attempt to identify genes associated with three disease phenotypes (invasive disease, otitis media and pneumonia) using genomes from the VICE study.

2. General Methods

This chapter describes the methods and materials more generally used in this thesis. Any methods specific to a particular results chapter are described in the relevant chapter.

2.1 Pneumococcal strain collections

Historical pneumococci

A unique collection of pneumococci dating back to the 1930s were collected by the Staten Serum Institute in Denmark and were made available by a collaborator, Steen Hoffman, for DNA extraction and whole genome sequencing (WGS). Sequences for the original 90 cps loci were obtained from this collection of pneumococci [15].

PMEN clones

The Pneumococcal Molecular Epidemiology Network (PMEN) was set up in 1997 in order to monitor the prevalence of antimicrobial-resistant pneumococci around the world as well as establish a standard nomenclature for globally circulating resistant pneumococcal lineages or clones [245]. Currently, 45 susceptible and non-susceptible PMEN clones have been designated and examples of each of these lineages were obtained for DNA extraction and WGS from the German Streptococcal Reference Laboratory.

South Africa

High-level penicillin resistant strains first emerged in South Africa in the late 1970s [191]. DNA was extracted from 83 penicillin non-susceptible pneumococci (dating from 1977-1989) provided by a South African collaborator, Anne von Gottberg, and sent for whole genome sequencing (WGS).

VICE study

Nasopharyngeal swabs were collected from healthy children attending 15 day care centres in the Greater Reykjavik area during April and May of the 2009-2014 period of the VICE study and cultured for pneumococci. Pure cultures were stored at -80^{o}C and every other pneumococcus was selected for DNA extraction and WGS.

All invasive pneumococci isolated from sterile sites (blood, cerebrospinal fluid, pleural fluid and joint fluid) and thus associated with invasive pneumococcal disease (IPD) have been collected and stored in Iceland since 1990 as part of national routine disease monitoring. All invasive pneumococci collected between 2009 and 2014 were selected for DNA extraction and WGS. Pneumococcal pneumonia isolates were those recovered from lung aspirates or sputum samples taken from patients with suspected pneumonia. Pneumococci associated with otitis media were cultured from swabs of discharging middle ear fluid from patients with suspected infections. All cases of suspected pneumonia and otitis media are sent to the Clinical Microbiology Laboratory at the Landspitali University Hospital in Reykjavik and approximately half of the pneumococcal-positive cultures obtained from suspected pneumonia and otitis media cases during the study period were selected for DNA extraction and WGS.

Previously sequenced genomes

A BIGS database was set up for the purposes of storing and analysing the pneumococcal genomes analysed in this thesis [20]. The sequence data were then uploaded to the database and the ribosomal, MLST and penicillin binding genes were tagged in an iterative process using BLASTN and TBLASTX [246]. Publically available reference genomes and other genomes from Genbank were downloaded pre-assembled from Genbank and uploaded to the BIGSdb database.

rMLST

Raw sequence reads for genomes sequenced as part of large population-level datasets (Massachusetts, Maela and Southampton [99, 100, 247]) were downloaded from the European Nucleotide Archive [248] and assembled using Velvet [249] before the final assemblies were uploaded to the rMLST database (pubmlst.org/rmlst) [20, 250]. Currently, the rMLST database contains over 150,000 bacterial genomes including ~23,000 pneumococcal genomes (August 2016).

2.2 Antimicrobial susceptibilities

VICE study

The VICE study pneumococci were tested for their susceptibility to the following antimicrobials: penicillin, erythromycin, tetracycline, chloramphenicol, clindamycin, ceftriaxone, oxacillin and trimethoprim/sulfamethoxazole. Two different disk-diffusion standards were used for determining the antimicrobial susceptibilities of pneumococci isolated in Iceland: CLSI (2009 - May 2012) and Eucast (June 2012 – present) [251, 252]. Isolates were first cultured onto Columbia agar with sheep blood and incubated overnight at 37ºC in 5% CO_2. A sterile swab was used to suspend bacterial growth in 1 ml phosphate buffered saline (PBS) to a turbidity matching a 0.5 McFarland standard. A new sterile swab was then used to spread the bacterial cell suspension onto Mueller Hinton agar with sheep blood. Disks containing the antimicrobial being tested were then placed on the agar plates and the cultures were incubated overnight at 37ºC in 5% CO_2. The diameter of the zone of inhibition around each disk was then measured and compared against the published interpretive criteria.

Antimicrobial susceptibility data for other published genomes included in analyses in this thesis were obtained from the original publications and antimicrobial susceptibility data for unpublished pneumococci including the historical, PMEN and South African genomes were provided by our collaborators.

2.3 Serotyping

The pneumococci collected as part of the VICE study were initially serotyped using either latex-agglutination or PCR. Following the development and validation of seqSerotyper.R (see Serotyping based on the cps locus sequence below), sequence-based serotyping replaced latex-agglutination and PCR for serotyping the VICE study pneumococci.

Serotypes for all other genomes included in this thesis were obtained from the original publications or from the collaborators who provided the samples. The original serotypes for these pneumococci were obtained using a variety of serotyping methodologies including Quellung, latex agglutination and PCR. For the analyses in Chapters 4 and 5, all of the genomes were re-serotyped using seqSerotyper.R.

2.4 DNA extraction protocols

DNeasy DNA extraction for whole genome sequencing (South African pneumococci [23])

Sterile disposable loops were used to subculture freezer stocks of isolates on to half of a Columbia agar with sheep blood (Oxoid) plate. Plates were incubated overnight at 37ºC in 5% CO_2. A third of the resulting growth on each half-plate was then subcultured onto three tryptic soy agar (Oxoid) with 1000 U/ml catalase (bovine liver, Sigma-Aldrich) plates and distributed with a swab before further incubation overnight at 37ºC in 5% CO_2. The addition of catalase is required for growth on tryptic soy agar (TSA) so as to neutralise the hydrogen

peroxide produced during pneumococcal growth. Without the presence of catalase, hydrogen peroxide release would lead to premature cell lysis followed by DNA denaturation. Pneumococcal growth from the three TSA plates was then suspended in 1 ml phosphate buffered saline (PBS; pH 7.4, Fisher Scientific) and centrifuged at 7600 rpm (Eppendorf 5417C) for ten minutes. The supernatants were removed and the remaining pellets resuspended in 180 µl enzymatic lysis buffer + 2% lysozyme (chicken egg white, Sigma-Aldrich) and incubated in a water bath at 37ºC for 30 minutes. 25 µl Proteinase K and 200 µl Buffer AL (Qiagen) were added to each of the extracts which were then well mixed. This was followed by a further water bath incubation at 70^{o}C for 30 minutes. 200 µl ethanol (96-100%) was added before thorough mixing by inversion. The mix was then pipetted into a DNeasy mini column sitting in the provided 2 ml collection tube and centrifuged at 10,000 rpm (Eppendorf 5417C) for 1 minute. The flow-through and collection tube was then discarded and the DNeasy mini column placed in a fresh 2 ml collection tube. 500 µl of Buffer AW1 was added and the mix was centrifuged for 1 min at 10,000 rpm (Eppendorf 5417C). Again, the flow-through and collection tube was discarded and the DNeasy mini column placed in a fresh 2 ml collection tube. 500 µl of Buffer AW2 was added and the mix was centrifuged for 3 min at 13,000 rpm (Eppendorf 5417C) to dry the DNeasy membrane (this centrifugation step ensures that no residual ethanol is carried over to the following elution). The flow-through and collection tube was discarded and the DNeasy mini column was placed in a clean 1.5 ml microcentrifuge tube where 115 µl of TE buffer was added before a final incubation at room temperature for 1 minute took place. The mix was then centrifuged at 10,000 rpm (Eppendorf 5417C) for 1 minute to produce a ~110 µl elute containing the purified DNA.

Promega DNA extraction for Whole-genome sequencing (VICE study, PMEN clones and historical pneumococci [253])

Isolates from Iceland were sent to Oxford on charcoal swabs, which were then cultured on Columbia agar with sheep blood (Oxoid) and incubated overnight at 37ºC in 5% CO_2. 1-3 colonies were picked from the blood agar plates and subcultured to 7 ml of Todd-Hewitt broth (Fisher) in a sterile 15 ml conical tube before being incubated overnight at 37ºC in 5% CO_2. Following incubation, the conical tubes were centrifuged at 4000 rpm (Eppendorf 5810R) for 15 min and the supernatant was then discarded. 300 µl of lysis buffer and 30 µl of Proteinase K was added to the conical tubes. The tubes were vortexed for 10 seconds to thoroughly mix the pellet and buffer before being placed in a water bath for 30 minutes at 70^{o}C. The mix was transferred to a Promega Buccal Swab column placed in a 1.6 ml microcentrifuge tube. The microcentrifuge tubes were centrifuged at 13,000 rpm (Eppendorf 5417C) for 2 minutes. The columns were discarded and the flow-through added to well 1 of the Maxwell 16 LEV cartridge. The cartridges were placed in the Promega Maxwell 16 robot and the LEV Blood protocol was selected. Each run on the robot took 45 minutes and the final extracts were eluted to 100 µl of elution buffer.

Quantification of DNA extracts using the Nanodrop

The protocol provided by the manufacturer (ThermoFisher Scientific) was used for the quantification of DNA using the Nanodrop spectrophotometer. Initialization of the Nanodrop spectrophorometer was begun by adding 1 µl of distilled water to the Nanodrop stage and analysing the water sample before wiping the stage clean. 1 µl of TE buffer was placed on the stage of the Nanodrop and analysed as a blank sample before wiping the stage clean. 1 µl of each DNA extract was placed on the stage of the Nanodrop and quantified. A DNA

concentration of at least 20 ng/ul was considered the minimum concentration necessary for whole-genome sequencing.

Quantification of DNA extracts using the Qubit

The protocol provided by the manufacturer (Invitrogen) was used for the quantification of DNA using the Qubit fluorometer. For DNA quantification on the Qubit platform, the Quant-IT BR (Broad Range) dsDNA assay kit (Invitrogen) was used. The Quant-IT working solution was prepared by adding 199 x n µl Quant-IT buffer to n µl Quant-IT reagent, where n = the number of samples to be quantified. The resulting solution was mixed by inversion. 2 µl of each DNA extract was added to 198 µl working solution in a clear 0.5 ml microcentrifuge tube and vortexed for 2-3 seconds. The samples were incubated at room temperature for 2-3 minutes. Following calibration with the Qubit standards, concentrations for each DNA extract were determined by the Qubit machine. A DNA concentration of at least 20 ng/ul was the minimum concentration necessary for whole-genome sequencing.

2.5 Whole Genome Sequencing

Standard Illumina multiplex library construction and paired-end sequencing was undertaken at the Wellcome Trust Sanger Institute (Hinxton, UK). The first step in library construction involved fragmenting the DNA extracts and selecting fragments <800 bp. These fragments were blunt-ended and phosphorylated in preparation for ligation to adapter sequences that included index tags, allowing for multiple genomes to be included in one sequencing library (96 or 192 genomes). Adapter-ligated fragments were PCR amplified and gel purified before being loaded onto a flow cell where the fragments were captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment was amplified to form distinct, clonal clusters through bridge amplification. Each element of a cluster was near-identical to its nearby neighbours and therefore the signal from a single molecule increased linearly,

which allowed for robust, reliable detection. The clusters were sequenced using a proprietary sequencing-by-synthesis method: each cluster was labelled using fluorescent reversible terminators that only allow one nucleoside to be added in a single cycle of sequencing, which is imaged using a camera. The sequencing cycles were repeated in order to obtain the desired read length of the final sequences (100 or 150 bp) before the fluorescent labels were removed and the sequences separated *in silico* using the index tags.

2.6 Whole Genome Sequence Analysis

Genome Assembly

Raw sequencing reads were assembled using Velvet [249], a de Bruijn graph assembler, to obtain preliminary assemblies. Contigs <300 bp were filtered out and SSPACE [254] was used to scaffold the assembled contigs based on information obtained from the original paired-end sequencing reads. GapFiller [255] was then run on these scaffolded contigs in an attempt to close gaps between the contigs and improve the overall quality of the assemblies. SMALT 0.5.7 [256] was used to assess the quality of the improved assemblies by mapping the raw sequencing reads against the assemblies generated through the pipeline in order to obtain quality metrics. These improved assembly files were assessed based on the number of contigs and total length of the assembly: assemblies with more than 100 contigs or with a total length outside the typical range of a pneumococcal genome (1.9 Mb – 2.2 Mb) were assumed to be contaminated, either by another pneumococcal sequence or another bacterial species, and were removed. The remaining assemblies were uploaded to BIGSdb.

MLST

As described in Chapter 1, MLST alleles are defined based on the sequences of seven housekeeping genes. CCs were defined by running goeBurst [21] on the whole

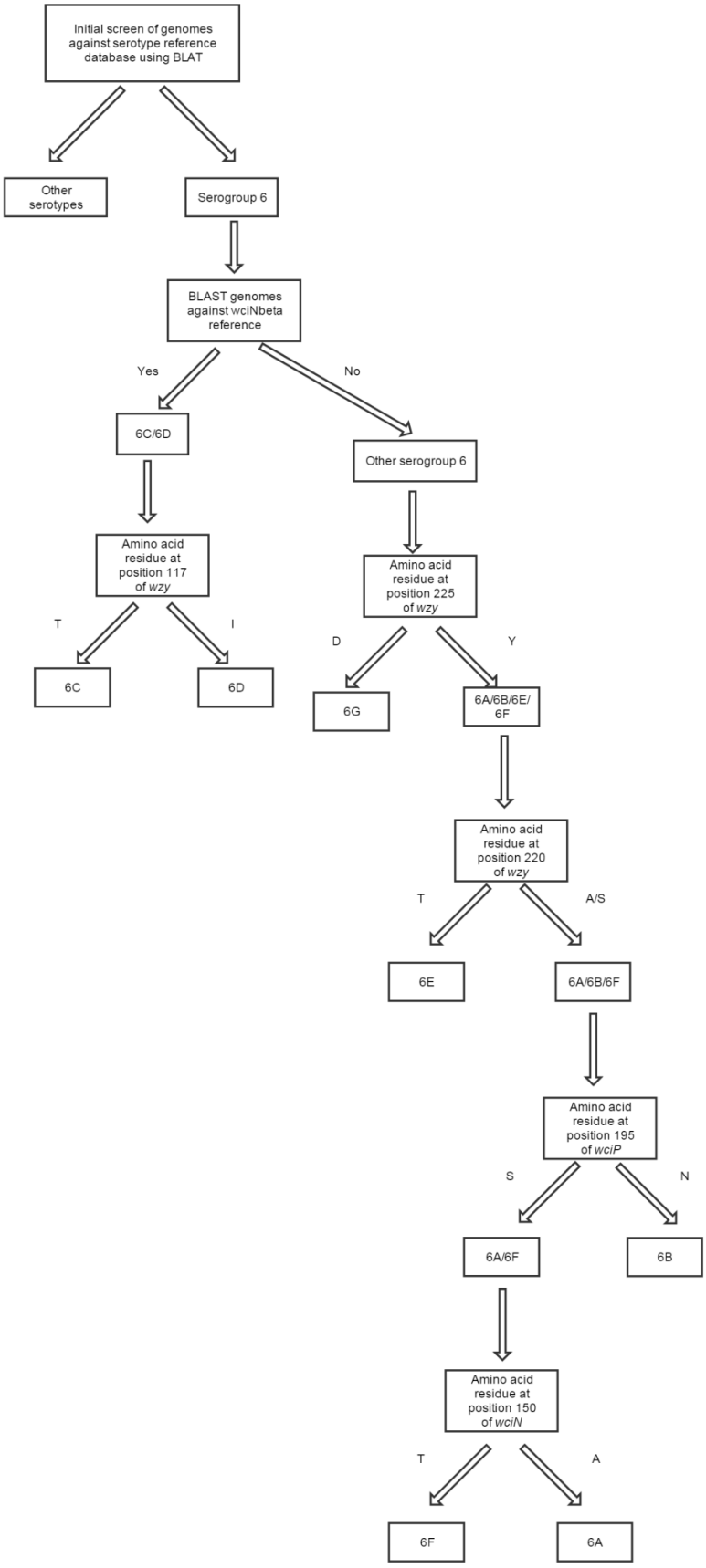


Figure 2.1 Flowchart outlining the pipeline for serotyping serogroup 6 pneumococci

pneumococcal MLST database downloaded from the MLST website (http://spneumoniae.mlst.net/). When goeBurst could not resolve the group founder, the group was assigned to CC NoneX where X is the ST with the lowest numerical value in the group. When a lack of closely related STs meant that a CC could not be assigned, they were named SingletonX where X is the isolate ST.

Serotyping based on the *cps* locus sequence

A custom program, SeqSerotyper, was written in R for the purposes of using WGS to assign serotypes to pneumococcal samples. The first step of the serotyping pipeline is a BLAT screen of the genomes against a collection of reference serotype sequences obtained from previously published studies [15, 107, 108, 110, 257-260]. Samples belonging to serogroups with serotypes defined by particular amino acid residues and/or specific alleles are separated out and re-analysed to more accurately determine the serotype. Figure 2.1 shows a flowchart of the serogroup 6 specific part of the pipeline with the specific amino acid residues and alleles used to assign each of the serogroup 6 serotypes indicated.

BLAST/BLAT

BLAST (Basic Local Alignment Search Tool) is an algorithm designed to compare nucleotide or amino acid sequences [246]. Typically, a BLAST search involves the comparison of a query sequence against a library of sequences to identify sequences that match the query sequence above a certain threshold. The BLAST package includes a number of different programs including BLASTN (for querying nucleotide sequences), BLASTP (amino acid queries) and TBLASTX (compares the six-frame nucleotide translations of a nucleotide query sequences against the six-frame translations of a nucleotide sequence library). BLAT (BLAST-like alignment tool) works in a similar manner to BLAST but was designed to be faster and more efficient than BLAST [261].

Gene prediction and annotation

WGS were annotated using Prokka [262], a tool designed for the rapid annotation of prokaryotic genomes. Using external programs, Prokka predicts coding sequences and transfers annotation from a custom database based on pneumococcal gene sequences downloaded from Genbank. Prokka also provides options for predicting mRNAs, ribosomal RNA and signal peptides and creates a number of output files in standard format that can be used as inputs for other programs.

Pan-genome analyses

Genome Comparator

Genome Comparator is a module of BIGSdb that allows a number of genomes to be compared against an annotated reference [20, 250]. Users can specify BLASTN thresholds for nucleotide identity and the proportion of a gene in the query dataset that needs to match the equivalent gene in the reference sequence (alignment). Each gene in each of the query genomes that is also present in the reference is assigned an allele number based on how much the query sequence differs at a nucleotide level from the reference sequence and an X if the reference gene is not present in the query genome. Genome Comparator identifies potential paralogues (genes with more than one copy within the same genome that may indicate a problem with sequence assembly) and these are generally removed from any downstream analyses. Genome Comparator can also create gene alignments for each of the query genome genes that are present in the reference genome as well as neighbour-net phylogenetic diagrams [263] based on the matrix created by comparing the presence or absence and nucleotide sequences of each of the reference genes in the query dataset.

Roary

Roary is a program that enables the pan-genome for large numbers of genomes to be calculated on an average desktop computer and takes as input annotated genome files (gff format) such as those produced by Prokka [264]. Coding sequences are extracted from the gff files and then converted to protein sequences. After filtering out partial sequences, cd-hit [265] is used to iteratively cluster the protein sequences based on sequence identity. This set of clustered protein sequences is then used as input for BLASTP in order to run an all against all BLAST comparison using a user-defined percentage sequence identity. The results of a final clustering step using orthoMCL [266] are merged together with the cd-hit clustering results and clusters of homologous genes are defined. The frequency of each defined cluster of homologous genes in the set of isolates is determined and from here the pan-genome for a given group of isolates is calculated. Roary also has options for calculating a core genome based on a user-defined cut-off as well as for creating alignments based on the core genome.

Phylogenetics

FastTreeMP

FastTreeMP [267] is a tool for that uses multiple cores or processors to quickly and accurately construct maximum likelihood phylogenetic trees using large numbers of genomes. FastTreeMP allows users to specify which model to use to construct phylogenetic trees and all trees constructed using FastTreeMP in this thesis used a nucleotide general time-reversible substitution model.

ClonalFrameML

Ignoring recombination when constructing phylogenetic trees may lead to erroneous structuring in the tree that does not accurately depict the true clonal relationship of the

genomes (ref). It is therefore necessary that recombination be accounted for when constructing phylogenetic trees based on sequences from bacterial genomes. ClonalFrameML [268] is an updated version of the original ClonalFrame [269] software that attempts to reconstruct phylogenetic trees to show the true clonal relationships between different genomes by detecting the location of recombination regions on each branch of a prior phylogenetic tree generated using a tree-building program such as FastTreeMP.

Interactive Tree of Life (iTOL)

All phylogenetic trees created using FastTreeMP and ClonalFrameML in this thesis were annotated using the web-based service, the Interactive Tree of Life (iTOL) [270]. iTOL allows users to upload trees in a number of different formats as well as text-based files that can be used to annotate clades or else add rings around the tree to display relevant metadata such as serotype or antimicrobial susceptibility.

3. Defining the estimated core genome of bacterial populations using a Bayesian decision model.

This chapter has been published as:

van Tonder, A.J., Mistry, S., Bray, J.E., Hill, D.M.C., Cody, A.J., Farmer, C.L., Klugman, K.P., von Gottberg, A., Bentley, S.D., Parkhill, J., Jolley, K., Maiden, M.C.J. and Brueggemann, A.B. 2014. Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Comput Biol* **10**(8): e1003788.

The development of the Bayesian decision core genome model was a collaboration with Shilan Mistry and Chris Farmer from the Mathematical Institute at the University of Oxford. Shilan Mistry was also a member of the Brueggemann group and was responsible for the mathematical aspects of the Bayesian decision model, whilst I was responsible for providing the raw data and testing of the model. The process of developing and testing the model was included in Shilan Mistry's DPhil thesis.

Martin Maiden's group in the Department of Zoology provided support and genome data for this chapter: the *Neisseria meningitidis* genomes analysed in this chapter formed the basis of Dorothea Hill's DPhil thesis; the *Campylobacter jejuni* genomes were taken from work done by Alison Cody; and James Bray downloaded raw sequence data for the *Staphylococcus aureus* and *Helicobacter pylori* samples, assembled the data and uploaded the genomes to the rMLST database. Keith Jolley developed the BIGSdb infrastructure, including the Genome Comparator tool, which allowed for the genomes and associated metadata used in this chapter to be easily accessible for analysis.

Anne von Gottberg provided South African pneumococcal isolates which were included in the pneumococcal dataset analysed in this chapter; the DNA was extracted in Oxford and

whole genome sequencing was performed at the Wellcome Trust Sanger Institute with the assistance of Stephen Bentley and Julian Parkhill. I downloaded pneumococcal genomes from Genbank and uploaded them to the Brueggemann BIGSdb database.

3.1 Abstract

The bacterial core genome is of intense interest and the volume of whole genome sequence data in the public domain available to investigate it has increased dramatically. The aim of this study was to develop a model to estimate the bacterial core genome from next-generation whole genome sequencing data and use this model to identify novel genes associated with important biological functions. Five bacterial datasets were analysed, comprising 2,096 genomes in total. A Bayesian decision model was developed to estimate the number of core genes, pairwise evolutionary distances (p-distances) based on nucleotide sequence diversity were calculated, and the median p-distance for each core gene relative to its genome location was plotted. Visually-informative genome diagrams were designed to depict areas of interest in genomes. Case studies demonstrated how the model could identify areas for further study, e.g. 25% of the core genes with higher sequence diversity in the *Campylobacter jejuni* and *Neisseria meningitidis* genomes encoded hypothetical proteins. The core gene with the highest p-distance value in *C. jejuni* was annotated in the reference genome as a putative hydrolase, but further work revealed that it shared sequence homology with beta-lactamase/metallo-beta-lactamases (enzymes that provide resistance to a range of broad-spectrum antibiotics) and thioredoxin reductase genes (which reduce oxidative stress and are essential for DNA replication) in other *C. jejuni* genomes. This Bayesian model of estimating the core genome is principled, easy to use and can be applied to large genome datasets. This study also highlighted the lack of knowledge currently available for many core genes in bacterial genomes of significant global public health importance.

3.2 Introduction

The advent of next-generation sequencing (NGS) has greatly increased the number of bacterial genomes sequenced and made available for study in public databases such as GenBank, the Sequence Read Archive and European Nucleotide Archive (ENA) [248, 271, 272]. Increasing computational power allows for comparative genomics studies involving hundreds or even thousands of sequences, but large scale computational resources are not available to all researchers. Developing methods for analysing large datasets that capitalise on the computational power of modern desktop computers will make comparative genomics analyses much more accessible to the wider research community, allowing this vast quantity of data to be analysed more extensively.

Identifying the core complement of genes in a bacterial species is often the first step in population genomics studies and the core genome can be defined in different ways. The most conservative and most frequently employed method is to only include genes present in 100% of isolates within the study population; however, this presents problems related to both biological sampling and the sequencing technology. Any collection of isolates is a subset of the entire population for the species of interest, and if the subset of isolates has limited genetic diversity then the number of "core" genes shared by all isolates in that sample will be higher than in a dataset which is genetically more diverse. This is not necessarily a problem, unless the intention is to extrapolate the findings to the wider bacterial population. Another biological limitation to using a 100% cut-off for inclusion in the core genome is that there may be rare variant strains which are missing genes that would otherwise be considered core genes. These variant strains may survive long enough to be sampled, potentially skewing the analyses. More generally, the size of the core genome is dependent on the size of the data set, with the core genome decreasing in size as more genomes are added to the analysis [233].

A large proportion of the bacterial genome sequences available at the time of writing are produced using next-generation sequencing platforms such as Illumina or Roche 454, so that even high-quality assemblies remain as incomplete or "draft" genomes. This is acceptable for most studies, but analyses of these genomes may exclude a gene from a list of core genes simply because it contains a sequence gap or is otherwise incomplete at that locus in the assembly of one or a few genomes. This assumes that the sequences being compared are all full-length: if an analysis accepts less than full-length coding sequences then gaps may not be an issue, but there will be other challenges with using incomplete sequence data, *e.g.* calculating pairwise distance (p-distance) measures.

If a definition of core genes as those found complete in all isolates in the dataset is too conservative, then the problem becomes that of determining an acceptable limit to the number of isolates missing any particular gene. One approach is to plot a frequency distribution that indicates how many genes are present in all isolates, or are missing in one or more isolates within the study population (Figure 3.1). For some bacterial species, there is a reasonably clear delineation between genes present in a large proportion of the study population versus those that are infrequent or rare, but for other bacterial species it is not clear.

Rather than make an arbitrary decision, a statistical model for estimating the core genome was developed that can be applied to different bacterial species by formalising the decision in the language of probability. The aim was to develop a Bayesian decision model to identify the genes found in what will be called the "estimated core genome" and apply this decision model to several large bacterial genome datasets. The nucleotide sequence diversity for each gene in the estimated core genome was described and used to consider how core genome sequence diversity varied across unrelated bacterial species. Finally, the data was depicted in a way that allowed the exploration of the sequence data in greater detail and generated testable hypotheses about the estimated core genome.

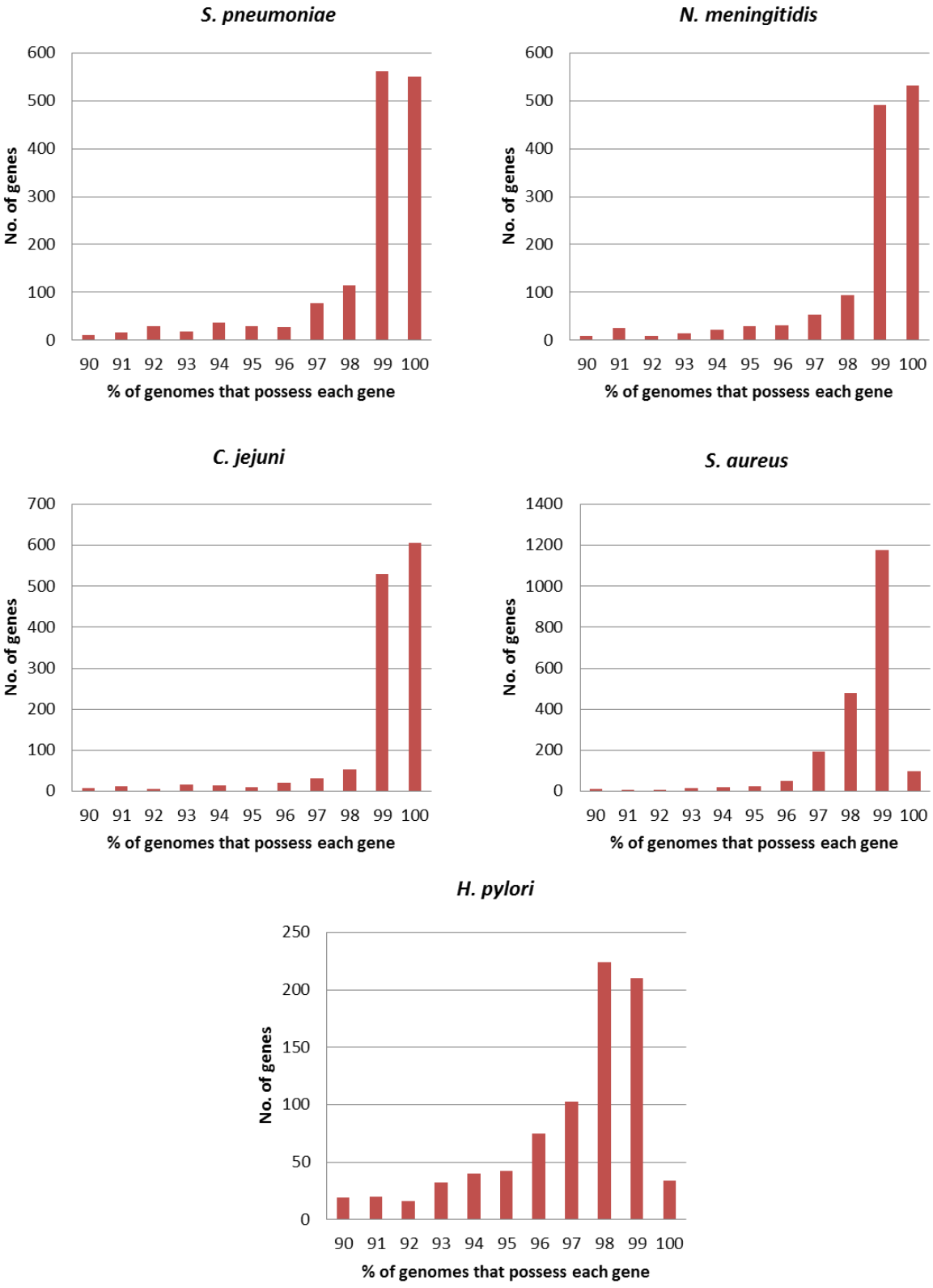


Figure 3.1 Bar charts depicting the number of genes present in 90-100% of isolates in each bacterial genome dataset.

The five bacterial species chosen for inclusion in this study were disease-causing organisms responsible for a large proportion of the global bacterial disease burden: *Streptococcus pneumoniae* (respiratory disease, the most important cause of infectious disease mortality); *Campylobacter jejuni* (gastrointestinal disease); *Neisseria meningitidis* (meningitis); *Staphylococcus aureus* (skin and soft tissue infections); and *Helicobacter pylori* (gastrointestinal ulcers).

3.3 Methods

Datasets selected for analyses

Complete lists of the bacterial genome data included in this study, with accession numbers and available metadata (obtained by consulting the relevant published papers or websites) are listed in Supplementary File 3.1. Publicly-available whole genome sequence data for *H. pylori* (N = 107 genomes) and *S. aureus* (N = 534 genomes) were collected in two ways: 1) raw sequence trace files from the ENA [248] were downloaded via an in-house genome assembly pipeline, assembled using Velvet [249] and uploaded to the rMLST BIGSdb database [20, 250]; and 2) finished reference genomes for each species were downloaded from GenBank and uploaded to the rMLST BIGSdb database. Only genomes that were already published in the scientific literature were included in the analyses. By comparison to MRSA genomes, few methicillin-susceptible *S. aureus* (MSSA) genomes are currently available (27 MSSA genomes were available at the time of our analysis, 18 of which were ST398) and the analyses was restricted to MRSA genomes only. Approximately 1000 *S. aureus* genomes were publicly available and published at the time, and a diverse dataset of ~600 genomes was selected such that the final MRSA dataset was similar in size to the *C. jejuni* and *N. meningitidis* datasets. Many of the available *S. aureus* genomes were ST239 or ST22, thus selection proceeded as follows: i) any non-ST239/ST22 genomes were

automatically included; ii) among the 456 ST239/ST22 genomes available, 167 ST239 and 186 ST22 were selected (duplicates or re-sequenced genomes were removed); and iii) any genomes that were MSSA or an unknown ST (n = 54) were subsequently removed.

The Global Historical *S. pneumoniae* dataset (N = 336 genomes) included 85 assembled genomes from a previously published study [273]; sequences for 25 published genomes [97] downloaded from the ENA and assembled using Velvet; 134 genomes downloaded from GenBank; and 92 isolates sequenced and assembled as described in Supplementary File 3.2. Raw sequence data for the 616 pneumococcal genomes comprising the comparison Massachusetts data set [99] were downloaded from the ENA, assembled and uploaded to the rMLST BIGSdb database as described above.

Data for *N. meningitidis* were collected largely as part of the Meningitis Research Foundation Meningococcus Genome Library database (MRF GL; n = 514 genomes) and an additional four historical isolates were included (http://www.meningitis.org/genome-library). The *C. jejuni* isolates included in this study (N = 601 genomes), all of human origin, were collected at the John Radcliffe Hospital in Oxford and form part of the Oxfordshire Human Surveillance collection [274]. Sequence data for *C. jejuni* and *N. meningitidis* can be found in the PubMLST (http://pubmlst.org) and rMLST (http://pubmlst.org/rmlst) databases.

Sequence types (STs) and clonal complexes (CCs)

STs were assigned to genomes either by retrieving the ST information from previously published papers or by extracting the sequences corresponding to the MLST loci and looking up the ST on the MLST website (*S. pneumoniae* and *S. aureus*) [16]. STs and CCs for *N. meningitidis* and *C. jejuni* were extracted from the relevant BIGSdb databases. All other CCs were defined using goeBURST [17] and the species-specific MLST databases downloaded from the MLST website. Supplementary File 3.3 provides a summary of the STs and CCs

included in this study for each bacterial species apart from *H. pylori*. Although an MLST scheme is available for *H. pylori*, the high genetic diversity of the species means that virtually every new strain has new alleles and new STs, making the interpretation of such data difficult and thus we have not defined STs and CCs for *H. pylori*.

Genome Comparator analyses and nucleotide sequence alignments

The genomes in each dataset were analysed using Genome Comparator is a component of the BIGSdb genome analysis database and software suite [13]. The BLASTN parameters selected used a cut-off of 70% identity over a 100% alignment with a word size of 15. Potential paralogues were removed from the analyses by identifying which of the coding loci were found in more than one copy in any query genome and excluding these sequences from any further analyses. For each coding locus in the reference genome, ClustalW [275] sequence alignments were generated for all of the query genomes containing that particular sequence. Neighbor-Net diagrams were also created by Genome Comparator as part of its standard analysis and figures were created using SplitsTree [263].

Definition of the core genome

For each of the collections of bacterial genomes, a reference genome was chosen, consisting of a set of K genes $\{g_1, \ldots, g_K\}$. Each gene is considered independently during the analysis. Let N be the number of isolates under consideration. For each isolate, $n = 1, \ldots, N$, let $c_k^n = 1$ if the k-th gene is present in isolate n or zero if it is not present. Then, for the k-th gene, $c_k = \sum_n c_k^n$ is the number of times the gene is found in N isolates. c_k^n is modelled as a sequence of binomial random variables with the probability parameter θ_k. Letting π denote a probability density, the probability of observing the k-th gene c_k times in N isolates given the model parameter θ_k is:

$$\pi(c_k|\theta_k) = \binom{N}{c_k} \theta_k^{c_k}(1-\theta_k)^{N-c_k} \quad (1)$$

The above equation, viewed as a function of θ_k, is the binomial likelihood function. For the parameter θ_k, a prior probability density $\pi(\theta_k)$ is specified. Then, using Bayes' rule, the posterior density of θ conditional on the observed frequency is computed:

$$\pi(c_k|\theta_k) \propto \pi(c_k|\theta_k)\pi(\theta_k) \quad (2)$$

If the prior density $\pi(\theta_k)$ is assumed to be a beta density with parameters (α_k, β_k) then the prior with the binomial likelihood (1) combined with Bayes' rule (2) can be used to find that the posterior density is also a beta distribution with parameters (α_k^*, β_k^*) [276]. The parameters of the prior (α_k, β_k) are related to the posterior parameters by $\alpha_k^* = \alpha_k + c_k$ and $\beta_k^* = \beta_k + N - c_k$.

Decision rule

The posterior density $\pi(\theta_k|c_k)$ represents the uncertainty of the parameter θ_k. If the density $\pi(\theta_k|c_k)$ has a greater value near $\theta_k = 1$ then that the gene is believed to be in the core genome. In light of this observation, the following decision rule for each gene can be introduced:

$$\text{If } \pi(\theta_k|c_k) = 0 \text{ at } \theta_k = 1 \text{ then the } k\text{-th gene is } \textbf{not} \text{ in the core} \quad (3)$$

The set of genes not rejected according to (3) can be defined as the estimated core genome.

Choice of prior

The selection of a prior amounts to specifying the prior belief of whether a gene is or is not in the core genome. The belief that there is an equal uncertainty of whether a gene is present in the core or not might also be adopted. Priors that reflect this type of belief are known as near

ignorance priors [277]. Rather than selecting a near ignorance prior it can be argued that a prior should be selected to reflect the nature of the decision process. Prior to analysing any of the N isolates there is no reason to believe that any of $\{g_1, \ldots, g_K\}$ is or is not in the core. As each strain is analysed, evidence is accumulated to suggest that the k-th gene is **not** a core gene. To reflect this process the prior belief that every gene is a core gene is adopted and then an attempt to falsify this statement using the decision rule is made. Formally, this corresponds to selecting as the prior a beta density with parameters $(\alpha_k, \beta_k) = (1,0)$.

Calculating p-distance values and classification of estimated core genes into functional groups

Custom Perl scripts were used to split the merged sequence alignment files generated by Genome Comparator into separate sets of nucleotide sequence alignment files by estimated core gene, and then the nucleotide distance between each pair of sequences for each estimated core gene was calculated. The number of sites at which the nucleotides differed between each pair of genes was counted. $p_k(n,m)$ was the proportion of sites that differed between isolate n and isolate m for the k-th gene in the estimated core. Then for each gene the matrix of pairwise evolutionary distances D_k was calculated using the Jukes-Cantor model [278] where the (n,m) entry of the matrix D_k was given by:

$$D_k(n,m) = -\frac{3}{4}\log\left(1 - \frac{4}{3}p_k(n,m)\right) \quad (4)$$

The median p-distance value was then calculated as a summary statistic for each estimated core gene:

$$d_k = median\{D_k(n,m); 1 \leq n,m \leq N\} \quad (5)$$

For each species the median distances for each gene in the estimated core was collected. This information was plotted against the genome position of each gene (relative to the position of that locus in the reference genome) and depicted in a circular diagram created using Circos [279]. In addition, the estimated probability density function of the median pairwise evolutionary distances for each species, or for individual genes within a species, was plotted using ksdensity (Kernel smoothing function estimate) [280]. Finally, the COG functional groups of the genes with p-distance values greater than the 95th percentile were determined using eggNOG [281].

Computing requirements and open source code

The computationally intensive part of the analysis is the Genome Comparator run (because it creates sequence alignments for every gene), but this runs via a publicly-available web interface on a cluster of servers hosted at the University of Oxford. Output files are stored for one week on the server. The Bayesian model for estimating the number of core genes, the calculation of p-distance values for all reference genes, creation of the genome diagrams (Figure 4) and the generation of estimated probability graphs (Figure 5) can be implemented using freely available scripts written in the open source R software package (https://www.R-project.org/). Along with wrapper scripts that prepare the Genome Comparator outputs and a detailed manual, the relevant code is available at: https://sourceforge.net/projects/bayesianestimatedcoregenome/. As a frame of reference, the Genome Comparator analysis of the largest dataset, *C. jejuni* (601 genomes) took 90 hours to run, including all sequence alignments and generation of the Neighbor-Net diagram, but the subsequent steps took approximately 2 hours and can be run on any modestly-powered computer.

Comparison of core genomes (*S. pneumoniae* and *S. aureus*)

For case studies 2 and 3 the names of the core genes for each of the datasets included in the comparisons were compared to each other using the VLOOKUP function in Microsoft Excel, which matches cells containing the same text (the same reference genomes were used for each dataset so the gene names were the same) in order to generate numbers of shared and unique genes. Functional groups for each of the sets of unique genes were assigned using eggNOG [281].

3.4 Results

Description of the datasets used in analyses

In total, 2096 genomes were analysed across the 5 different bacterial species (Table 3.1 and Supplementary File 3.1). A phylogenetic network for each dataset was derived using Neighbor-Net [282] as part of the initial Genome Comparator [20] analyses; these diagrams demonstrated the overall diversity of the genomes in each study dataset (Figure 2).

The *S. pneumoniae* (pneumococcal) dataset consisted of 336 genomes for isolates of 39 different serotypes collected over 90 years (1916-2008) from at least 32 countries around the world. The isolates were recovered from individuals of a wide range of ages, including isolates from patients with disease and isolates recovered from healthy individuals. The multilocus sequence typing (MLST) data revealed 163 sequence types (STs), which could be clustered into 74 different clonal complexes (CCs) indicative of isolates with shared ancestry (Table 3.1 and Supplementary File 3.1).

Table 3.1 Summary of study datasets.

Bacterial species	Study dataset	No. of genomes	Years of isolation	No. of countries represented	No. of STs[a]	No. of CCs[a]
S. pneumoniae	Global Historical Collection	336	1916-2008	32	163	74
C. jejuni	Oxfordshire Human Surveillance 2011	601	2011	1 (UK)	134	29
N. meningitidis	MRF Meningococcus Genome Library	518	2010-2011[b]	3[c]	198	24
S. aureus	rMLST BIGSdb database[d]	534	1955-2011	27[e]	25	11
H. pylori	rMLST BIGSdb database	107	--[f]	10	--[f]	--[f]

[a]STs = sequence types; CCs = clonal complexes. [b]Four historical isolates from 1976, 1983 (n = 2) and 1986 were also included. [c]One historical isolate was from The Gambia and another was from Norway. All other isolates were from the UK. [d] http://pubmlst.org/rmlst/. [e]39% (n = 210) of the genomes were from the UK; 18% (n = 94) genomes were from unknown locations. [f]The year of isolation for virtually all genomes was unknown; STs and CCs for *H. pylori* were not defined (see main text).

The largest genome dataset analysed was that of *C. jejuni* (N = 601 genomes). Isolates were recovered from human stool samples collected from patients in Oxfordshire, United Kingdom (UK) with gastroenteritis during 2011. 134 STs from 29 CCs were characterised in this collection, which was representative of the broader *C. jejuni* population genetic diversity [274, 283].

The *N. meningitidis* (meningococcal) dataset was comprised of 518 genomes and these isolates were collected nearly exclusively from patients residing in England, Wales and Northern Ireland in the 2010/11 epidemiological year, apart from 4 historical isolates from Norway (1976), The Gambia and UK (1983) and UK (1986). The 2010/11 genomes are part

of the Meningitis Research Foundation Meningococcus Genome Library, which contains genomes from all culture-confirmed cases of meningococcal disease submitted to the Meningococcal Reference Unit in 2010/11 and 2011/12. Isolates of seven serogroups were included, mostly serogroup B (n = 394), Y (n = 74) and W-135 (n = 27). 198 STs were represented by the isolates and the STs clustered into 24 CCs. Culture-confirmed cases of meningococcal disease are largely representative of the England and Wales disease-causing *N. meningitidis* population as described previously [284].

The *S. aureus* dataset was large (N = 534 genomes) but genetically less diverse (25 STs and 11 CCs; Table 3.1 and Supplementary File 3.1) than other datasets, since the analyses were restricted to methicillin-resistant *S. aureus* (MRSA) only. Most of the publicly-available genomes that are already published are of a limited number of CCs, predominantly the MRSA CCs that are epidemiologically the most important. The MRSA isolates were recovered from patients in 27 countries, although 39% of isolates were recovered in the UK. The *H. pylori* dataset included 107 genomes and 82% of the collection was from the USA, Canada or Japan. Only limited additional metadata were available for these isolates.

Estimated core genomes for each bacterial species dataset

The size of the reference genomes used in the Genome Comparator analyses for each dataset varied from 1.6 to 2.8 Mb, and the total number of genes in each reference genome ranged from 1,566 to 2,547 (Table 3.2). There were small numbers of unique loci, i.e. genes found only in one genome: *S. pneumoniae* (n = 6); *C. jejuni* (n = 6); *N. meningitidis* (n = 7); *S. aureus* (n = 4); and *H. pylori* (45).

An initial BLASTN criteria of 70% identity was chosen, which allowed for the identification of variable sequences among conserved gene classes [285] and avoided bias towards reference-specific sequences, and 100% sequence alignment, which means that coding

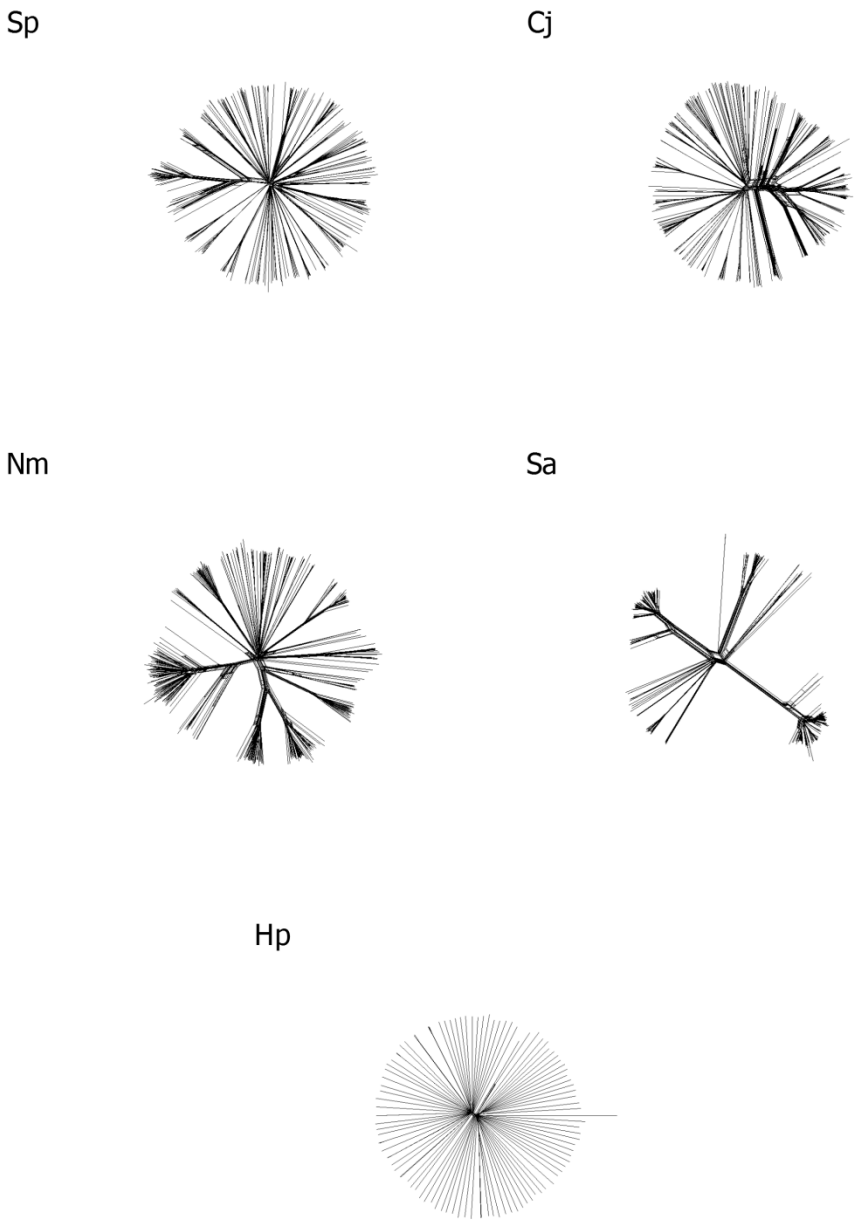


Figure 3.2 Neighbor-Net phylogenetic networks for the five species included in this study. Neighbor-Net diagrams for each species as follows: Sp) *S. pneumoniae*; Cj) *C. jejuni;* Nm) *N. meningitidis*; Sa) *S. aureus*; and Hp) *H. pylori.* Scale bars for each individual diagram were too small to be visualised and thus were removed for illustrative purposes.

sequences occurring at the ends of contigs or with gaps were therefore not included. Lowering the BLASTN criteria to 70% identity and 90% alignment increases the number of genes in the estimated core, as partial gene sequences will be detected and included (Table 3.3), which may be more suitable for other user-specific analyses but is not ideal for the calculation of p-distances to estimate sequence diversity.

Table 3.2 Summary of estimated core genome analyses.

Bacterial species	**Reference genome**	**Reference genome size (Mb)**	**Reference genes (n)**	**% of genomes that possess each core gene**	**Estimated core genes (n)**
S. pneumoniae	ATCC700669	2.22	1990	≥99.7	851
C. jejuni	NCTC11168	1.64	1623	≥99.8	866
N. meningitidis	FAM18	2.19	1917	≥99.8	744
S. aureus	HO5096_0412	2.83	2547	≥99.8	242
H. pylori	26695	1.67	1566	≥99.1	244

The smallest estimated core genomes were those of MRSA and *H. pylori* (242 and 244 genes, respectively; Table 3.2) and the *S. pneumoniae*, *C. jejuni* and *N. meningitidis* core genomes were similar in size, ranging from 744 to 866 genes. The percentage of genomes within a dataset that possessed each estimated core gene ranged from ≥99.1% to ≥99.8%. The number of putative paralogues identified in the initial Genome Comparator analyses varied from 1 in *C. jejuni* to 40 among the *S. pneumoniae* genomes, and these genes were removed from further analyses (lists of putative paralogues for each species are provided in Supplementary File 3.4). If putative paralogues had not been removed, one would have been included in each of the estimated core genomes of *S. pneumoniae* and *N. meningitidis*.

Table 3.3 Comparison of estimated core genomes using different BLASTN alignment criteria.

			70% identity / 90% alignment				70% identity / 100% alignment			
Bacterial species	Reference genome	Reference genome size (Mb)	Unique genes (n)	Potential paralogues in whole genome (n)	% of genomes that possess each core gene	Estimated core genes (n)	Unique genes (n)	Potential paralogues in whole genome (n)	% of genomes that possess each core gene	Estimated core genes (n)
S. pneumoniae	ATCC700669	2.22	3	60	≥99.7	1113	6	40	≥99.7	851
C. jejuni	NCTC11168	1.64	3	6	≥99.8	1183	6	1	≥99.8	866
N. meningitidis	FAM18	2.19	4	50	≥99.8	1104	7	21	≥99.8	744
S. aureus	HO5096_0412	2.83	1	38	≥99.8	537	4	13	≥99.8	242
H. pylori	26695	1.67	21	15	≥99.1	766	45	2	≥99.1	244

Calculated pairwise evolutionary distance values (p-distances)

Within each genome dataset, median p-distance values were calculated for each of the estimated core genes and the estimated probability density function was plotted for each bacterial species (Figure 3.3). The estimated probability density function plotted a

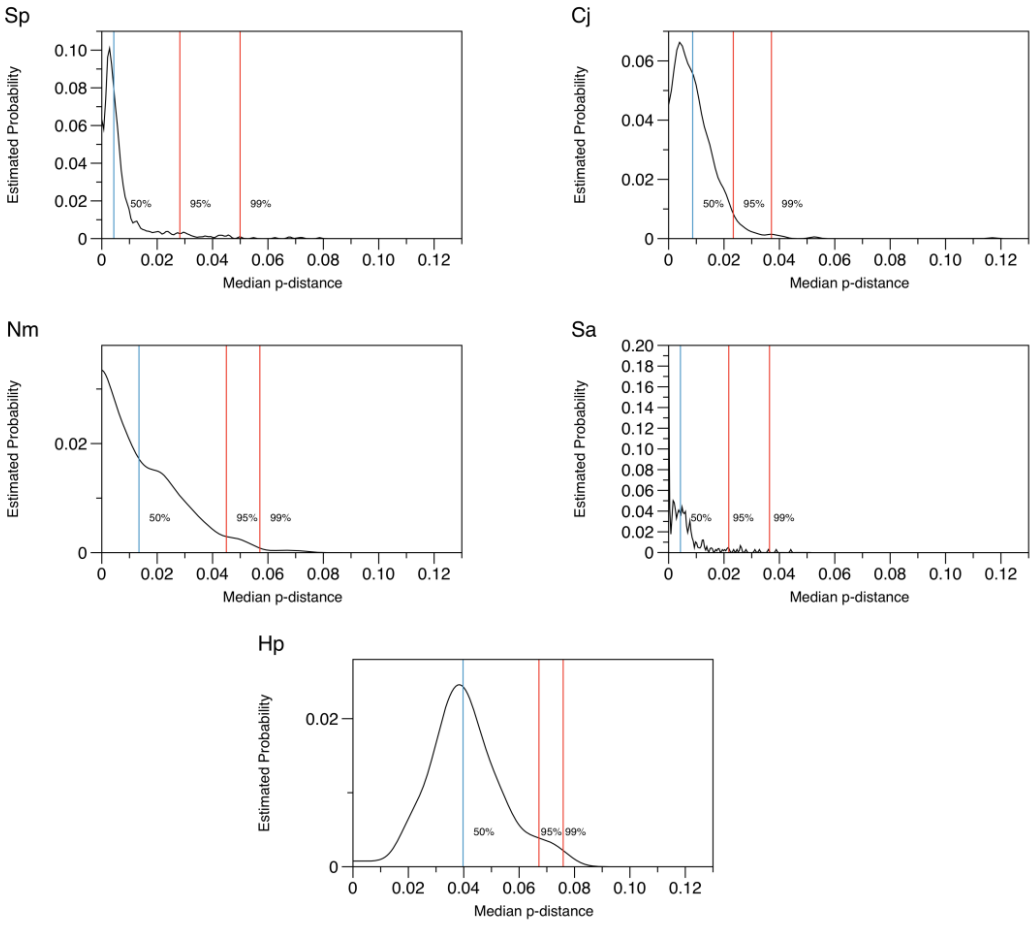


Figure 3.3 Estimated probability densities of median p-distances. The estimated probability density function was plotted for each bacterial species as follows: Sp) *S. pneumoniae*; Cj) *C. jejuni;* Nm) *N. meningitidis*; Sa) *S. aureus*; and Hp) *H. pylori*.

smoothed histogram of the median p-distances vs. the estimated probability (relative frequency) of each p-distance value. The shape of the graphs for *S. pneumoniae* and *C. jejuni* were similar, showing a large peak of very small p-distance values, i.e. highly conserved genes, but these were entirely different from the graphs depicting the data for

the other genes for that particular genome dataset. Each graph is an indication of the overall sequence diversity of the set of estimated core genes for that particular genome dataset.

The median p-distance value for each estimated core gene was then plotted against its position in the reference genome and illustrated as a circular bacterial chromosome (Figure 3.4). The length of each line indicates the median p-distance value for that gene. The estimated core genes were distributed around each genome and accessory regions in the reference genomes (e.g. ICE elements or phage genes), were observed as gaps where no core genes clustered. Estimated core genes with p-distances above 0 but less than the 95th percentile (blue lines) and those above the 95th percentile (red lines) stood out in a pattern on each genome diagram and allowed for an evaluation of specific genes and gene clusters in the genome, as demonstrated below. Supplementary File 3.5 lists all the estimated core genes and p-distances for each bacterial species. All genes with a p-distance value greater than the 95th percentile for each bacterial species, and the Cluster of Orthologous Groups (COG) functional category for each of those genes, are listed in Supplementary File 3.6.

Three case studies that expanded the initial analyses of the estimated core genome

Case study 1: Investigate the core genes in C. jejuni *and* N. meningitidis *with p-distances above the 95th percentile.*

43 estimated core genes in *C. jejuni* were over the 95th percentile (Figure 3.4-Cj). 26% (n = 11; range of p-distances 0.026-0.118) of these genes were hypothetical proteins or of unknown function, and 19% (n = 8) were associated with coenzyme metabolism (Table 3.3 and Supplementary File 3.6). The *C. jejuni* genome diagram also clearly identified a gene in the estimated core that had a much higher p-distance value (0.118) than the rest of

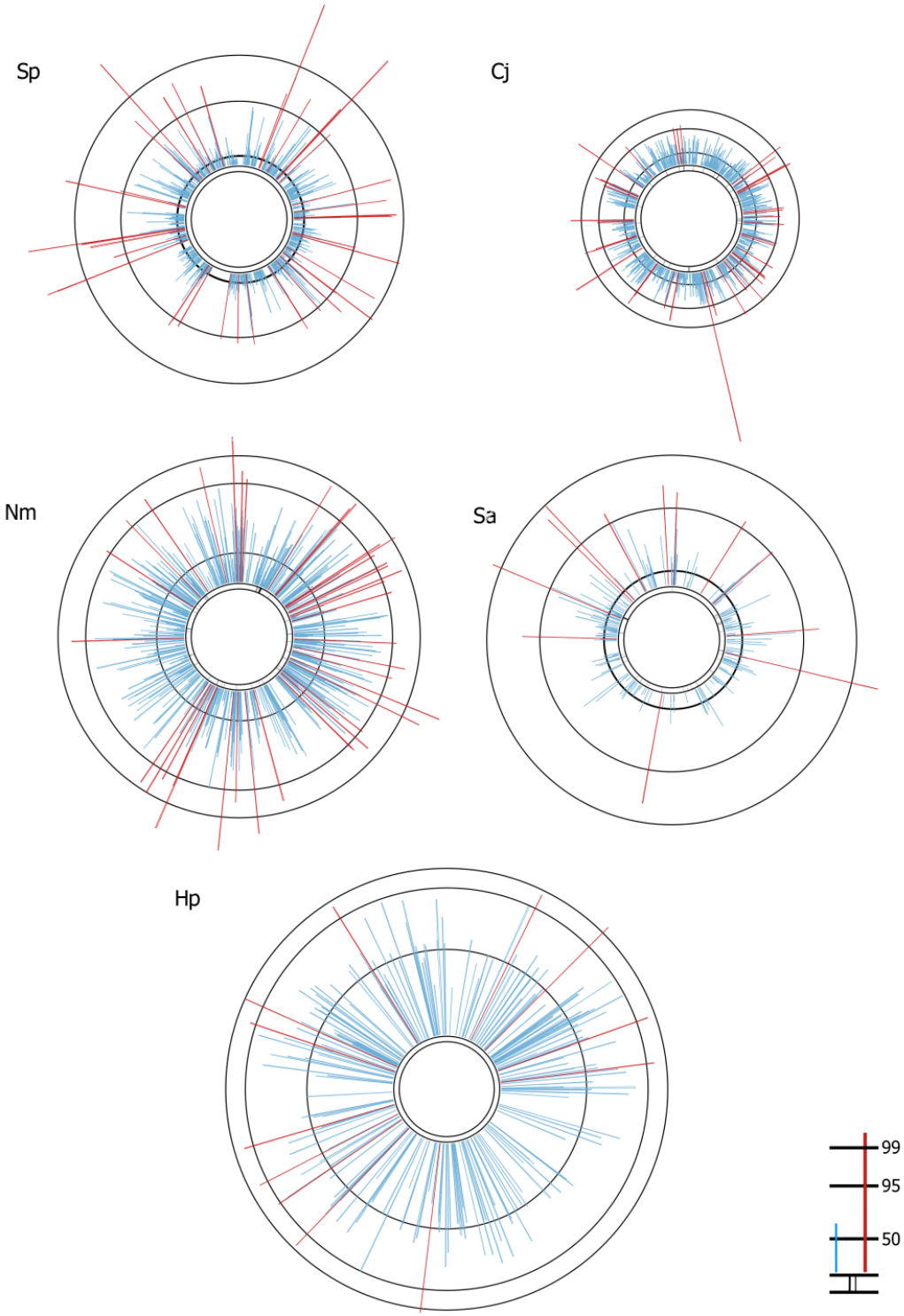


Figure 3.4 Distribution of core genes and median pairwise distances. Median p-distance plotted against genome position, relative to the reference genome, for Sp) *S. pneumoniae* (ATCC700669, 2.2 Mb); Cj) *C. jejuni* (NCTC11168, 1.64 Mb); Nm) *N. meningitidis* (FAM18, 2.19 Mb); Sa) *S. aureus* (HO5096_0412, 2.83 Mb); and Hp) *H. pylori* (26695, 1.67 Mb). The length of each line on a genome diagram was scaled according to the maximum p-distance value calculated for that particular bacterial species. The three outer rings mark the 99th, 95th and 50th percentiles. Median p-distances above the 95th percentile were coloured red; median p-distances above zero but less than the 95th percentile were coloured blue. The innermost pair of rings marks the position of genes with a median p-distance of zero.

the core genome (the longest red line in Figure 3.3-Cj). Cj0809c was annotated in the reference genome as a putative hydrolase. A Pfam [286] search of the Cj0809c amino acid sequence returned a match to the Lactamase_B superfamily and a BLAST search against the NCBI non-redundant database found orthologues of the gene annotated as metallo-beta-lactamase in other *C. jejuni*. A BLAST search of the KEGG database [287] resulted in hits to sequences annotated as beta-lactamase, metallo-beta-lactamase and thioredoxin reductase (NADPH) genes in other *C. jejuni* genomes.

Beta-lactamases and metallo-beta-lactamases are enzymes that provide resistance to a wide range of broad-spectrum beta-lactam antibiotics. Thioredoxin reductase is part of the thiol redox system, which reduces oxidative stress and is essential for DNA replication; it has also recently been proposed as a possible novel antimicrobial target site [288, 289]. The high level of sequence diversity observed in Cj0809c means that it is likely under selective pressure and/or is a site of frequent recombination, and further investigation of the true function of Cj0809c as well as the different sequence variants of this particular gene is necessary.

Overall, 37 estimated core genes in *N. meningitidis* had median p-distance values over the 95th percentile threshold and these were categorised into 12 different COG functional categories (Tables 3.3 and Supplementary File 3.6). Similar to the *C. jejuni* data, 24% (n = 9) encoded unknown hypothetical proteins or genes with only a general predicted function and their calculated p-distances ranged from 0.045-0.073. The two core genes with the highest p-distance values (NMC1154 and NMC0661) also encode putative proteins with unknown functions. NMC1154 is adjacent to *ribA*, which encodes a GTP cyclohydrolase II (*ribA* has been proposed as a novel antimicrobial target site [290]) and NMC0661 is a hypothetical protein, possibly in the YicC-like family of bacterial proteins, which are poorly characterised but may play a role in stationary phase survival [291].

Table 3.3 Cluster of Orthologous Groups (COG) functional groups for the estimated core genes with p-distances greater than the 95th percentile in the *C. jejuni* and *N. meningitidis* genomes.

		C. jejuni		*N. meningitidis*	
COG functional group		**no.**	**%**	**no.**	**%**
R/S	hypothetical protein[a]	11	25.6	9	24.3
H	coenzyme metabolism	8	18.6	7	18.9
J	translation	1	2.3	5	13.5
M	cell wall membrane envelope biogenesis	4	9.3	4	10.8
E	amino acid metabolism and transport	2	4.7	4	10.8
L	replication and repair	0	0.0	3	8.1
C	energy production and conversion	4	9.3	2	5.4
T	signal transduction	1	2.3	1	2.7
F	nucleotide metabolism and transport	1	2.3	1	2.7
P	inorganic ion transport and metabolism	0	0.0	1	2.7
G	carbohydrate metabolism and transport	3	7.0	0	0.0
O	post-translational modification, protein turnover, chaperone	3	7.0	0	0.0
I	lipid metabolism	2	4.7	0	0.0
D	cell cycle control and mitosis	1	2.3	0	0.0
K	transcription	1	2.3	0	0.0
N	cell motility	1	2.3	0	0.0
Total		**43**	**100**	**37**	**100**

a. Includes genes in COG categories R (general functional prediction only) and S (function unknown), and genes otherwise annotated as hypothetical proteins.

There were also several clusters of red lines in the *N. meningitidis* genome, indicating regions of the genome with higher sequence diversity among the estimated core genes (Figure 3.4-Nm). One gene cluster encodes proteins in the NADH dehydrogenase complex (p-distances for each core gene in this region range from 0.021-0.055; Supplementary Files 3.5 and 3.6). NADH dehydrogenase is involved in oxidative phosphorylation and has recently been identified as a region of the *N. meningitidis* genome that has undergone intragenic recombination; authors have speculated that sequence differences in this gene region may be related to differences in fitness and possibly virulence [292, 293].

The matrix of calculated p-distances across all nucleotide sequences for any particular core gene, which were summarised as a median p-distance and plotted on the genome diagram were explored. This provided insight into the range of calculated p-distances for a core gene and an indication of whether there was any particular clustering of p-distances within that gene. To investigate this, the estimated probability density graph for the core gene with the highest median p-distance value for each of the bacterial species was generated (Figure 3.5).

The graphs varied widely in shape and height, and the graphs for Cj0809c and HP0203 suggested that there were clusters of sequences that were likely to be biologically informative. Overall, these graphs suggest that the median p-distance value is a suitable summary statistic to represent the sequence diversity of a particular core gene across all genomes analysed, and as such it provides a useful starting point for exploring a very large set of core genes; however, a gene-specific plot of the range of p-distances should also be considered as it may reveal informative biological patterns in the data. Statistical analyses of the pairwise p-distance matrix would be required for further inference to be made.

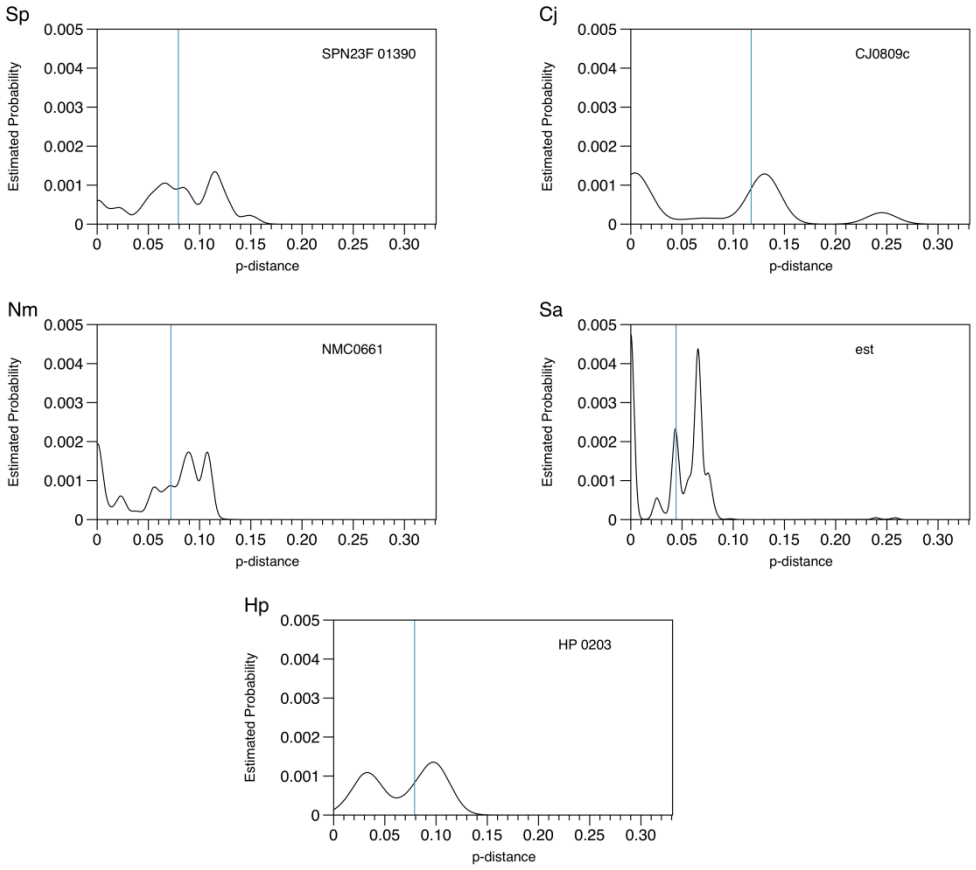


Figure 3.5 Estimated probability density of p-distances for the core gene with the highest median p-distance. Probability distributions of p-distances for the core gene with the highest median p-distance for: Sp) *S. pneumoniae*; Cj) *C. jejuni*; Nm) *N. meningitidis*; Sa) *S. aureus*; and Hp) *H. pylori*. Gene names are given in the corner of each graph.

Case study 2. A comparison of Bayesian vs COGs methodologies to estimate the core genome of S. pneumoniae.

The Bayesian model methodology was compared to that of a recently published study that estimated the core genome among *S. pneumoniae* using a COGs-based methodology [99]. Croucher and colleagues generated whole genome sequence data for 616 pneumococcal carriage isolates collected during three discrete time periods between 2001-2007 from healthy children <7 years of age in Boston, Massachusetts. Putative

protein-coding sequences from all genome assemblies were extracted and grouped into COGs; COGs present in 100% of genomes in a single copy were defined as core genes. Overall, the COGs-based method estimated the size of the core genome in the Massachusetts *S. pneumoniae* dataset to include 1194 genes, whereas the Bayesian analysis of the same dataset estimated 948 core genes. 840 genes were common to both estimations and 108 and 354 genes were unique to each methodology (Figure 3.6A; Supplementary File 3.7).

An explanation for the difference in core genome estimates was that the Genome Comparator analysis was set up to only consider full-length genes (100% alignment of the putative coding sequence), whereas partial sequences were included in the COGs-based methodology. To test this, the Genome Comparator analysis was re-run with a lowered threshold of 90% sequence alignment (*i.e.* up to 10% of the sequence could be missing and the gene would still be included in the estimated core genome). The result was that the sets of estimated core genes were very similar: 1206 vs. 1194 genes, of which 1027 genes were found in both lists (Figure 3.6B).

It is useful to note here that the output generated using the Bayesian methodology includes a list of the estimated core genes (named according to the reference genome), their predicted functions (including hypothetical proteins) and location in the genome (relative to the reference genome). The COGs-based analysis outputs data for COG groups, which requires the user to undertake additional processing steps to acquire the corresponding list of gene names, sequences and genome position.

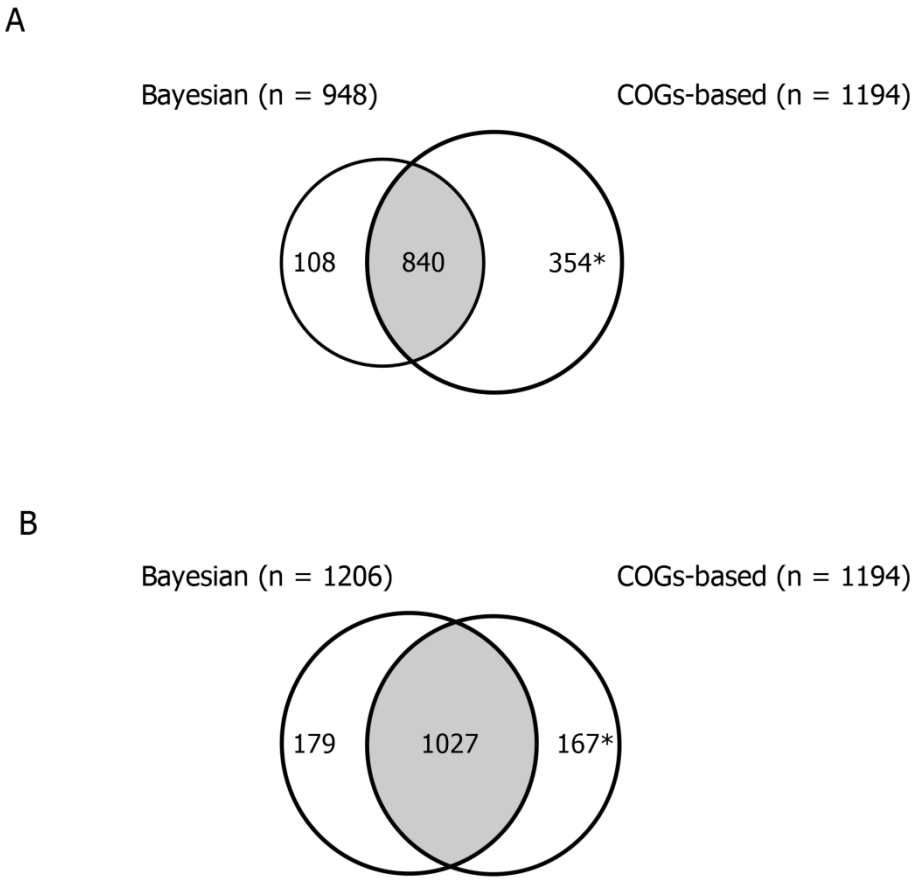


Figure 3.6 A comparison of the results of two different methodologies for estimating the core genome, using the 616 *S. pneumoniae* genomes from Massachusetts [11] as the assessment dataset. Comparison of the number of core genes identified using the Bayesian method versus the number of core genes identified using the COGs-based method. A) Results obtained using an initial BLAST cut-off of 70% identity and 100% sequence alignment for the Bayesian analysis; and B) Results obtained using an initial BLAST cut-off of 70% identity and 90% sequence alignment for the Bayesian analysis. *Note that the total includes 10 predicted coding regions that are not in the ATCC 700669 reference genome.

Case study 3. Comparison of the estimated core genome of specific MRSA CCs and STs.

The estimated core genome for the MRSA dataset derived from the Bayesian model was much smaller than previous estimates of the number of core genes in *S. aureus*/MRSA.

Prior estimates of the *S. aureus* core genome varied from 1492 to 2266 genes [294-296], whilst the figure obtained using the Bayesian model was only 242 genes. This was due to a number of factors: i) recently published analyses of larger MRSA genome datasets were restricted to a single CC, in which case the set of genes in the estimated core will be larger since the genomes share common ancestry [294, 297, 298]; ii) the much larger sample size of dataset in this study (534 genomes) in comparison to previous estimates using datasets of 14-63 MRSA genomes [294-296]; and iii) the exclusion of partial gene sequences (Supplementary File 8) which means that the estimate of the genes found in the MRSA core genome is very conservative. To test whether the small estimated MRSA core genome may also be associated with major core genome differences between unrelated CCs, the MRSA genome data was reanalysed in Genome Comparator by running the two largest CCs, 5 and 22, individually. The CC-specific data were then compared with respect to the number of estimated core genes and the COG functional groups associated with each unique set of core genes.

CC5 and CC22 were comprised of 207 and 235 genomes, respectively, in the MRSA dataset. 81% (167/207) of genomes in CC5 were ST239, a widely-distributed, multidrug-resistant clone that is among the most common MRSA lineages worldwide [299, 300]. The predominant genotype in CC22 was ST22 (186/235; 79%), also known as EMRSA-15 and responsible for 80% of healthcare-associated MRSA infection in the UK [301].

Core genomes of 692 and 1114 genes were estimated for CC5 and CC22, respectively, and 396 genes were common to both CCs (Figure 3.7; Supplementary File 3.8). A total of 296 genes were defined as core genes in CC5 but not in CC22: 45% of these were associated with metabolic functions, 22% with information storage/processing, 12% with cellular processes and signalling, and 21% were poorly categorised or with unknown

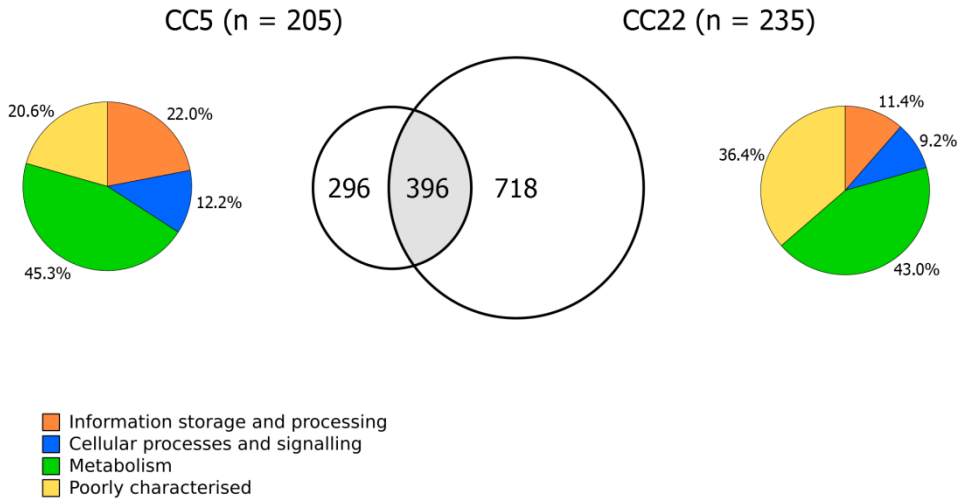


Figure 3.7 Comparison of *S. aureus* clonal complexes CC5 and CC22. Venn diagram of the comparison between the estimated core genes for CC5 and CC22, along with pie charts showing the COG functional categories associated with the genes that were unique to each dataset.

function. In contrast, 718 estimated core genes in CC22 were not part of the core genome of CC5: the proportions of unique genes associated with metabolism or cellular processes/signalling were similar to that of CC5, but half as many were associated with information storage/processing and notably, over one-third (n = 261) of unique core genes in CC22 were poorly categorised or with unknown function. Given that CC22 is a major healthcare-associated MRSA clone, an argument could be made for identifying and discerning the functions of these unique core genes.

3.5 Discussion

This study exploited the large volume of publicly-available whole genome sequence data to outline a method for analysing bacterial genomes in a straightforward way, using web-based tools and computer programmes that run on modestly-powered computers. The

analyses described here do not require access to supercomputers. The resulting data can be explored in a biologically relevant manner and there is flexibility to change the analysis parameters to suit different datasets and different questions.

As an example, the core genes from among the coding loci were defined as those present at full sequence length so that complete gene sequence information was included and p-distances could be reliably calculated. The model is designed to allow for the inclusion of genes present in <100% of genomes, which adjusts for arbitrary contig assembly issues. It is important to note that by excluding partial genes, many of which will be incomplete due to breaks in gene sequences based on sequencing technicalities and/or the genome assembly, the core genome estimates generated using the Bayesian model are a conservative estimate for each bacterial species. However, as was demonstrated, simply lowering the sequence alignment to <100% will increase the number of estimated core genes, which may be appropriate for some datasets and analyses. In other words, the unit of count was the complete gene, but if the unit of count was nucleotides or the aim was to generate a list of full plus incomplete estimated core genes, then including partial sequences would be appropriate and requires the user to simply lower the sequence identity threshold in the initial analysis.

Most importantly, the process described here is completely transparent and the assumptions are easily understood, *e.g.* a list of genes is exported that includes the locus name, putative product, sequence length and genome position, which allows for detailed user inspection. The initial Genome Comparator output includes information about which genes are truncated (found at the ends of contigs) in each query genome and the user should evaluate these data carefully in conjunction with an assessment of the overall quality of the genome assemblies and consider whether or not to explore partial genes as part of a separate analysis. Furthermore, if a user wished to analyse amino acid

sequences as opposed to nucleotide sequences, this is possible by selecting the appropriate option at the start of the Genome Comparator analysis; however, this will also significantly increase the run time and memory requirements and will become an issue with large genome datasets. A simpler option would be to convert the aligned nucleotide sequences for genes of interest into protein sequences after the Genome Comparator run is completed.

The estimated core genome sizes that were obtained using the Bayesian model were expected to be lower than previous estimates, since the number of genes common to all genomes in a dataset decreases as the number of genomes increases, and the study datasets are much larger and more diverse than the great majority of those previously analysed (see Table S8 for a list of relevant references). That is not a criticism of the previous studies and the analyses of small numbers of whole genomes; it is simply an indication of what data were available at the time each study was undertaken and how much the genomics field has changed in a short span of time.

The utility of using this Bayesian model was demonstrated by determining which core genes were more diverse than others, but other investigators may wish to focus on the core genes that are highly conserved in a particular dataset. It is important to note that there is no one definitive "core genome" – the estimates of core genes will vary from dataset to dataset and between different methodologies.

A Bayesian model for calculating the core genome was chosen in this study but a frequentist model, based on a hypothesis test, could also have been applied. The advantage to using a Bayesian model is that it allows for the formalisation of the decision rule by a particular choice of prior. Furthermore, the Bayesian model implemented here does not account for correlations between the genes and each gene is considered to be

statistically independent of each other; however, many genes are known (or likely) to be linked and operate as part of an operon or cassette. A revision of the proposed model could assess correlations between specific genes and/or gene regions. Such modifications would be a significant computational and statistical challenge given the large volume of genome data one would potentially wish to analyse, but a correlation model could provide useful biological information from the sequence data. Moreover, the median was selected as the summary of the distribution of gene pairwise distances since the underlying distributions of p-distances are not Gaussian. A better representation of the underlying probability distributions could be achieved by considering a number of different percentiles, but for this study the analyses were restricted to just the median.

Finally, the case studies that were highlighted demonstrated how the data outputs and initial analyses may be used to derive more focussed analyses on specific genes or gene regions in the genome and generate new hypotheses to test. The genome diagrams are a useful way of depicting areas of potential interest in the genome and the bacterial species evaluated here presented a huge range of possibilities for further study, from which a few were selected to highlight as examples.

The field of bacterial genomics is advancing rapidly and it is now possible to generate enormous quantities of sequence data (albeit currently incomplete) at a low cost; therefore, it is also essential to find and develop suitable, widely-accessible and inexpensive methods of processing and analysing these data in order to maximise the utility and benefits of whole genome sequence data. This model formalises the estimation of the core bacterial genome as a Bayesian decision problem and the resulting outputs reveal many areas for further exploration of the bacterial core genome.

3.6 Appendixes

Appendix 3.1 List of references to previously published papers that estimated a bacterial core genome.

Species	First author	Year	Reference	No. of estimated core genes	No. of whole genomes	Comments
S. pneumoniae	Donati	2010	[235]	1666	44	
	Hiller	2011	[302]	1324	52	
	Croucher	2013	[99]	1194	616	
C. jejuni	Lefubure	2010	[303]	~1300	43	
	Friis	2010	[304]	1295	13	
	Biggs	2011	[305]	1001	2	
N. meningitidis	Hotopp	2006	[306]	1706	3	
	Schoen	2008	[307]	1337	6	
	Schoen	2009	[308]	1330	7	
	Rusniok	2009	[309]	1736	5	
	Budroni	2011	[310]	1630	20	
	Krauland	2012	[311]	1776	2	
	Kong	2013	[293]	1090	36	
S. aureus	Hall	2010	[296]	1923	14	
	Harris	2010	[297]	???	63	ST239 only
	Castillo-Ramirez	2011	[294]	1492	63	ST239 only, follow on from Harris study above
	Boissy	2011	[295]	2266	17	
	Castillo-Ramirez	2012	[312]	???	165	ST239 only; includes 62 from Harris study above
	Holden	2013	[298]	???	193	ST22 only
H. pylori	McClain	2009	[313]	1237	5	
	Fischer	2010	[314]	1223	7	
	Kawai	2011	[315]	1079	20	
	Lara-Ramirez	2011	[316]	1186	9	
	Gunaletchumy	2012	[317]	760	10	

3.7 Supplementary information (see attached USB drive)

Supplementary File 3.1 Datasets

Supplementary File 3.2 DNA extraction protocol

Supplementary File 3.3 Genome STs and CCs

Supplementary File 3.4 Putative paralogues

Supplementary File 3.5 Core genes

Supplementary File 3.6 Core genes with p-distance > 95^{th} percentile

Supplementary File 3.7 Bayesian core genome versus COGs core genome

Supplementary File 3.8 MRSA results

4. Genomics reveals the worldwide distribution of multidrug-resistant serotype 6E pneumococci

The results described in this chapter were presented at the 12th European Meeting on the Molecular Biology of the Pneumococcus (Europneumo 2015), Oxford July 2015.

This chapter has been published as:

van Tonder, A.J., Bray, J.E., Roalfe, L., White, R., Zancolli, M., Quirk, S.J., Haraldsson, G., Jolley, K.A., Maiden, M.C.J., Bentley, S.D., Haraldsson, A., Erlendsdóttir, H., Kristinsson, K.G., Goldblatt, D. and Brueggemann, A.B. 2015. Genomics reveals the worldwide distribution of multidrug-resistant serotype 6E pneumococci. *J Clin Microbiol*. **53**(7): 2271-85.

Sequence data for two published datasets (USA and Thailand), were downloaded from the ENA, assembled using Velvet and, following a quality control step, uploaded to the rMLST database where the pneumococcal genomes were assigned MLST profiles by James Bray in the Department of Zoology, University of Oxford. I extracted metadata for the published pneumococcal datasets from the original publications.

Pneumococci from Iceland were collected by my collaborators as part of the VICE study and were sent to Oxford on swabs where I cultured and stored the samples before extracting the DNA for sequencing. DNA extracts were sent to the Wellcome Trust Sanger Institute where sequence template libraries were constructed by the Illumina library construction team before sequencing was performed using the Illumina Hiseq platform. The resulting sequence data was quality controlled before being inputted into the Sanger assembly and improvement pipeline which assembled the genomes using Velvet and closed gaps in the assemblies computationally using SSPACE and GapFiller.

I downloaded the resulting assemblies from the Sanger ftp site, assessed the genome quality based on contig length and number and excluded any substandard genomes from further analyses. I then uploaded the genomes to my group's BIGSdb database where the MLST profiles were scanned and assigned. I downloaded pneumococcal genomes from Genbank and uploaded them to the Brueggemann BIGSdb database. The genomes were serotyped based on the whole genome sequence data using a script I developed, seqSerotyper.R.

The opsonophagocytosis assays (OPA) used to evaluate functional antibody responses to serotype 6E were undertaken in Professor David Goldblatt's laboratory at the Institute of Child Health, University College London. Strains were sent from Oxford to London where members of Prof. GoldbIatt's group, in particular Lucy Roalfe, Rebecca White and Marta Zancolli, performed the assays.

4.1 Abstract

Serogroup 6 pneumococci are a frequent cause of serious disease and a common coloniser of the nasopharynx in children. Serotypes 6A and 6B have been recognised for many decades, but more recently serotypes 6C and 6D, which are genetically similar to serotypes 6A and 6B, were discovered. Serotype 6E was first reported in 2004 but was thought to be rare; however, serotype 6E has been detected among recent pneumococcal collections by a number of groups. Therefore, a diverse dataset of ~1000 serogroup 6 genomes was analysed, the prevalence and distribution of serotype 6E was assessed, the genetic diversity among serogroup 6 pneumococci was analysed and the inhibition of serotype 6E pneumococci by pneumococcal conjugate vaccine (PCV)-induced serotype 6A and 6B antibodies was investigated. This study found that 43% of all genomes were serotype 6E and they were recovered worldwide from healthy children and patients of all

ages with pneumococcal disease. Four genetic lineages, three of which were multidrug-resistant, described ~90% of the serotype 6E pneumococci. Serological assays demonstrated that vaccine-induced serotype 6B antibodies were able to kill serotype 6E pneumococci. This study also revealed three major genetic clusters of serotype 6A capsular sequences, discovered a new hybrid 6C/6E serotype, and identified 44 examples of serotype switching. Therefore, while vaccines appear to offer protection against serotype 6E, genetic variants may reduce vaccine efficacy in the longer term due to the emergence of serotypes that can evade vaccine-induced immunity.

4.2 Introduction

Serogroup 6 is a particularly important serogroup, as it is one of the most common serotypes found in the nasopharynx of unvaccinated young children and is a major cause of serious pneumococcal disease among all age groups [318, 319]. Serotypes 6A and 6B have been recognised for many decades, but more recently serotypes 6C and 6D, which are genetically similar to serotypes 6A and 6B, were discovered [109, 258, 320]. From a vaccine perspective, it was shown that the 6B antibodies elicited by PCV7 were partially protective against serotype 6A, but not serotype 6C [321-323]. However, PCV13 was shown to protect against 6A, 6B and 6C [323, 324] and PCV10 is protective against 6B and possibly 6A [321, 325]. Serotype 6D pneumococci have been reported infrequently, although the prevalence in South Korea was estimated to be 10% of clinical strains, and a PCV7-induced cross-protective immune response to serotype 6D was demonstrated [326, 327]. Serotypes 6F, 6G and 6H have been described very recently and whether PCVs provide any protection against these serotypes is unknown [110, 328].

The first report of serotype 6E pneumococci was by Mavroidi *et al*, in a study that explored sequence diversity and evolution among serogroup 6. Internal fragments of the

three serotype-specific genes were sequenced in a diverse collection of 102 isolates of serotype 6A and 6B pneumococci. While they found little sequence divergence between serotype 6A and most of the 6B isolates, they did identify a group of what they called 'class 2' serotype 6B sequences, which were >5% divergent from the majority of serotype 6B isolates [128]. Two subsequent studies of serogroup 6 diversity and evolution in other pneumococcal collections confirmed the existence of 'class 2' serotype 6B or what one report called '6B-III' or possible 'serotype 6E' strains [258, 329]. Very recently, investigators have reported serotype 6E pneumococci in several Asian countries [259, 330, 331]. The ongoing VICE study characterising Icelandic pneumococci pre- and post-PCV10 implementation also discovered serotype 6E strains. Furthermore, when the genome sequences of several serotype 6B PMEN reference strains were interrogated it was found that they all possessed a serotype 6E *cps* locus sequence.

Therefore, a large and diverse dataset of ~1000 serogroup 6 pneumococcal genomes was compiled to investigate three aims: (i) determine the prevalence, distribution and epidemiology of serotype 6E; (ii) examine the genetics of the serogroup 6 *cps* locus and the molecular epidemiology of serogroup 6 lineages; and (iii) assess whether the serotype 6B polysaccharides in PCV7 and PCV13 induce the production of protective antibodies to serotype 6E.

4.3 Methods

Pneumococcal genome dataset.

A dataset of 1,059 assembled serogroup 6 pneumococcal genome sequences was compiled using previously published genome datasets [99, 100, 236], GenBank sequence data (http://www.ncbi.nlm.nih.gov/GenBank) and genome data from the VICE study (Figure 4.1; Supplementary File 4.1).

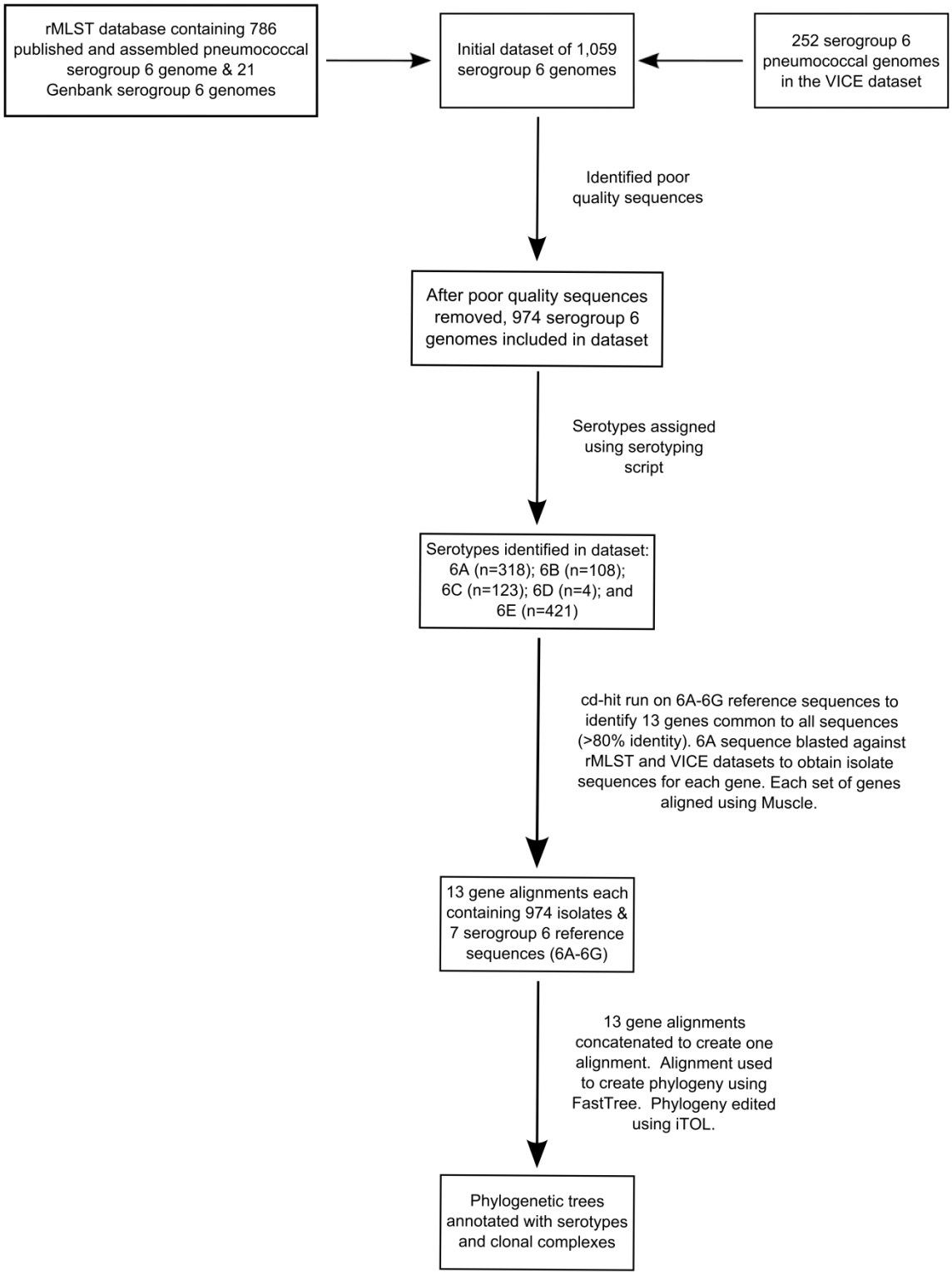


Figure 4.1 Flowchart outlining the compilation of the pneumococcal genome data set and corresponding cps locus sequences that were analysed in this dataset.

The genome dataset included four serotype 6B pneumococcal reference strains from the Pneumococcal Molecular Epidemiology Network (PMEN) collection [245], for which genome sequences were available (Table 4.1). The genomes from Genbank were downloaded directly, and all other genomes in the dataset were downloaded as raw sequence reads from the European Nucleotide Archive (ENA), assembled using Velvet [249] and deposited in the rMLST database, which is powered by BIGSdb [20, 250]. Corresponding metadata was manually acquired from the original publications and matched to the genome data. Genome sequence quality was assessed and poor quality sequences (e.g. those with gaps or non-full length gene sequences in the *cps* locus) were removed, leaving 974 genomes for analyses. All genome assemblies analysed in the current study, with corresponding metadata, are available from the pneumococcal PubMLST website (http://pubmlst.org/spneumoniae/).

Sequence types (STs), clonal complexes (CCs) and serotypes

Multilocus sequence type (MLST) data were automatically extracted from each genome using BIGSdb [20] and STs were clustered into clonal complexes (CCs) using Phyloviz [332]. Serotypes for each of the 974 pneumococci included in the study were assigned based on the genome sequence, using seqSerotyper.R; briefly, the genomes were serotyped by identifying amino acid residues and/or alleles specific for each serotype: serotypes 6A/6B, $wciP_{195}$; serotypes 6C/6D, presence of *wciNβ* and wzy_{117}; serotype 6E, wzy_{220}; 6F, $wciN_{150}$; and 6G, wzy_{225}.

Table 4.1 Reference PMEN[a] genomes and *cps* locus sequences used in this study.

Reference	Other name	ST[a]	Accession number	Year	Other info	Reference(s)
PMEN2	Spain^{6B}-2	ST90	ATCC700670; ENA	1988	Abscess in human patient, Madrid	[245, 333]
PMEN8	S. Africa^{6B}-8	ST185	ATCC700675; ENA	1990	Human blood culture, South Africa	[245, 334]
PMEN12	Finland^{6B}-12	ST270	ATCC700903; ENA	1987	Isolated in Finland	[245, 335]
PMEN17	Maryland^{6B}-17	ST384	ATCC BAA342; ENA	1997	Invasive disease isolate, USA	[336]
6A *cps* locus	34351 Rodrigues 6A/2	ST7188	CR931638	1952	Isolated in USA	[15]
6B *cps* locus	2616/39 6B/3	ST7199	CR931639	1939	Isolated in Denmark	[15]
6C *cps* locus	CHPA388	--	EF538714	1999-2002	Nasopharyngeal isolate, USA	[257]
6D *cps* locus	MNZ21	ST282	HM171374	2008	Isolated from nasopharynx of healthy child, South Korea	[337]
6E *cps* locus	HK02-14	ST90	Multiple (accession numbers for each sequenced gene, see reference)	2008	Isolated from sputum sample, Hong Kong	[259]
6F *cps* locus	MNZ1135 PS6657	--	KC832411	1992-2012	Clinical adult sample, Germany	[110]
6G *cps* locus	MNZ1136 PS16864	--	KC832410	1992-2012	Clinical adult sample, Germany	[110]

a. PMEN = Pneumococcal Molecular Epidemiology Network; ST = sequence type as determined by multilocus sequence typing.

Analyses of the *cps* locus sequences.

Thirteen *cps* locus genes were identified among all seven serogroup 6 reference sequences using cd-hit [265] and a sequence identity threshold of >80%: *wzg*, *wzh*, *wzd*, *wze*, *wchA*, *wciO*, *wciP*, *wzy*, *wzx*, *rmlA*, *rmlC*, *rmlB*, *rmlD* (Figure 4.3 depicts phylogenetic trees for each of the 13 genes). Note that *wciN* was also present among all *cps* loci, but α and β versions of *wciN* were highly divergent, as noted previously [257, 258]. The serotype 6A reference sequence for each of the 13 genes was then BLASTed [246] against the study genome dataset to extract the sequences for the 13 genes from each genome. Pairwise estimates of evolutionary distance (p-distance; number of nucleotide sites that differ between two sequences divided by the total number of nucleotides compared) were calculated for each of the 13 genes in the 974 genomes, stratified by serotype.

The extracted *cps* locus sequences were aligned gene-by-gene using Muscle [338] before being concatenated together to obtain a 12,271 bp *cps* locus alignment for each of the 974 genomes. The concatenated sequences were then input into FastTree2 to construct a *cps* locus phylogeny, using a nucleotide general time-reversible model [267]. The resulting phylogeny was annotated using iTOL [270]. P-distances between the serogroup 6 *cps* locus reference sequences were calculated using MEGA5 [339]. Input sequences were 13,416 bp in length and spanned the *cps* locus from the start of *wzg* through the end of *rmlD*, including intergenic regions, but excluding *wciN*, *HG262* (present in the *cps* locus of serotypes 6A, 6F and 6G only) and *HG263* (present in serotype 6B and 6E only).

Genome-wide assessment of sequence diversity.

The Genome Comparator module in BIGSdb was used to compare all 974 serogroup 6 genomes to the annotated reference genome 670-6B (NC_014498.1), also known as

PMEN2 or Spain^{6B}-2. The BLASTn parameters were set to ≥70% sequence identity and 100% sequence alignment length [20]. The data were exported to an Excel spreadsheet

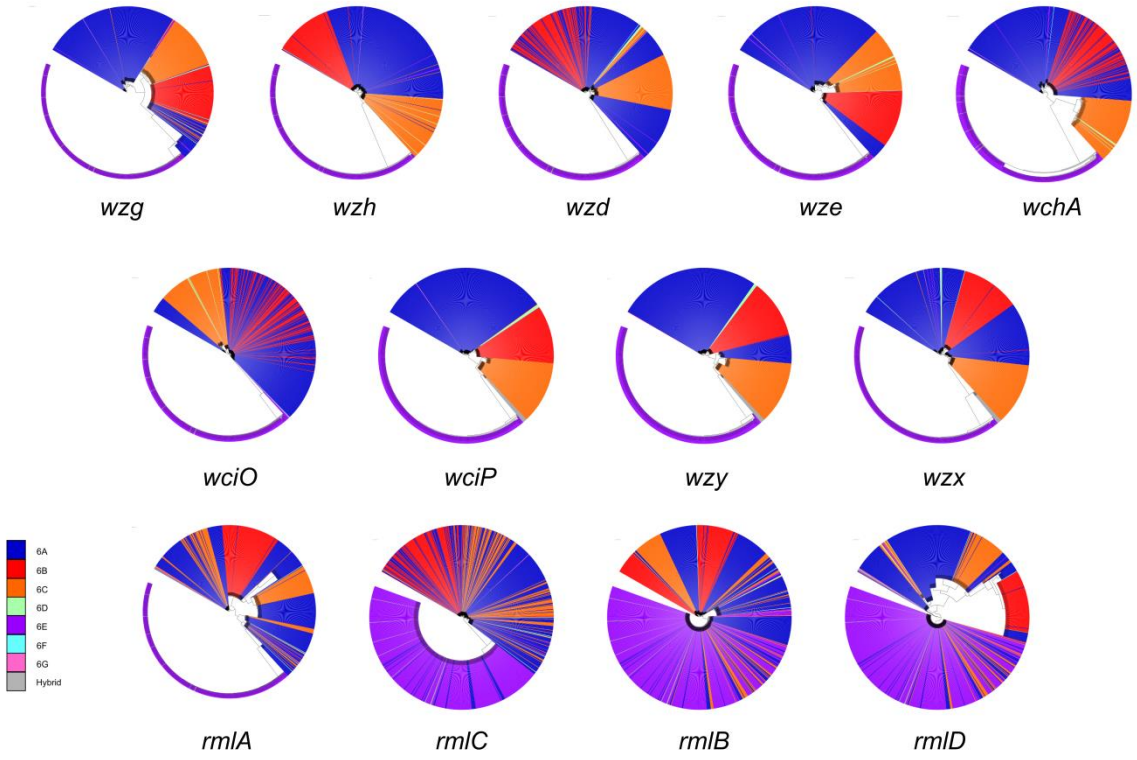


Figure 4.2 Phylogenetic trees for each of the 13 common genes in the cps locus, constructed using the entire set of 974 genomes.

that depicted the results of sequence comparisons of each annotated coding sequence (hereafter referred to as 'gene' for simplicity) in the reference 670-6B genome to each of the 974 query genomes (Supplementary File 4.2). Data were output on a gene-by-gene basis across the entire genome for every query genome. 670-6B reference gene sequences were marked as allele '1' and corresponding sequences in each query genome were marked as: 'X' (not present); '1' (identical to the reference); 'N' (sequence present

but non-identical to the reference, marked with allele numbers to indicate unique sequences); or 'T' (sequence present but truncated). The exported Genome Comparator gene-by-gene data analysis for the full 2.24 Mb genome was read into R (http://www.r-project.org/) to create a pseudo-heat map to compare gene presence/absence and sequence diversity across all 974 genomes.

The Genome Comparator analysis also revealed that 432 gene sequences were found in all 974 genomes in full coding length, thus for each of the genomes these 432 genes were concatenated (292 Kb in total) and FastTree2 was used to assemble a phylogenetic tree that represented all 974 genomes. Multilocus sequence type (MLST) data were available for each genome and the STs were clustered into CCs using Phyloviz [332] (see Supplementary File 4.1). The 432-gene phylogenetic tree was annotated with CC and serotype data for each genome.

Serological analyses of serotype 6E pneumococci.

To evaluate functional antibody to serotype 6E, post-primary vaccination sera from PCV7 and PCV13 recipients (n = 10) who participated in previous vaccine studies [340, 341] were selected. Sera were analysed by an opsonophagocytic assay (OPA, titre ≥1:8 considered positive) [342] utilising 5 different isolates of serotype 6E pneumococci (PMEN2 and PMEN8 plus three recent Icelandic isolates). To assess the contribution of serotype-specific antibody to killing of cross-reactive antigens, sera were retested by OPA after adsorbing sera with purified 6A, 6B (ATCC, USA) or 6C capsular polysaccharide (Statens Serum Institut, Denmark) one at a time and reported as the titre where 50% of bacteria were killed.

4.4 Results

Epidemiology of serogroup 6 pneumococci.

The study genome dataset represented a diverse set of 974 serogroup 6 pneumococci recovered in 16 different countries across five continents between 1972 and 2014 (Table 4.2). 69% (n = 675) of pneumococci were recovered from carriage, 27% (n = 267) from disease and 4% (n = 32) from an unknown source, from individuals spanning a wide range of ages. 78% (n = 760) of the pneumococci were collected prior to any PCV introduction in the country of origin. Antibiogram data demonstrated a range of antimicrobial-susceptible and -resistant isolates among the serotypes, although susceptibility data were missing for many of the pneumococcal genomes. The full list of genomes and associated metadata are detailed in Supplementary File 4.1.

Based upon the molecular analysis of the dataset, 421 of 974 pneumococci were found to be serotype 6E. They were recovered from individuals between the ages of 6 months and 87 years residing in 15 different countries in Europe, North America, South America, Africa and Asia (Table 4.2). The serotype 6E pneumococci were isolated from 1981 onwards and 90% were isolated prior to the use of any PCVs. Serotype 6E isolates were recovered from both healthy young children and individuals of all ages with pneumococcal disease. Diseases caused by serotype 6E pneumococci spanned the range of typical pneumococcal diseases, *i.e.* otitis media, sinusitis, empyema, pneumonia, bacteraemia and meningitis (Supplementary File 4.1).

One-third of the genomes (n = 318) were serotype 6A pneumococci recovered between 1972 and 2013 from patients of all ages residing in six different countries. The majority of serotype 6A pneumococci were recovered from healthy children, although 67 isolates from patients with invasive and non-invasive diseases were also included (Table 4.2;

Table 4.2. Epidemiological characteristics of individual serotypes within 974 serogroup 6 pneumococcal genomes.

Serotype	6A	6B	6C	6D	6E	Hybrid	Total
Isolate n (%)	318 (33)	108 (11)	115 (12)	4 (0.4)	421 (43)	8	974
Years of isolation	1972-2013	2001-2013	2001-2014	2004	1981-2013	2004-2006	1972-2014
Country of origin:							
Thailand	125	0	61	4	202	8	400
Iceland	118	100	14	0	130	0	362
USA	53	8	40	0	20	0	121
South Africa	19	0	0	0	10	0	29
Portugal	0	0	0	0	13	0	13
Germany	0	0	0	0	10	0	10
South Korea	0	0	0	0	8	0	8
Spain	0	0	0	0	7	0	7
China	0	0	0	0	6	0	6
Peru	0	0	0	0	6	0	6
Turkey	2	0	0	0	1	0	3
France	0	0	0	0	2	0	2
Finland	0	0	0	0	1	0	1
Oman	0	0	0	0	1	0	1
PNG[e]	1	0	0	0	0	0	1
Unknown	0	0	0	0	4	0	4
Recovered from:							
Carriage	247	80	107	4	229	8	675
Disease	67	28	8	0	164	0	267
Unknown	4	0	0	0	28	0	32
Patient age (y)[b]	<0.5-87	0.5-83	<0.5-82	unknown	<0.5-87	unknown	<0.5-87
Vaccine status[c]							
pre-PCV	227	73	69	4	379	8	760
post-PCV	89	35	46	-	35	-	205
unknown	2	-	-	-	7	-	9
Susceptibility (µg/ml) [d]							
Penicillin	<0.03-16	<0.03-	<0.03-1	S[d]	<0.03-4	0.06-1	<0.03-16
Erythromycin	<0.03-16	≤0.03-	<0.03-2	--	<0.03-	--	<0.03-
Tetracycline	≤0.5->4	≤0.5-0.25	≤0.5-0.25	--	≤0.5-64	--	≤0.5-64
Chloramphenicol	2	2	2-4	--	2->8	--	2->8

a) Serotypes were determined from the nucleotide sequence of the *cps* locus. No pneumococci of serotypes 6F or 6G were identified in the study genome dataset. The hybrid serotype is genetically a combination of both 6C and 6E *cps* locus sequences (see Results); b) Age data were missing for 625 genomes, but available data indicated that isolates of serotypes 6A, 6B, 6C and 6E were recovered from both children and adults; c) PCV = pneumococcal conjugate vaccine; vaccine status refers to whether any PCV was being used in the country of origin at the time of pneumococcal isolation; d) Susceptibility data were missing for many genomes, but the ranges of available data are given here. S = susceptible. See Supplementary Table 1 for more details; e) PNG=Papua New Guinea

Supplementary File 4.1). Pneumococci with a serotype 6C *cps* locus sequence comprised 12% (n = 115) of the dataset and were predominantly recovered from healthy children in Thailand, Iceland and the USA since 2001.

11% (n = 108) of the pneumococci possessed a serotype 6B *cps* sequence. Serotype 6B pneumococci were isolated between 2001 and 2013 from both carriage and disease (otitis media, pneumonia and bacteraemia), from patients of a wide range of ages. Four serotype 6D pneumococci from Thailand were identified among the genomes, but no serotype 6F or 6G pneumococci. Eight carriage pneumococci from Thailand with a hybrid serotype 6C/6E were also identified and these are discussed in more detail below. Figure 4 provides a comparison of the serotype distribution among all 974 genomes using the original serotyping data versus sequence-based serotyping.

Serotype-specific prevalence estimates.

The study genome dataset was diverse and compiled from several different genome collections, thus estimates of serotype-specific prevalence based on the entire dataset may not be representative of the global pneumococcal population. However, the carriage datasets from the Maela refugee camp in Thailand and from Massachusetts could reliably be used to assess the prevalence of serotypes within specific geographical locations during a specified time period.

The Maela genome dataset was comprised of 3,085 pneumococci collected from infants and mothers living in a rural refugee camp in Thailand from 2007-2010 [100, 343]. PCVs had not been used in this setting prior to or during the study period. 398 serogroup 6 pneumococcal genomes were identified, of which 50% (n = 200) were serotype 6E, 31% (n = 125) were serotype 6A, 15% (n = 61) were serotype 6C, 1% (n = 4) were

serotype 6D and 2% (n = 8) were a hybrid 6C/6E serotype (Supplementary File 4.1). No pneumococci with a serotype 6B *cps* locus sequence were identified.

In contrast, 616 pneumococcal genomes were collected from healthy young children in Boston, Massachusetts, USA, during three time periods (2001, 2004 and 2007) after implementation of PCV7 [99]. 97 serogroup 6 pneumococci were identified, of which 47% (n = 46) were serotype 6A, 35% (n = 34) were serotype 6C, 9% (n = 9) were serotype 6E and 8% (n = 8) were serotype 6B (Supplementary File 4.1).

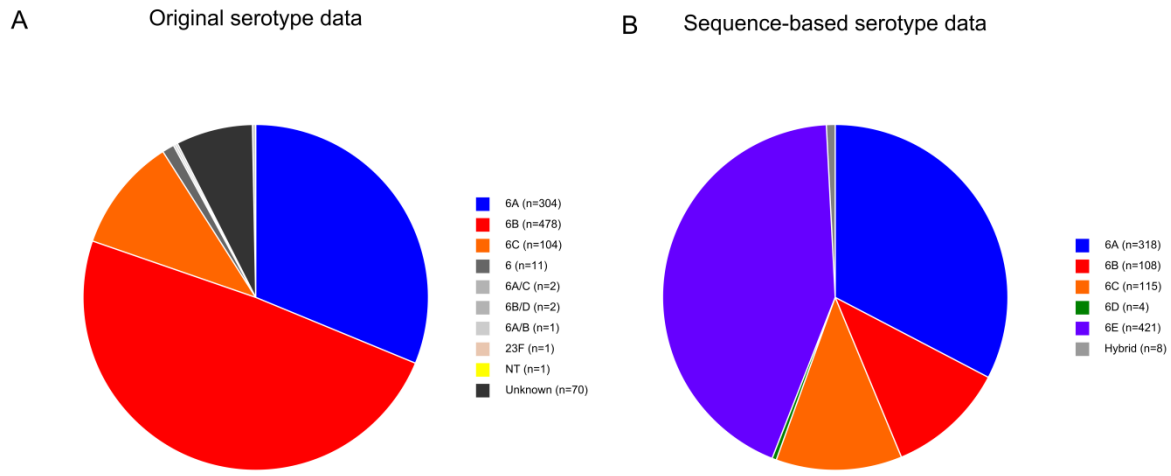


Figure 4.3. Using the entire dataset (n = 974 genomes), a comparison of the original serotyping data as previously published for each genome versus the sequence-based serotyping data generated in this study. The majority of the unknown serotypes in part A correspond to Icelandic pneumococci from the study currently in progress and for which serotyping data generated *via* traditional methods were not available.

Serotypes among PMEN clones.

Genome sequences for four PMEN reference strains were included in this study: PMEN2 (Spain^{6B}-2), PMEN8 (S. Africa^{6B}-8), PMEN12 (Finland^{6B}-12) and PMEN17 (Maryland^{6B}-17), as shown in Table 4.1 (44; http://pubmlst.org/spneumoniae/). All four

were previously identified as serotype 6B (presumably based on the Quellung reaction), but all possess a serotype 6E *cps* locus sequence. Moreover, one dataset included in this study was compiled specifically to study the PMEN2 lineage [344], which is a multidrug-resistant lineage of pneumococci detected in many countries around the world, but was originally identified in Iceland and Spain in the 1980s (http://pubmlst.org/spneumoniae/). 189 pneumococcal genomes were sequenced in the original PMEN 2 study. 172 of these had complete *cps* locus sequences and were thus included in the current study, and all possessed a serotype 6E *cps* locus sequence.

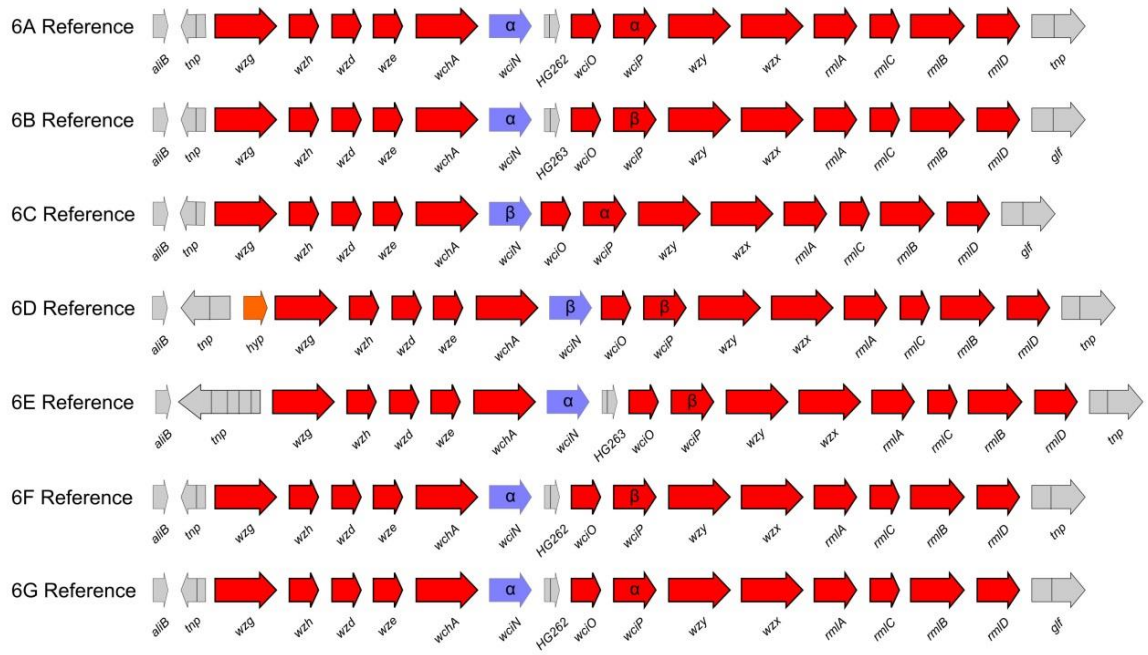


Figure 4.4 Organisation of the *cps* locus for each of the seven serogroup 6 reference sequences (see Table 4.2 for accession numbers). Genes coloured red are the 13 genes common to all serogroup 6 *cps* loci at >80% sequence similarity; *wciN* was coloured blue and the α or β allele was indicated; transposons and other pseudo-genes were coloured grey; and a hypothetical gene was coloured orange. α or β versions of *wciP* were also indicated.

Table 4.3 Estimates of pairwise evolutionary distances between serogroup 6 *cps* locus sequences, and among individual genes within the *cps* locus of each serotype.

	Serotype						
***cps* locus[a]/gene[b]**	**6A**	**6B**	**6C**	**6D**	**6E**	**6F**	**6G**
6A							
6B	0.008						
6C	0.016	0.011					
6D	0.011	0.016	0.012				
6E	0.068	0.065	0.066	0.068			
6F	0.009	0.008	0.010	0.012	0.067		
6G	0.010	0.004	0.013	0.017	0.067	0.009	
wzg	0.013	0	0	0	0	--	--
wzh	0.008	0	0	0	0	--	--
wzd	0.001	0	0.001	0	0	--	--
wze	0.004	0	0.009	0	0	--	--
wchA	0.004	0	0.007	0	0.001	--	--
wciNα	0.001	0	--	--	0	--	--
wciNβ	--	--	0.002	0	--	--	--
wciO	0	0	0.006	0	0	--	--
wciP	0	0	0.002	0	0	--	--
wzy	0.001	0	0.004	0	0	--	--
wzx	0.001	0	0.001	0	0	--	--
rmlA	0.046	0	0.062	0	0	--	--
rmlC	0.005	0	0.002	0	0	--	--
rmlB	0.016	0	0.014	0	0	--	--
rmlD	0.005	0	0.006	0	0	--	--

a) Pairwise comparisons were made between the serogroup 6 references using 13,416 bp *cps* locus sequences, which spanned the *cps* locus from the start of *wzg* through the end of *rmlD*, but excluded *wciN*, *HG262* and *HG263* from the analyses (see Methods and Figure 1); b) Median pairwise distance measures for each *cps* locus gene, estimated using the entire 974 pneumococcal genome dataset but stratified by serotype. No serotype 6F or 6G pneumococci were identified among the study genomes.

Genetics and phylogeny of the *cps* locus.

Thirteen *cps* locus genes were common to all 974 serogroup 6 pneumococci at a similarity threshold of >80% (Figure 4.4) and gene synteny was preserved among all the common genes. Pairwise distance measures (p-distances) of nucleotide variation between the *cps* loci of the seven reference strains were assessed and notably, serotype 6E was 6.7% divergent (p-distance range 0.065-0.068) from all other serotypes, whereas p-distances between all other serotypes were 0.4-1.7% (Table 4.3). P-distances were also calculated individually for each of the 13 common genes using the entire 974 genome dataset stratified by serotype, and the median p-distance value was used to provide a simple summary statistic of gene-specific sequence diversity within each serotype (Table 4.3). Most notably, the serotype 6B, 6D and 6E *cps* gene sequences were highly conserved within each serotype (median p-distance per gene = 0), whereas nearly all of the *cps* locus genes in the serotype 6A and 6C *cps* loci varied to some extent, with *wzg*, *rmlA* and *rmlB* being the most diverse (in addition to *wciN*), as noted in a previous study [329].

A phylogenetic tree was constructed using concatenated sequences of the 13 common *cps* locus genes for each of the 974 genomes, and major serotype clusters were clearly delineated (Figure 4.5). There were three major serotype 6A clusters and one cluster each for serotypes 6B, 6C and 6D. No serotype 6F or 6G pneumococci were identified, but the serotype 6F/6G reference sequences were within serotype 6A clusters. This was consistent with the initial report describing these new serotypes as being nearly identical to the serotype 6A *cps* sequence [110]. The serotype 6E pneumococci clustered in one group with minor within-cluster genetic variation, apart from a few pneumococci from South Africa in 1984-85 and Massachusetts in 2004 (long purple branches) that had switched serotypes and are discussed below, and three Thai pneumococci at the tips of

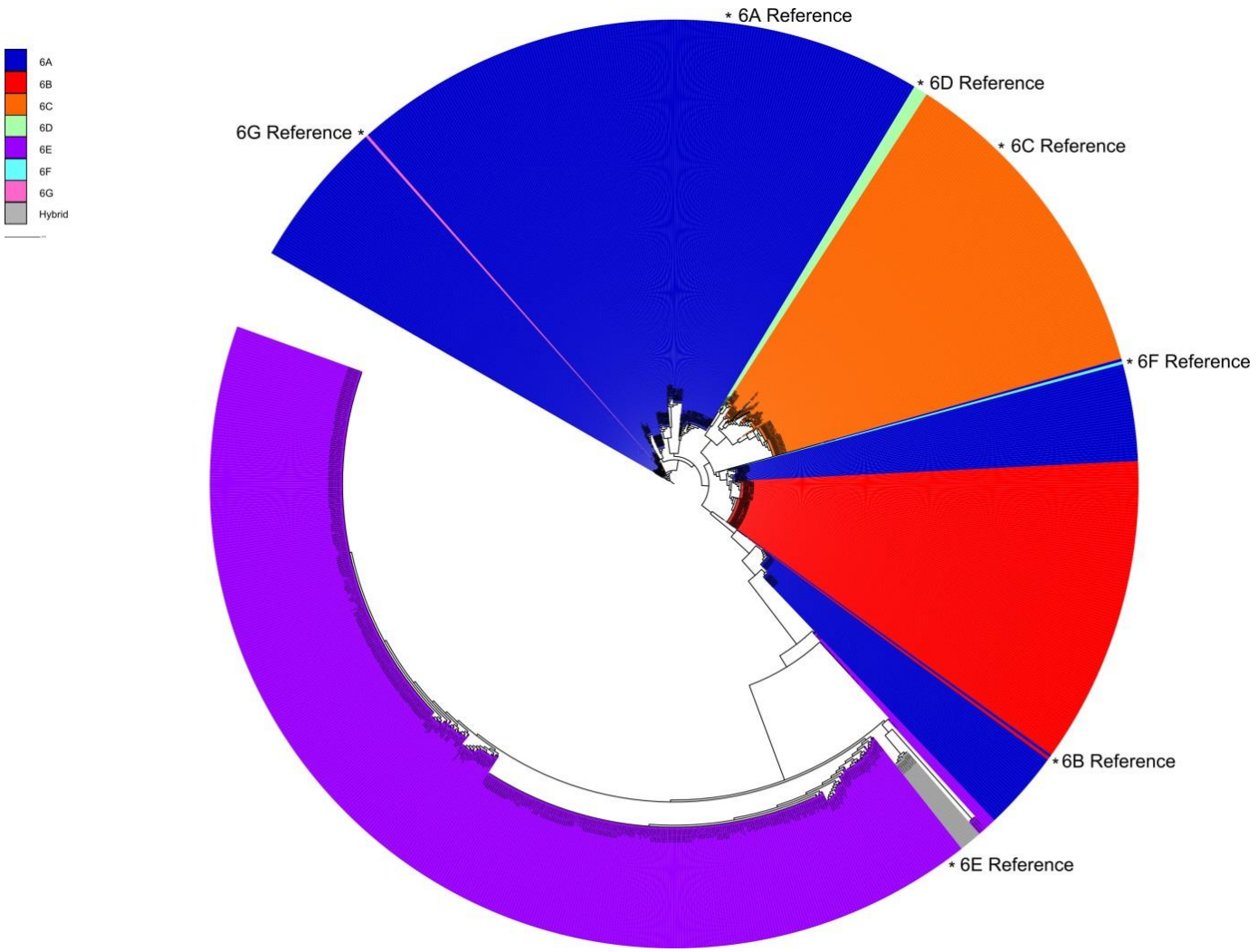


Figure 4.5. Phylogenetic tree depicting the relationships between the concatenated sequences of 13 common *cps* locus genes (12.3 Kb) among 974 serogroup 6 pneumococcal genomes.

very short branches that predominantly possess serotype 6E *cps* locus sequence but also have sequence regions that match serotype 6C or 6D sequence.There was also a small cluster (coloured grey) of eight pneumococci collected in Thailand from 2004-2006. This small cluster represented a hybrid serotype 6C/6E *cps* locus, with sequence differences as shown in Figure 4.6. The hybrid failed to type properly as either serotype 6C or 6E in the serotyping pipeline because it possessed a *wciNα* like serotype 6E but the *wzy* amino acid that differentiates serotype 6C (Figure 4.6A). The reason for the serotyping failure was clearly apparent when the sequences for each of the *cps* locus genes were examined (Figure 4.6B). The hybrid predominantly matches the
serotype 6E sequence, but contains a region of mainly serotype 6C-like sequence from *wciP* through roughly the first half of *wzx*.

Molecular epidemiology of serogroup 6 pneumococci.

A phylogenetic tree was constructed based upon 432 concatenated gene sequences (292 kb) present in full coding length in all 974 pneumococci (Figure 4.7). The tree was annotated with clonal complex (CC) and serotype data, which clearly defined the major CCs (those with ≥12 members) within the dataset. The serotype data, represented by the outer coloured ring, demonstrated which serotypes were associated with each CC and where serotype switching had occurred within CCs.

Serotype 6E pneumococci were associated with 20 CCs (43 sequence types (STs)) in the study dataset, although 89% of the serotype 6E pneumococci were members of one of four lineages: CC90 (n = 203); CC315 (n = 117); CC4405 (n = 55); and CC273 (n = 20), as shown in Figure 4.7 and Supplementary File 4.1. CC90, CC315 and CC273 are multidrug-resistant lineages associated with PMEN2, PMEN20 (Poland^{6B}-20) and

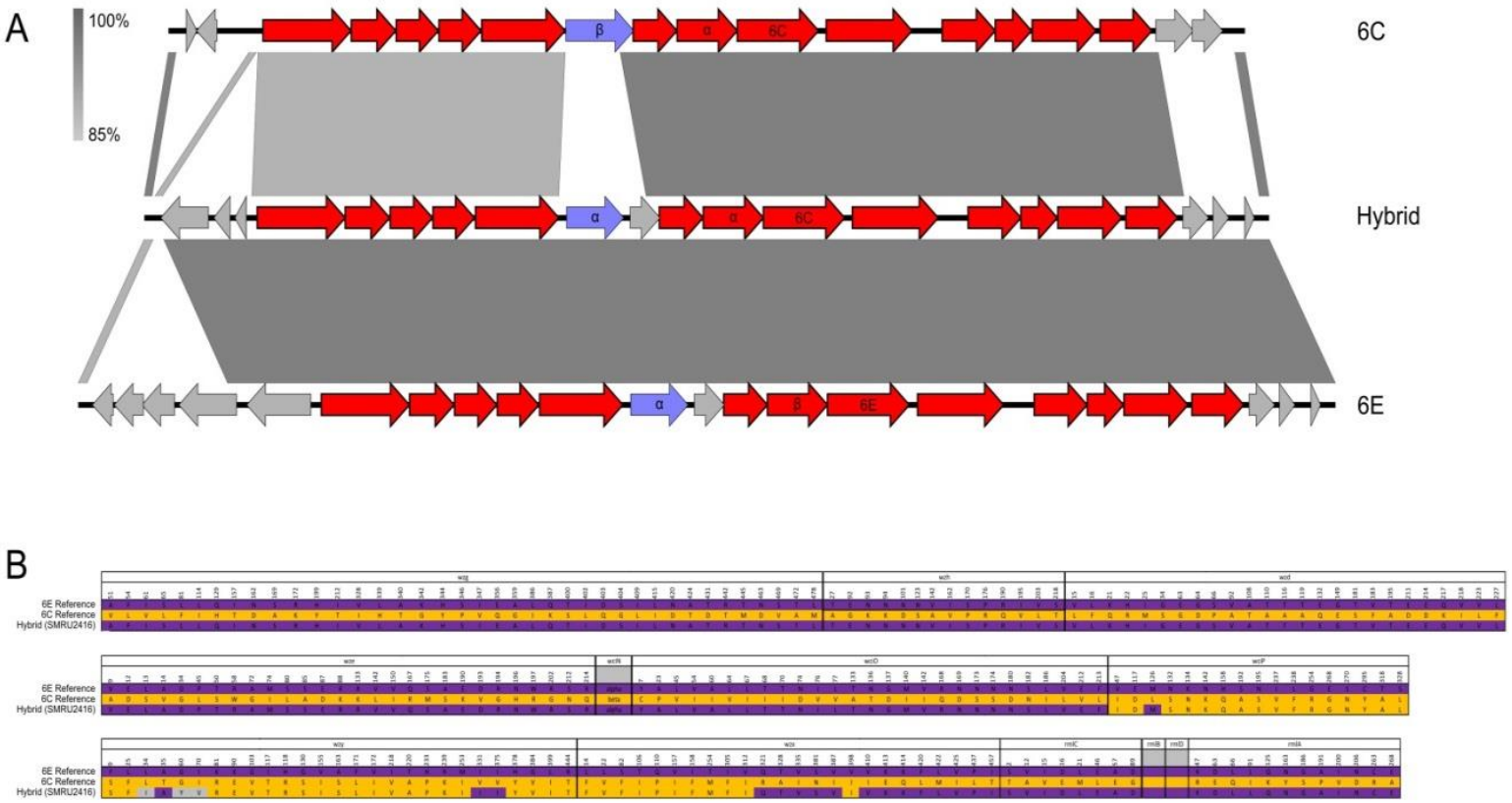


Figure 4.6 Illustration of the genetic organisation, nucleotide similarity and variable amino acids of the hybrid serotype 6C/6E *cps* locus sequence. A) Comparison of the hybrid *cps* locus sequence to the reference serotype 6C and 6E sequences. The results of pairwise BLAST nucleotide sequence comparisons are shown; darker grey colour highlights greater sequence conservation between the pair of sequences. B) Variable amino acid residues identified between the serotype 6C, 6E and hybrid sequences. The position of each variable residue in the corresponding gene sequence was indicated by the number above the residue. The residues associated with serotypes 6E and 6C were coloured purple and orange, respectively, and those not found in either 6E or 6C were coloured grey. Note that the *wciN* allele was simply indicated as α or β, and that *rmlB* and *rmlD* were identical across all three sequences.

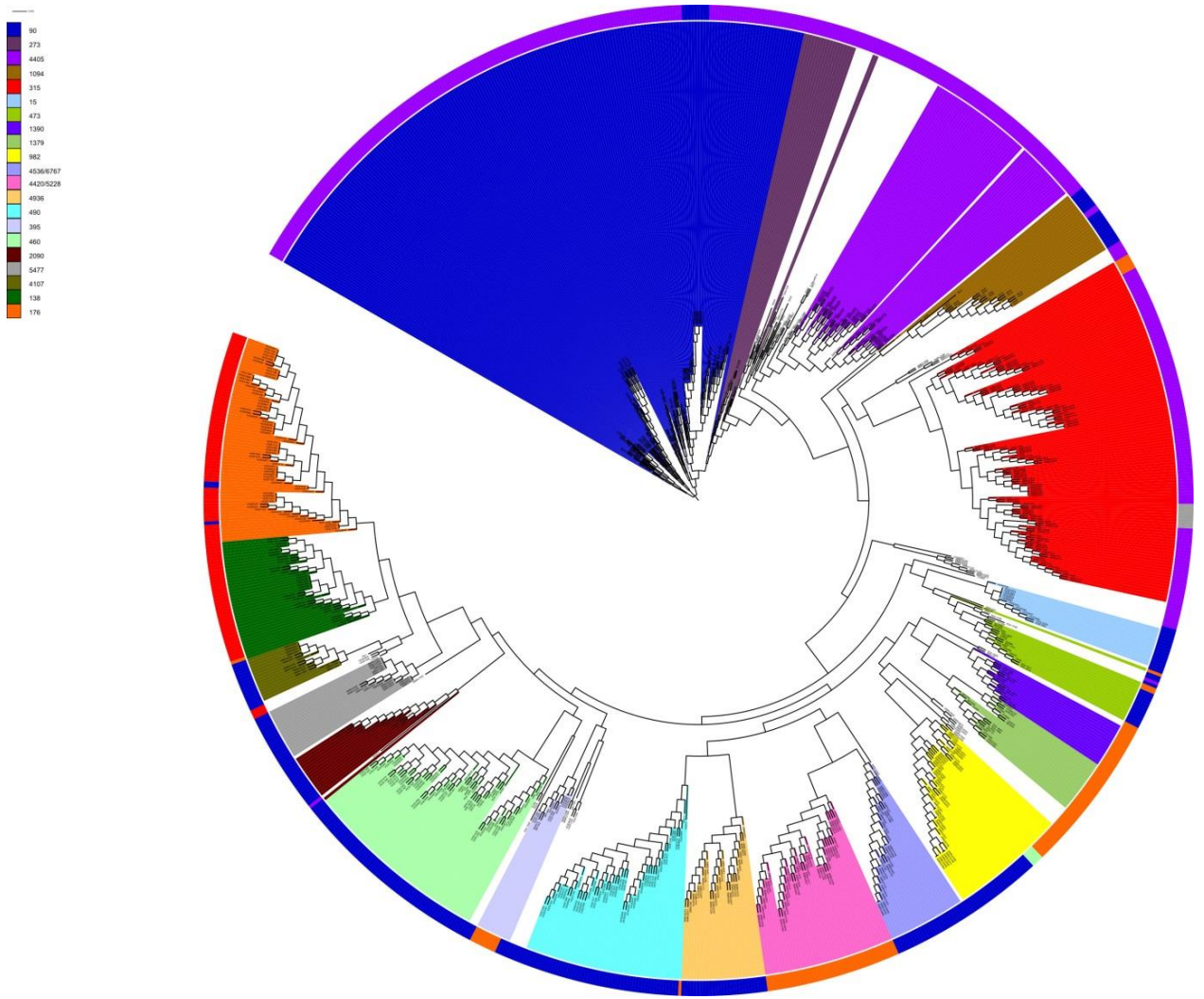


Figure 4.7. **Phylogenetic tree describing genome-wide relationships among serogroup 6 pneumococci.** The tree was constructed using the concatenated sequences of 432 full-length coding loci found in all 974 genomes and annotated with clonal complex (CC) designations and serotype. CCs with ≥12 isolates were coloured as listed in the key. The outer ring indicates the serotype by colour: 6A, blue; 6B, red; 6C, orange; 6D, light green; 6E, purple; 6F, light blue; and 6G, pink.

PMEN22 (Greece^{6B}-22), respectively (Supplementary File 4.1; http://pubmlst.org/spneumoniae/). All 8 hybrid serotype 6C/6E genomes were ST315.

Notably, pneumococci of these four serotype 6E lineages have been detected in at least 32 different countries on six continents, as detailed in the PubMLST database. However, the majority of the isolates of these four lineages in PubMLST were submitted as serotype 6B (including PMEN20 and PMEN22): these should be considered putative serotype 6E rather than 6B, and thus the PubMLST database expands the wide distribution of serotype 6E pneumococci.

Serotype 6A pneumococci were members of 24 different CCs (45 STs) in the study dataset, of which 11 CCs captured 90% of the serotype 6A population (Supplementary File 4.1). Serotype 6C pneumococci were associated with 12 CCs, five of which defined 83% of the serotype 6C genomes. All but three strains of serotype 6B pneumococci were members of either CC176 (n = 66) or CC138 (n = 39), and all four serotype D genomes were ST4407.

Serotype switching among serogroup 6 pneumococci.

There was clear evidence for serotype switching (horizontal genetic exchange of all or part of the *cps* locus sequence, conferring a change in serotype) within 11 CCs, since pneumococci within the CC were not exclusively defined by a single serotype (Figures 4.7 and 4.8). Notably, CC315 was represented by three serotypes: 6E, 6C and the 6C/6E hybrid. CC90 was predominantly defined by serotype 6E pneumococci, except for 9 genomes from the Maela refugee camp that were ST5127 (a double-locus variant of ST90) and had a serotype 6A *cps* locus. CC1094 was a major South African lineage of serotype 6A, although three genomes isolated in the mid-1980s were serotype 6E.

The *cps* locus sequences of all 44 genomes associated with putative serotype switches were manually inspected and confirmed. All serotype switches were clearly evident from the sequence, either by an exchange of the entire *cps* locus or by mosaic patterns of DNA sequence fragments indicative of recombination events within the *cps* locus, as shown in Figure 4.9 [168, 170, 345-348]. No other hybrid *cps* loci were identified apart from the 6C/6E described above.

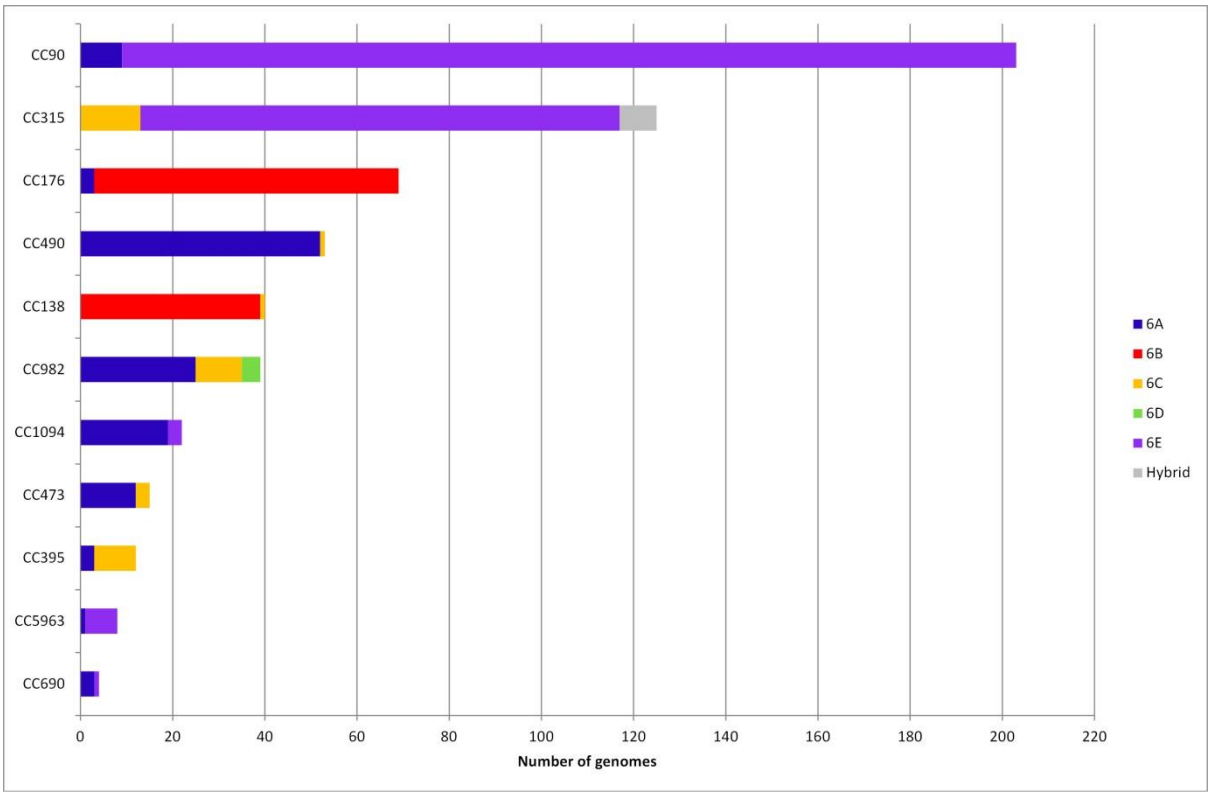


Figure 4.8. Clonal complexes (CCs) identified among the 974 pneumococcal genomes that were not of a single serotype.

Genome-wide diversity among serotypes.

Finally, all 974 genomes were compared to the PMEN2 genome reference using Genome Comparator, to investigate the presence/absence and diversity of 2352 genes across each genome. These data were depicted as a pseudo-heat map, ordered by serotype and CC

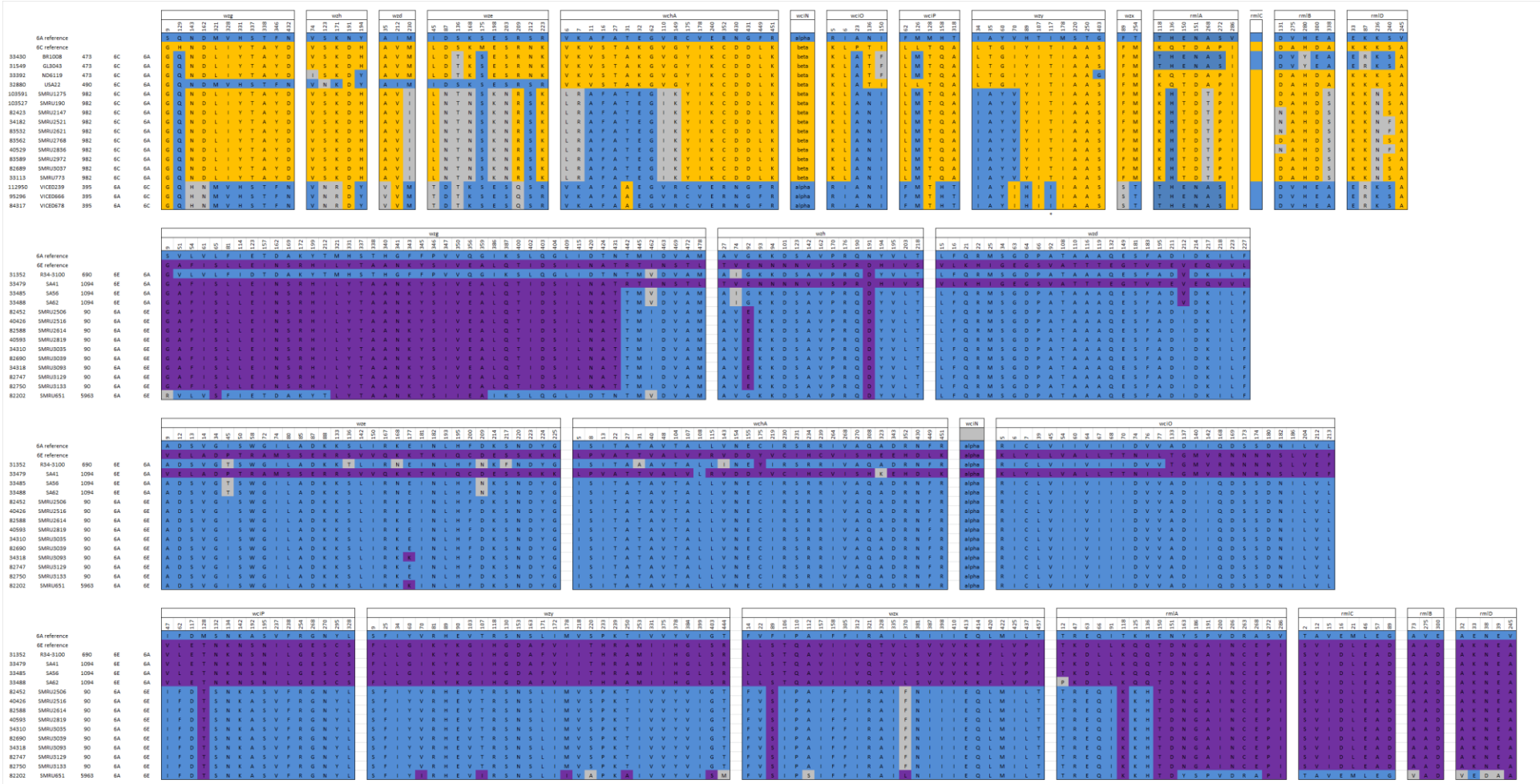


Figure 4.9a. Gene-by-gene depiction of the variable amino acids for all pneumococci that demonstrated evidence for capsular switching. The appropriate serotype reference strains for each comparison are at the top of each section (see Table 2 for accession numbers). Residues were coloured by serotype: 6A (blue), 6B (red), 6C (orange), 6D (pale green), and other (grey). Numbers above each variable residue mark the location of that residue in that gene.

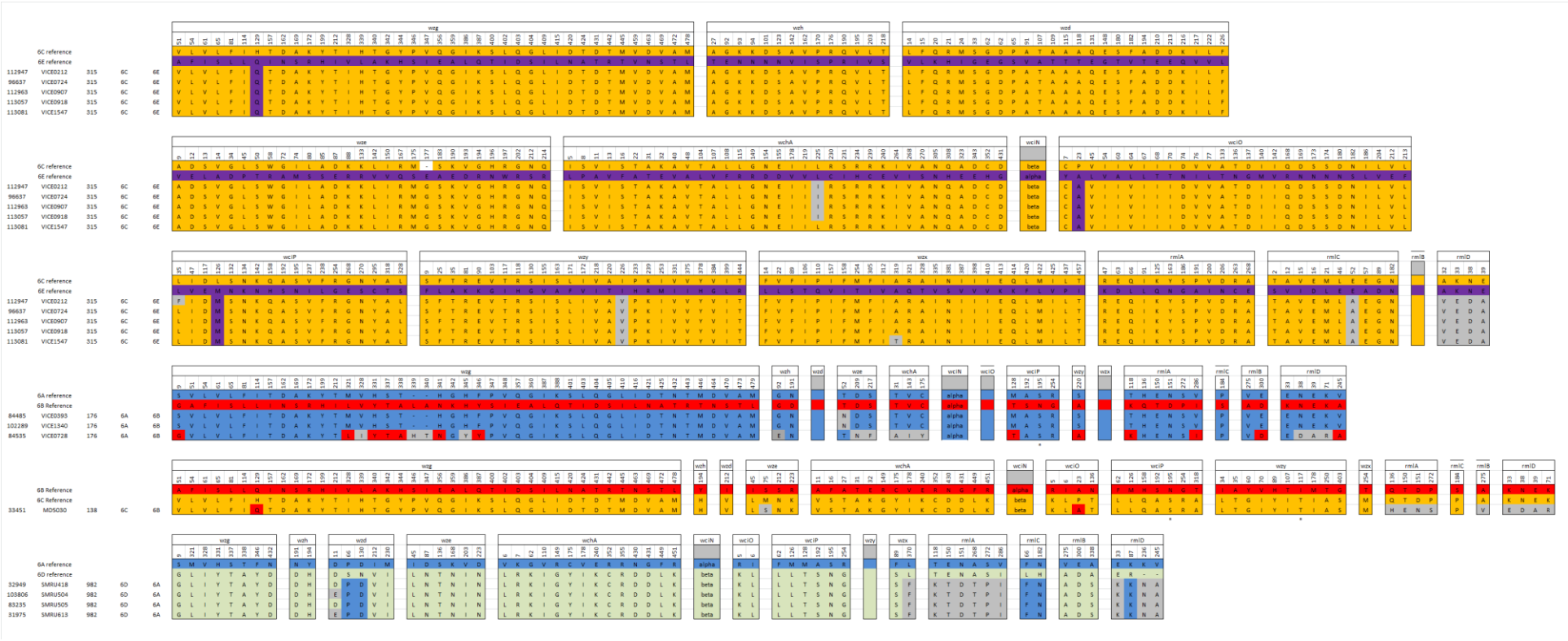


Figure 4.9b. Gene-by-gene depiction of the variable amino acids for all pneumococci that demonstrated evidence for capsular switching. The appropriate serotype reference strains for each comparison are at the top of each section (see Table 2 for accession numbers). Residues were coloured by serotype: 6A (blue), 6B (red), 6C (orange), 6D (pale green), and other (grey). Numbers above each variable residue mark the location of that residue in that gene.

(Figure 4.10). A number of observations were immediately apparent. As expected, the genomes of CC90^{6E} (CC^{serotype}) pneumococci were very similar to the PMEN2 reference (ST90^{6E}) genome used for comparison (largest white horizontal band). However, they were not identical – the bacteriophage sequences in the PMEN2 reference were either not present or of a different sequence in the CC90^{6E} study pneumococci, and there was a region of variable genes (blue) in the latter half of the genome in addition to several smaller variable regions in some genomes within CC90^{6E}. The serotype switch CC90^{6A} pneumococci were highly similar across the genome to the CC90^{6E} pneumococci, although CC90^{6A} genomes also did not have the PMEN2-like bacteriophages and smaller variable regions were identified. The STs within CC273^{6E}, CC4405^{6E} and CC490^{6A} shared some MLST alleles (5, 1-4, and 1, respectively), with the STs in CC90, but all possessed genes across the genome that matched identically to CC90^{6E} (genes coloured white). In contrast, across the genome the CC315^{6E} genes were mainly of different alleles (i.e. genes mainly coloured blue). Future studies will use these genome-wide data to investigate whether specific genotypic differences among serogroup 6 lineages relate to phenotypic differences between lineages and/or serotypes.

Vaccine-induced inhibition of serotype 6E.

Finally, a key question was whether or not PCVs would provide immunological protection against serotype 6E pneumococci. Stored sera from infants in the United Kingdom, previously vaccinated with either PCV7 or PCV13, were available and used to test for killing of serotype 6E pneumococci. Five strains were tested and results are shown in Table 4.4. Antibodies induced by both PCV7 and PCV13 killed the five strains tested. Removal of antibodies specific for 6B polysaccharide abolished the killing completely in most sera tested. Removal of 6A antibodies varied by serum analysed but in general was similar irrespective of the type of vaccine used (6B alone in PCV7 or both

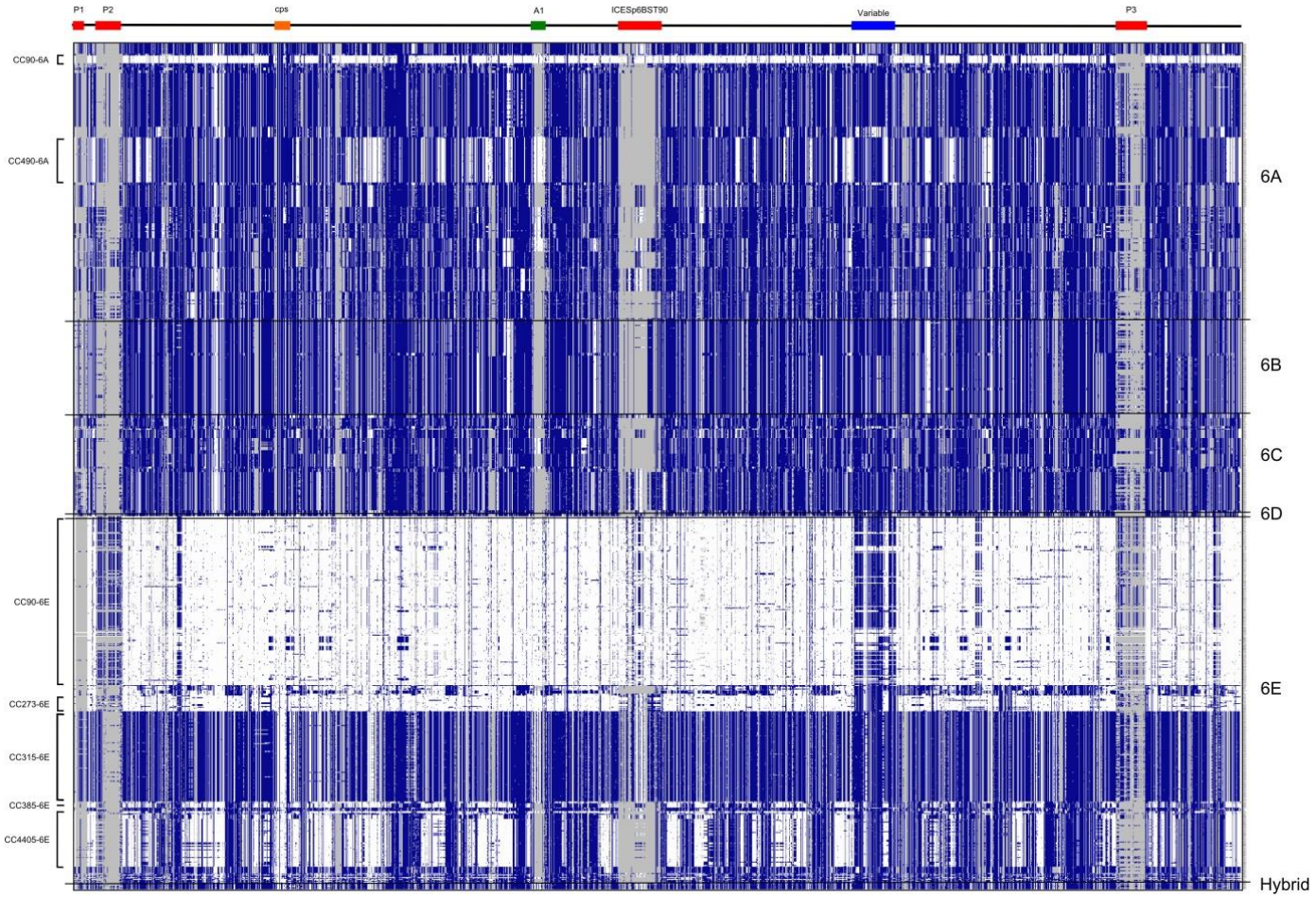


Figure 4.10. Visual representation of the Genome Comparator output for all 974 genomes as a pseudo-heat map. Each genome was depicted horizontally and in the gene order defined by the reference PMEN2 genome sequence, which has 2352 coding sequences (genes). Colours indicate gene-by-gene presence/absence and sequence similarity for each query genome as compared to the PMEN2 reference: grey = gene not present in the query genome; white = gene was present in the query genome and had an identical sequence to that of the reference genome; and blue = gene was present in the query sequence but the sequence was not identical to that of the reference. Several regions of the PMEN2 reference genome were highlighted: P1 = phage remnant; P2 = 11865-like phage; P3 = 2167-like phage; cps = capsular locus; A1 = ATP-synthase operon; ICESp6BST90 = ICE element; and Variable = variable region of gene sequences in CC90-6E.

Table 4.4 (part 1). Results of serological assays using paediatric sera collected after primary PCV7 and PCV13[a] immunisation to inhibit serotype 6E pneumococci.

Serotype 6E strain	Vaccine	Sample	Titer without competitor	Titer with 6A PnPs[b]	% inhibition	Titer with 6B PnPs[c]	% inhibition	Titer with 6C PnPs	% inhibition
PMEN 2	PCV7	209(6)	266	na[d]	na	<8	98.5	298	-12.0
		214(6)	165	117	29.1	<8	97.6	144	12.7
		190(6)	3462	3738	-8.0	<8	99.9	2999	13.4
		154(6)	2005	1361	32.1	<8	99.8	1397	30.3
		042(7)	213	70	67.1	<8	98.1	258	-21.1
	PCV13	010A	766	708	7.6	<8	99.5	684	10.7
		208A	261	147	43.7	<8	98.5	193	26.1
		226A	1995	1684	15.6	<8	99.8	1337	33.0
		240A	2030	1760	13.3	496	75.6	1835	9.6
		243A	3069	433	85.9	na	na	2473	19.4
PMEN 8	PCV7	209(6)	246	na	na	<8	98.4	237	3.7
		214(6)	235	221	6.0	<8	98.3	237	-0.9
		190(6)	2382	3811	-60.0	<8	99.8	2692	-13.0
		154(6)	2151	1755	18.4	<8	99.8	1894	11.9
		042(7)	477	90	81.1	<8	99.2	585	-22.6
	PCV13	010A	484	536	-10.7	<8	99.2	456	5.8
		208A	111	72	35.1	<8	96.4	75	32.4
		226A	1677	1446	13.8	<8	99.8	1245	25.8
		240A	1979	1803	8.9	533	73.1	2153	-8.8
		243A	2769	328	88.2	<8	99.9	2117	23.5
VICE 0629	PCV7	209(6)	na	na	na	na	na	na	na
		214(6)	180	54	70.0	<8	97.8	112	37.8
		190(6)	2024	1877	7.3	<8	99.8	1858	8.2
		154(6)	923	790	14.4	<8	99.6	703	23.8
		042(7)	<8	<8	0.0	<8	0.0	<8	0.0

Table 4.4 (part 2). Results of serological assays using paediatric sera collected after primary PCV7 and PCV13[a] immunisation to inhibit serotype 6E pneumococci.

VICE 0629	PCV13	010A	420	199	52.6	<8	99.0	193	54.0
		208A	na	na	na	na	na	na	na
		226A	681	509	25.3	<8	99.4	566	16.9
		240A	1316	1011	23.2	249	81.1	977	25.8
		243A	1419	117	91.8	<8	99.7	1109	21.8
VICE 1004	PCV7	209(6)	198	na	na	<8	98.0	194	2.0
		214(6)	359	334	7.0	<8	98.9	333	7.2
		190(6)	2514	2424	3.6	<8	99.8	2238	11.0
		154(6)	863	1141	-32.2	<8	99.5	762	11.7
		042(7)	233	76	67.4	<8	98.3	239	-2.6
	PCV13	010A	421	488	-15.9	<8	99.0	461	-9.5
		208A	233	100	57.1	<8	98.3	148	36.5
		226A	2626	241	90.8	na	na	1855	29.4
		240A	1791	1394	22.2	<8	99.8	1257	29.8
		243A	2099	1701	19.0	371	82.3	1739	17.2
VICE 1150a	PCV7	209(6)	230	192	16.5	<8	98.3	247	-7.4
		214(6)	367	336	8.4	<8	98.9	416	-13.4
		190(6)	2426	2418	0.3	<8	99.8	2368	2.4
		154(6)	1259	1066	15.3	<8	99.7	1470	-16.8
		042(7)	500	161	67.8	<8	99.2	390	22.0
	PCV13	010A	540	293	45.7	<8	99.3	304	43.7
		208A	145	106	26.9	<8	97.2	92	36.6
		226A	1674	1173	29.9	<8	99.8	879	47.5
		240A	1602	931	41.9	238	85.1	944	41.1
		243A	2324	301	87.0	<8	99.8	1849	20.4

a) PCV7 and PCV13: 7-valent and 13-valent pneumococcal conjugate vaccines, respectively; b)PnPs: pneumococcal polysaccharide tested at 1 µg/ml; c) Titer of 4 used to calculate % inhibition for results <8; na: not available due to technical failure.

6B and 6A in PCV13) with only partial inhibition of killing demonstrated, findings consistent with a previous study of 6A inhibition of 6B killing [323]. Anti-6C antibody removal also varied depending on the serum used but in general had little effect on killing.

4.5 Discussion

This is the first in-depth large scale interrogation of the molecular epidemiology of serogroup 6 strains and it illustrates that serotype 6E pneumococci have been circulating for at least 33 years, preceding PCV introduction by nearly two decades. This study revealed that 43% of the genome collection (previously thought to predominantly contain serotypes 6A, 6B and 6C) in fact represented serotype 6E pneumococci of several major genetic lineages, three of which were multidrug-resistant PMEN lineages. They were distributed across 15 countries and 5 continents, and the pneumococcal PubMLST database provides evidence for an even wider geographical distribution. Serotype 6E pneumococci caused a range of diseases among all age groups, but were also frequently recovered from healthy young children. This study identified several major genetic clusters of serotype 6A *cps* locus sequences, discovered a new hybrid 6C/6E serotype and revealed many examples of serotype switching involving serotypes 6A-6E. Importantly, serological assays demonstrated that vaccine-induced serotype 6B antibodies were able to kill serotype 6E pneumococci.

For several decades the existence of serotype 6E pneumococci was obscured because they cross-react to the serotype 6B antisera used in the Quellung reaction, and it was only by inspection of the *cps* locus sequences that the existence of serotype 6E was realised. Initially serotype 6E were recognised by several research groups analysing key regions of gene sequences within small collections of isolates, and now the high prevalence and

worldwide distribution of 6E has been revealed unequivocally here by the interrogation of a global and historical collection of genome sequences. Basically, the majority of what for many years were thought to be serotype 6B isolates were in fact serotype 6E pneumococci. 'True' serotype 6B pneumococci were also identified in this study, but they were mainly of two genetic lineages, CC138 and CC176, both of which have also been detected in many countries around the world (http://pubmlst.org/spneumoniae/).

This study revealed the sequence diversity among serogroup 6 *cps* loci and of particular note was that there were three major serotype 6A *cps* locus sequence clusters. Do the sequence-based changes in the 6A *cps* locus result in changes to the polysaccharide and if so, do PCV-induced 6A antibodies differentially protect against alternative versions of serotype 6A polysaccharide? This is currently unknown but warrants further investigation. Moreover, in this study a hybrid 6C/6E serotype was discovered that is sufficiently divergent to presumptively consider it yet another serotype, as well as evidence for many distinct serotype switches among the genetic lineages. Yet it is important to recognise that serotype switching and the creation of genetic variants at the *cps* locus appears to be a normal biological process among pneumococci and is not a direct consequence of vaccine use [348]. However, vaccine-induced immune pressure does alter the pneumococcal population structure, which can select for the emergence of new genetic variants. The earliest reported evidence for vaccine escape pneumococci was the result of such a scenario [168, 170, 349]. PCVs perturb the pneumococcal population with unpredictable consequences; therefore, the importance of genomics and molecular epidemiology in any pneumococcal surveillance programme cannot be underestimated.

To date, the biochemical structure of serotype 6E polysaccharide is not yet known. It may be that the structures of the serogroup 6B and 6E polysaccharides are similar enough

to explain the inhibition of serotype 6E pneumococci by PCV-induced 6B antibodies, even though at a sequence level the *cps* loci are very different. However, Oliver et al recently reported a detailed investigation of the new serotypes 6F and 6G, and experimentally showed that single changes in the amino acid sequence of *wciNα* resulted in changes in the repeating units of the capsular polysaccharides, with concomitant changes in serological profiles. *wciNα* and *wciNβ* are highly divergent versions of *wciN*: serotypes 6A, 6B, 6E, 6F and 6G possess *wciNα* (encoding a galactosyltransferase), whilst serotypes 6C and 6D have the *wciNβ* allele (encoding a glucosyltransferase). The authors concluded that small changes in the sequence not only resulted in new capsular types, but they also posited that these changes could confer immunological changes in the human host response [110]. A detailed comparison of serogroup 6 polysaccharide biochemistry should be investigated as a matter of priority, to understand the biochemical structure of the polysaccharides in the context of the observed serotype-specific epidemiology and killing by serotype 6B antibodies.

The majority of these serotype 6E study isolates were collected pre-PCV, and this study showed that PCV7 and PCV13-induced antibodies to serotype 6B were protective. Therefore, the overall prevalence of 6E should be significantly reduced post-PCV implementation and the post-PCV vaccine impact studies from many countries have demonstrated a significant reduction in the prevalence of serotype 6B and near elimination of 6B carriage [319, 350]. This suggests two possibilities: (i) PCV7 and PCV13, which contain serotype 6B polysaccharides, inhibit serotype 6B and sufficiently cross-protect against serotype 6E at a population level; or (ii) that the serotype '6B' polysaccharides in the vaccines are actually serotype 6E polysaccharides. This study was unable to identify which strain(s) was specifically used to produce the serotype 6B

polysaccharides used in the PCVs, in order to confirm the serotype based on the *cps* locus sequence.

Whether or not current PCVs are in fact serotype 6E vaccines remains an open but important question. The overwhelming success of PCVs in reducing serotype 6B (6E) disease suggests that perhaps the question is purely an academic one; however, it would be relevant in cases of vaccine failures where there may be discordance between the vaccine serotype and the serotypes of pneumococci associated with vaccine failures. One clue to a mismatch might be if pneumococci responsible for 6B vaccine failures were ST138 or ST176 lineages, as these appear to be the predominant (true) serotype 6B lineages that circulate worldwide. Determining country-specific estimates of the prevalence of serotype 6E pre- and post-PCV implementation will be essential to assessing whether at a population level PCVs are protective against serotype 6E.

Genomics has revolutionised microbiological research and this study reinforces just how influential the change has been and will continue to be in the future. It is now relatively simple and cost-effective to use next generation sequencing to obtain the (nearly) complete bacterial genome sequence, and the early challenges of contig assembly, availability of databases in which to store and query genome sequences, and the development of tools to analyse large scale databases are being overcome. There are currently >10,000 pneumococcal genome sequences available in public databases and other genome sequencing projects are underway. Challenges remain, including making published genome assemblies widely accessible to all users, but the genomics field is moving apace.

4.6 Supplementary information (see attached USB drive)

Supplementary File 4.1. Study datasets.

Supplementary File 4.2. Genome Comparator output.

5. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5,405 pneumococcal genomes

These data were presented as an oral presentation and a poster at the 10th International Symposium on Pneumococci and Pneumococcal Diseases (ISPPD), Glasgow in June 2016.

This chapter was published as:

van Tonder, A.J., Bray, J.E., Quirk, S.J., Haraldsson, G., Jolley, K.A., Maiden, M.C.J., Hoffmann, S., Bentley, S.D., Haraldsson, A., Erlendsdóttir, H., Kristinsson, K.G. and Brueggemann, A.B. 2016. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5,405 pneumococcal genomes. *Microbial Genomics*. doi: 10.1099/mgen.0.000090

The genomes from the VICE study were prepared as described in the General Methods chapter and Chapter 4. Steen Hoffman provided pneumococci from a historical collection; DNA extractions were performed by Caroline Harrold before whole genome sequencing was performed at the Wellcome Trust Centre for Human Genetics (WTCHG) at the University of Oxford. The genomes were then assembled and improved using the WTCHG's automated pipeline before being uploaded to the Brueggemann BIGSdb database by Caroline Harrold. Genomes from previously published datasets were downloaded from the rMLST database by James Bray (see Chapter 4).

5.1 Abstract

Approximately 100 different serotypes are known, but the extent of sequence diversity within the *cps* locus of individual serotypes is not well understood. Investigating serotype-specific sequence variation is crucial to the design of sequence-based serotyping methodology, understanding pneumococcal conjugate vaccine (PCV) effectiveness, and the design of future PCVs. The analysis of serogroup 6 cps loci in Chapter 4 revealed a surprising amount of capsular diversity. The availability of large genome datasets makes it possible to assess population-level variation among pneumococcal serotypes and in this study 5,405 pneumococcal genomes were used to investigate *cps* locus diversity among 49 different serotypes. Pneumococci were recovered between 1916 and 2014 from people of all ages living in 51 countries. Serotypes were deduced bioinformatically, *cps* locus sequences were extracted and variation was assessed within the *cps* locus, in the context of pneumococcal genetic lineages. Overall, *cps* locus sequence diversity varied markedly: low to moderate diversity was revealed among serogroups/types 1, 3, 7, 9, 11 and 22; whereas serogroups/types 6, 19, 23, 14, 15, 18, 33 and 35 displayed high diversity. Putative novel and/or hybrid *cps* loci were identified among all serogroups/types apart from 1, 3 and 9. This study demonstrated that *cps* locus sequence diversity varied widely between serogroups/types. Investigation of the biochemical structure of the polysaccharide capsule of major variants, particularly PCV-related serotypes and those that appear to be novel or hybrids is warranted.

5.2 Introduction

The standard phenotypic methodologies for determining the pneumococcal serotype are the Quellung reaction and latex agglutination; however, PCR-based and microarray-based serotyping have increased in popularity in recent years [351]. Furthermore, there is

growing interest in deducing the serotype directly from the genome sequence, given that genome sequencing is now relatively straightforward and increasingly utilised. Genome sequence-based serotyping has the added advantage that putatively novel serotypes can be identified directly from the sequence, whereas all other methods result in a negative or equivocal test result that requires repeat testing. However, confirmation of putatively novel serotypes requires a biochemical assessment of the polysaccharide capsule to confirm novelty, which is not a trivial task.

Every serotyping method has its challenges, but for sequence-based methods a key consideration is the extent of sequence diversity within the *cps* locus (*e.g.* within the PCR primer-binding region) that might lead to erroneous results. Sequence-based methods are also reliant on having a representative reference set of *cps* loci with which to design the test method and ensure accurate serotype prediction. The first detailed genetic analysis of a single example of each of 90 pneumococcal serotypes was published in 2006, and since then new serotypes have been discovered [15, 107, 108]. Whether or not the original reference set of 90 *cps* loci are the best representatives of each serotype is unknown, even though they form the basis for the design of many sequence-based serotyping assays or algorithms.

A surprising amount of *cps* locus diversity was revealed among serogroup 6 pneumococci in Chapter 5. Therefore, in order to further assess pneumococcal serotype diversity, a large and diverse pneumococcal dataset of 5,405 genomes of 14 major serogroups/types, which corresponded to 49 different serotypes was compiled. Serogroups/types were selected based upon their overall prevalence, inclusion in current or future PCVs, and invasive disease potential [23]. The aim of the study was to assess the level of sequence diversity at the *cps* locus for each serogroup/type and identify any unusual serotypes within each serogroup, in the context of the pneumococcal genetic lineages.

5.3 Methods

Pneumococcal genome sequences and datasets

Table 5.1. Pneumococcal genomes included in *cps* locus analyses.

		Serogroup/type													
Genomes*	n	1	3	6	7	9	11	14	15	18	19	22	23	33	35
Thailand	1875	6	35	396	25	42	49	147	161	32	482	30	325	59	86
Iceland	1536	6	60	338	15	49	78	59	87	22	448	50	255	27	42
USA (MA)	540	0	11	97	14	10	49	4	82	5	104	20	73	5	66
UK	428	2	8	126	6	8	37	7	50	1	50	27	62	16	28
Serotype 1	317	317	0	0	0	0	0	0	0	0	0	0	0	0	0
PMEN2	172	0	0	172	0	0	0	0	0	0	0	0	0	0	0
PMEN1	126	0	0	2	0	0	0	0	0	0	0	0	124	0	0
Historical	100	5	6	2	11	12	5	8	5	7	15	3	9	6	6
Serotype 3	75	0	75	0	0	0	0	0	0	0	0	0	0	0	0
PMEN ref	21	4	0	5	1	1	0	3	0	1	3	0	3	0	0
Genbank	215	8	10	44	4	4	4	30	0	3	83	4	23	2	0
Total n	5405	348	205	1182	76	126	222	258	385	71	1185	130	874	115	228

* Pneumococcal genome studies from which genome sequences were obtained. Note: USA = United States of America; MA = Massachusetts; UK = United Kingdom; PMEN = Pneumococcal Molecular Epidemiology Network.

Datasets of assembled pneumococcal genomes (n=5,405; Table 5.1) from 14 serogroups/types were compiled using previously published genome datasets [97, 99, 100, 236, 247, 348, 352, 353], sequence data from GenBank (www.ncbi.nlm.nih.gov/genbank/), a new set of historical (1937-1996) pneumococcal genomes and new genome data from the VICE study [253]. The genome datasets included 21 Pneumococcal Molecular Epidemiology Network (PMEN) reference strains sequenced by the Brueggemann group or downloaded from GenBank (Appendix 5.1; [245]).

Table 5.2. Demographical and epidemiological information for serogroups/types with low to moderate diversity in the *cps* locus.

					Carriage/Disease			PCV status		
Serogroup/type		Years	Country (n)	Genomes (n)	Carr	Dis	Unk	Pre	Post	Unk
1	-	1943-2012	22	348	16	260	72	20	5	323
3	-	1961-2014	10	205	88	111	6	167	36	2
7	7A	1937	1	1	0	0	1	1	0	0
	7B	1952-2010	3	13	11	0	2	13	0	0
	7C	1939-2011	4	10	8	0	2	7	2	1
	7F	1962-2014	5	36	16	17	3	6	25	5
	7Hybrid	1952-2009	4	16	11	2	3	14	2	0
9	9A	1962-2009	2	3	0	2	1	2	1	0
	9L	1941-2010	3	11	8	0	3	11	0	0
	9Lii	2008-2010	1	7	7	0	0	7	0	0
	9N	1938-2014	5	36	24	7	5	18	14	4
	9V	1968-2014	6	58	30	25	3	44	11	3
	9Vii	2008-2011	2	11	11	0	0	10	0	1
11	11A	1939-2014	6	166	140	25	1	33	94	39
	11B	1940-2010	2	2	0	1	1	2	0	0
	11C	1957	1	1	0	0	1	1	0	0
	11D	1986	1	1	0	0	1	1	0	0
	11E	2012	1	1	0	1	0	0	1	0
	11F	1952	1	1	0	0	1	1	0	0
	11Hybrid	2001	1	1	1	0	0	1	0	0
	11X	2008-2010	1	49	49	0	0	49	0	0
22	22A	1939-2009	2	19	17	0	2	19	0	0
	22F	1940-2014	4	98	75	22	1	7	64	27
	22Hybrid	2008-2010	1	13	13	0	0	13	0	0

Genome sequences downloaded from GenBank were assembled by the original submitter. All other genome sequences were downloaded as raw sequence reads from the European Nucleotide Archive (http://www.ebi.ac.uk/ena), assembled using Velvet [249], assessed for quality and deposited in the rMLST database, which is publicly available and powered using BIGSdb (http://pubmlst.org/rmlst/; [20, 250]). Metadata for the isolates were manually extracted from the original publications (Supplementary File 5.1). Summaries of the metadata, including demographics, carriage/disease status and whether the pneumococci were recovered pre- or post-PCV are included in Tables 5.2 and 5.3.

Table 5.3. Demographical and epidemiological information for serogroups/types with high diversity in the *cps* locus.

					Carriage/Disease			PCV status		
Serogroup/type		Years	Country (n)	Genomes (n)	Carr	Dis	Unk	Pre	Post	Unk
6	6A	1952-2014	5	283	212	68	3	140	110	33
	6Bi	1939-2014	4	171	136	34	1	79	49	43
	6Bii (6E)	1981-2014	16	420	230	164	26	374	37	9
	6C	2001-2014	3	106	88	18	0	8	56	42
	6D	2006-2011	2	5	5	0	0	4	0	1
	6F	2006-2011	1	1	1	0	0	0	0	1
	6Hybrid	1978-2014	5	196	180	13	3	166	21	9
19	19A	1939-2014	5	210	141	65	4	56	144	10
	19B	1971	1	1	0	0	1	1	0	0
	19C	1939	1	1	0	0	1	1	0	0
	19F	1952-2014	9	745	470	271	4	555	178	12
	19Hybrid	1952-2014	9	222	166	43	13	154	33	35
	19A/F	2012-2014	1	6	4	2	0	0	6	0
23	23A	1945-2014	4	84	70	13	1	17	52	15
	23B	1941-2014	4	76	64	11	1	2	54	20
	23Fi	1940-2014	22	375	158	171	46	253	82	40
	23Fii	1983-2010	2	49	48	0	1	48	1	0
	23Hybrid	1997-2011	5	261	256	4	1	248	13	0
	23X	1996-2014	5	29	26	2	1	17	10	2
14	14i	1952-2013	8	58	19	33	6	35	17	6
	14wciY-fs	1967-2013	9	48	15	30	3	32	11	5
	14ii	2008-2010	1	88	88	0	0	88	0	0
	14Hybrid	1939-2010	4	64	59	2	3	63	1	0
15	15A	1939-2014	5	62	59	2	1	53	3	6
	15BC	1939-2013	4	20	15	2	3	6	7	7
	15F	1963	1	1	0	0	1	1	0	0
	15Hybrid	2001-2014	4	210	183	27	0	41	132	37
	15X	2008-2010	1	92	92	0	0	92	0	0
18	18A	1952-2010	2	4	3	0	1	4	0	0
	18B	1941-2011	2	2	0	1	1	1	1	0
	18C	1939-2013	5	31	20	8	3	21	9	1
	18F	1961	1	1	0	0	1	1	0	0
	18Hybrid	1945-2010	4	33	32	0	1	33	0	0
33	33A	1937-2011	3	4	1	0	3	3	0	1
	33B	1962	1	1	0	0	1	1	0	0
	33C	2007-2010	1	11	11	0	0	11	0	0
	33D	1979	1	2	0	0	2	2	0	0
	33F	1999-2014	3	40	23	17	0	11	20	9
	33Hybrid	2008-2014	3	52	52	0	0	43	3	6
	33X	2008-2010	1	5	5	0	0	5	0	0
35	35A	1939-2010	2	11	10	0	1	11	0	0
	35B	1939-2014	5	108	97	9	2	42	61	5
	35C	1941-2009	2	8	6	0	2	8	0	0
	35F	1939-2014	5	73	63	9	1	18	32	23
	35Hybrid	2007-2010	2	28	28	0	0	25	3	0

The serogroup/types selected for investigation in this study were 1, 3, 6, 7, 9, 11, 14, 15, 18, 19, 22, 23, 33 and 35, corresponding to 49 serotypes in total. Each genome sequence was screened against the relevant *cps* reference for the stated serotype and any genomes without full-length *cps* sequences were removed from further analyses. Genome sequences and associated metadata for all the pneumococci analysed in this study were deposited in the pneumococcal PubMLST website (http://pubmlst.org/spneumoniae/).

Sequence-based serotyping

With the exception of serotypes 3 and 37 (although serotype 37 was not investigated here), the polysaccharide capsule in pneumococcus is synthesized by the Wzy-dependent pathway. All Wzy-dependent serotypes have a *cps* locus located between *dexB* and *aliA*. Four conserved genes, *wzg*, *wzh*, *wzd* and *wze*, (also known as *cpsA-D*) at the start of the *cps* locus are responsible for the regulation of capsular synthesis while the remaining genes in the *cps* locus are serotype-specific and are involved in the production of the polysaccharide capsule. The serotype 3 *cps* locus shares a similar structure except that *wzg*, *wzh* and *wze* are truncated or frame-shifted [354].

Serotypes for each of the pneumococci included in the study were assigned based on the genome sequence, using a bespoke serotyping script called seqSerotyper.R (https://github.com/avantonder/seqSerotyper). The genome sequences were screened against published serotype reference sequences downloaded from GenBank, including the recently identified serotypes 6F, 6G and 11E (Appendix 2). Following the initial screen, where necessary, further differentiation of serotypes within serogroups was achieved

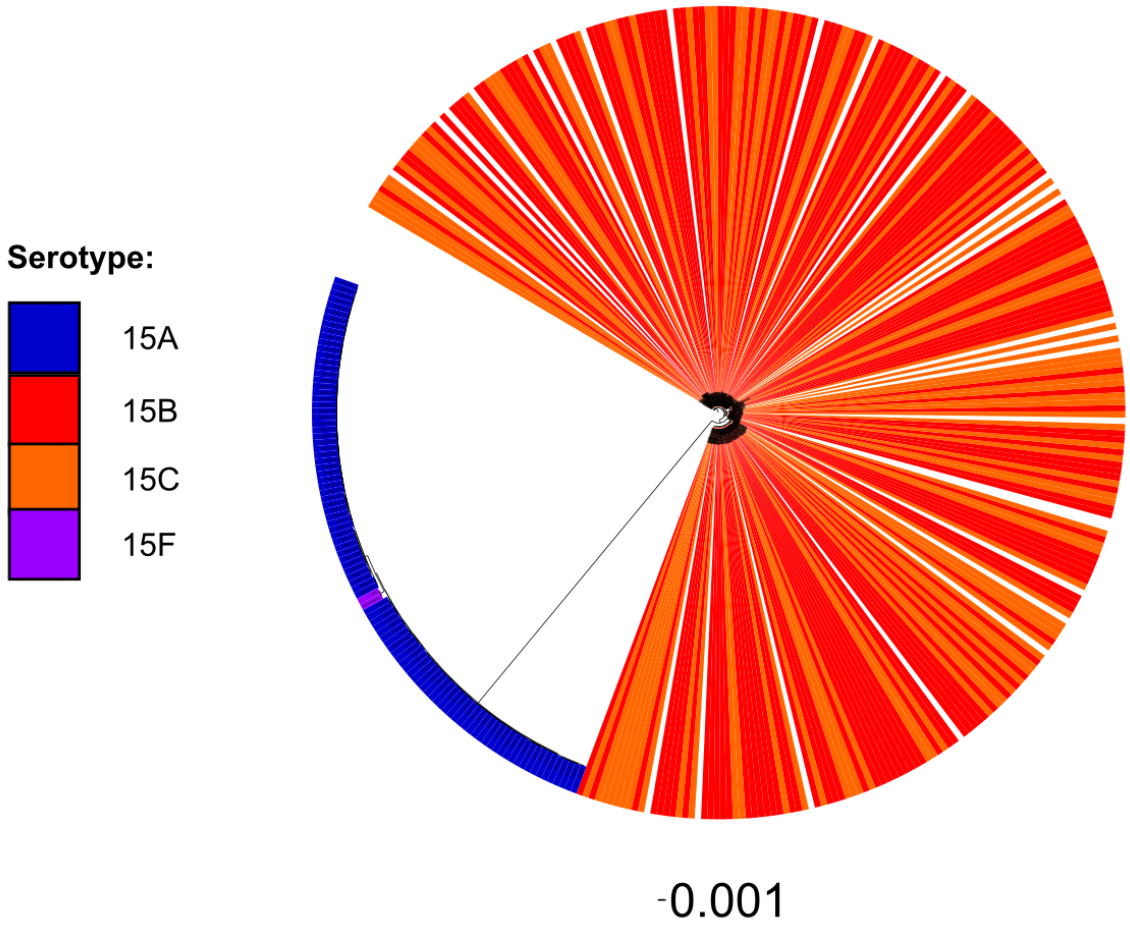


Figure 5.1 Serotypes assigned to serogroup 15 genomes using a previously published method of counting the number of TA repeats in the *wciZ* gene of serotypes 15B and 15C.

using serotype-specific markers such as key amino acid residues and/or alleles unique to a particular serotype.

Serotypes 15B and 15C are difficult to accurately differentiate using phenotypic methods. A published method for separating serotypes 15B and 15C based upon counting the number of TA repeats in *wciZ* was attempted [130]; however, this methodology failed to unambiguously resolve the two serotypes (Figure 5.1) and thus any pneumococci with a 15B- or 15C-like sequence were reported here as 15B/C.

Extraction and analysis of the *cps* locus sequences within each serogroup/type

For serotypes 1, 3 and 14 the *cps* locus sequences were extracted from the corresponding pneumococcal genome sequences and aligned using Muscle, excluding frameshifted genes and transposons [355]. To investigate the serogroups that contained multiple serotypes (6, 7, 9, 11, 15, 18, 19, 22, 23, 33 and 35), within each individual serogroup cd-hit was used to identify CDS that were common to all serotypes in that serogroup at a sequence identity threshold of ≥70%; ([265]; Appendix 5.3). A representative sequence from one of the serotype references within the serogroup was then BLASTed against the pneumococcal genome dataset to extract sequences for each of the shared CDS of that serogroup [246]. These extracted *cps* sequences were aligned gene-by-gene using Muscle before being concatenated to obtain a *cps* locus alignment for each serogroup.

Phylogenetic trees were created from the respective concatenated alignments using FastTreeMP and a general time-reversible nucleotide model [267]. ClonalFrameML was implemented to create final phylogenetic trees adjusted for recombination [268]. iTOL was used to annotate and colour the resulting diagrams [270].

Investigation of sequence diversity among serogroups/types

To examine the level of *cps* locus sequence diversity amongst the 49 different serotypes included in this study, the nucleotide sequence of each reference serotype was BLASTed against the pneumococcal genomes in BIGSdb and the corresponding sequences were exported. Using MEGA6, each set of serotype-specific nucleotide sequences was aligned, converted to amino acid sequences and the variable amino acids were exported [356]. The variable amino acid alignments were then exported to MS Excel and annotated: across the alignment amino acids that were identical to the reference serotype

were coloured the same colour as the reference whilst alternative amino acids were shaded grey.

Putative novel (*cps* sequence with unknown origin) and hybrid (sequence combinations of ≥2 different serotypes) *cps* loci were identifiable in the phylogenetic trees based on distinct clustering and in the variable amino acid alignments based upon significant divergence from the reference sequence. Variable regions of sequence were manually extracted and investigated (e.g. via BLAST searches to the study database, the rMLST database and/or GenBank) to determine the likely origin of the alternative sequence.

Determination of pneumococcal genetic lineages

Prokka was used to predict coding sequences (CDS) in each genome and add annotation using a bespoke pneumococcal sequence database compiled for this study based upon the available gene annotation data from all pneumococcal genomes in GenBank [262]. The resulting annotation files in gff format were then input into Roary and clustered using a sequence identity threshold of 90% [264]. A core genome threshold of 100% (i.e. those genes present in every genome were considered core genes) was implemented to select the core genes for each serogroup/type and these gene sequences were aligned. FastTreeMP and ClonalFrameML were used to construct phylogenetic trees. Multilocus sequence type (MLST) data were automatically extracted from each genome using BIGSdb and STs were clustered into clonal complexes (CCs) using Phyloviz [332]. Each core gene phylogenetic tree was annotated using iTOL.

5.4 Results

Serogroups/types with low to moderate diversity in the *cps* locus

Serotype 1

Serotype 1 pneumococci (n=348) were recovered in 22 countries from 1943-2012 (Table 5.2). Minor serotype 1 *cps* locus variation was revealed among 12 of 14 genes, but patterns of variation were most commonly identified in *wzd* and *wze* (*cpsC* and *cpsD*, regulation) and *ugd*, *rmlA* and *rmlB* (sugar biosynthesis) (Supplementary File 5.2; [354]). Three genetic lineages were represented although most serotype 1 pneumococci in this study were within CC217, given that many serotype 1 genomes were from a recent African study ([353]; Figure 5.2).

Serotype 3

Serotype 3 pneumococci (n=205) were collected in 10 countries from 1961-2014 (Table 5.2) and 81% of serotype 3 pneumococci were in CC180, a widespread serotype 3 lineage (Figure 5.2; http://pubmlst.org/spneumoniae/). Little variation within the *cps* locus was revealed and what was noted was mainly due to the fact that the serotype 3 reference sequence differed at five amino acids in *ugd* (sugar biosynthesis), *wchE* and *galU* (glycosyltransferases) from the majority of the serotype 3 pneumococci (Supplementary File 5.3).

Serogroup 7

Distinct clusters of serotype-specific *cps* loci were observed among serogroup 7 pneumococci (total n=76; Figure 5.2). Nearly half (n=36) were serotype 7F pneumococci of three major genetic lineages, CC191, CC218 and CC3545/5265, plus two singletons

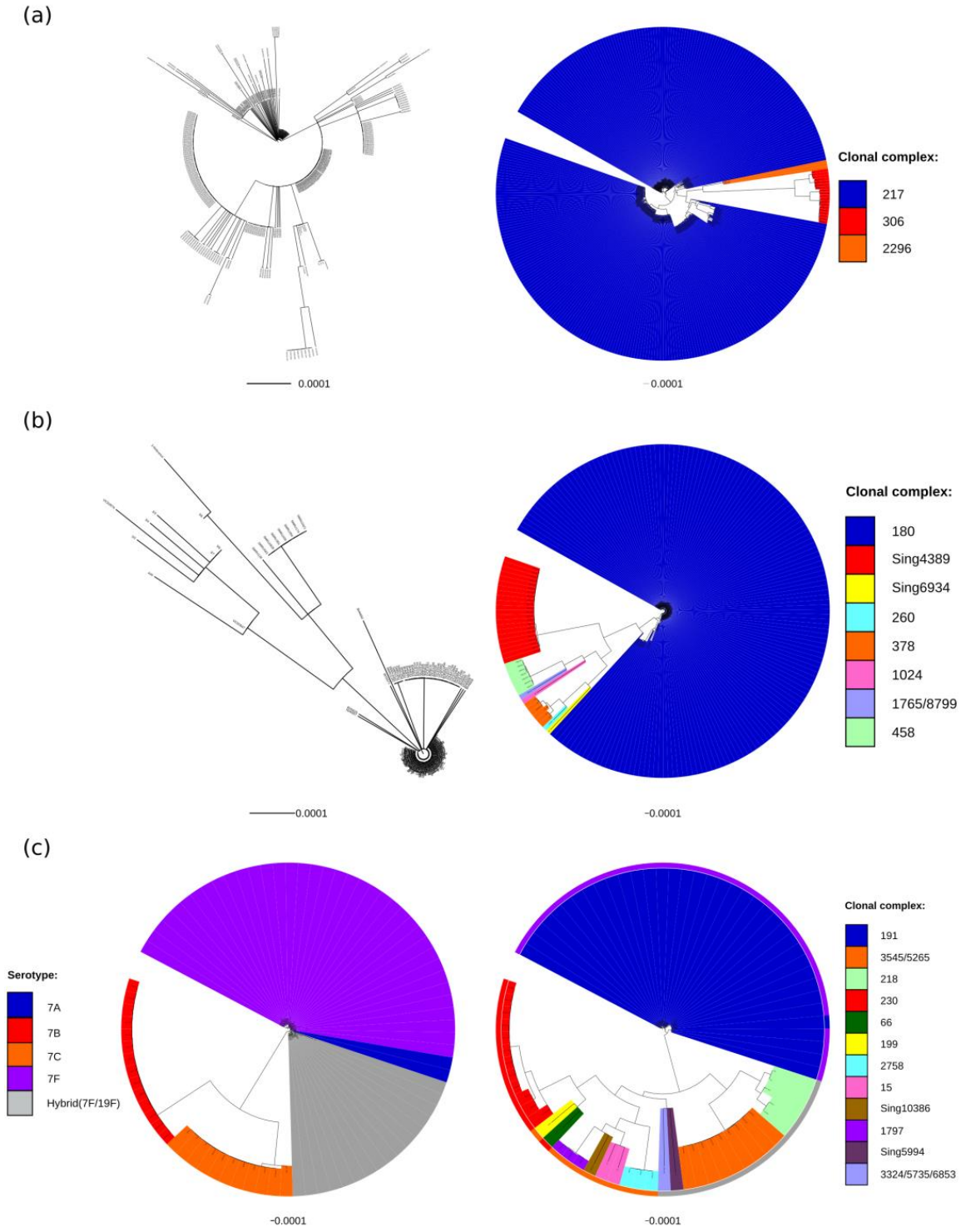


Figure 5.2 (a) Phylogenetic trees generated based on 15 *cps* locus genes amongst 348 serotype 1 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,000 full-length coding loci found in all 348 serotype 1 genomes (right). (b) Phylogenetic trees generated based on 4 *cps* locus genes amongst 205 serotype 3 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,160 full-length coding loci found in all 205 serotype 3 genomes (right). (c) Phylogenetic trees generated based on 10 *cps* locus genes amongst 76 serogroup 7 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,378 full-length coding loci found in all 76 serogroup 7 genomes (right). The outer rings of the trees on the right indicate the serotype by colour as detailed in the corresponding trees on the left.

(single genotypes with no closely related variants) that dated from 1962-2014. The *cps* locus genes were generally highly conserved, although minor variation at amino acids in *wzd* (*cpsC*, regulation), *wcwA* (glycosyltransferase) and *rmlD* (sugar biosynthesis) was noted (Supplementary File 5.4). Thirteen serotype 7B pneumococci were characterised and dated from 1952-2011 (Table 5.2), most of which were from Thailand (n=11; CC230; Fig. 1). Serotype 7C pneumococci (n=10; 5 different CCs) were recovered from 1939-2009 in four countries. The serotype 7B and 7C *cps* loci were highly conserved (Supplementary File 5.4). Serotype 7A (differentiated from 7F by a frameshift in *wcwD*) was rare: only one example (CC191) from 1937 was detected.

Additionally, a hybrid *cps* locus comprised of *wzg* (*cpsA*, regulation) sequence that matched both serotypes 7F and 19F was identified (Supplementary File 5.4). Carriage and disease-associated pneumococci with this hybrid locus (n=16) were isolated in four countries from 1952-2009 (Table 2). The pneumococci were of four genetic lineages, but mainly CC3545/5265 (n=9) from Thailand and CC218 (n=5), which is widespread (http://pubmlst.org/spneumoniae/).

Serogroup 9

Distinct clusters of serotype-specific *cps* loci were observed among 126 serogroup 9 pneumococci (Figure 5.3). Serotype 9V pneumococci were the most common (n=69), dated from 1968-2014, and 83% were members of CC156/162. One major serotype 9V *cps* locus sequence predominated, although eight minor Icelandic variants were identified with changes in *wzg*, *wze* and *wchO* (glycosyltransferase; Supplementary File 5.5). In addition, a variant serotype 9V *cps* locus (serotype 9Vii, Figure 5.3 and Supplementary File 5.5) with divergent alleles at 7 of 14 *cps* locus genes was revealed among 11 serotype 9V isolates (10 CC280 genomes from Thailand plus one singleton genome from

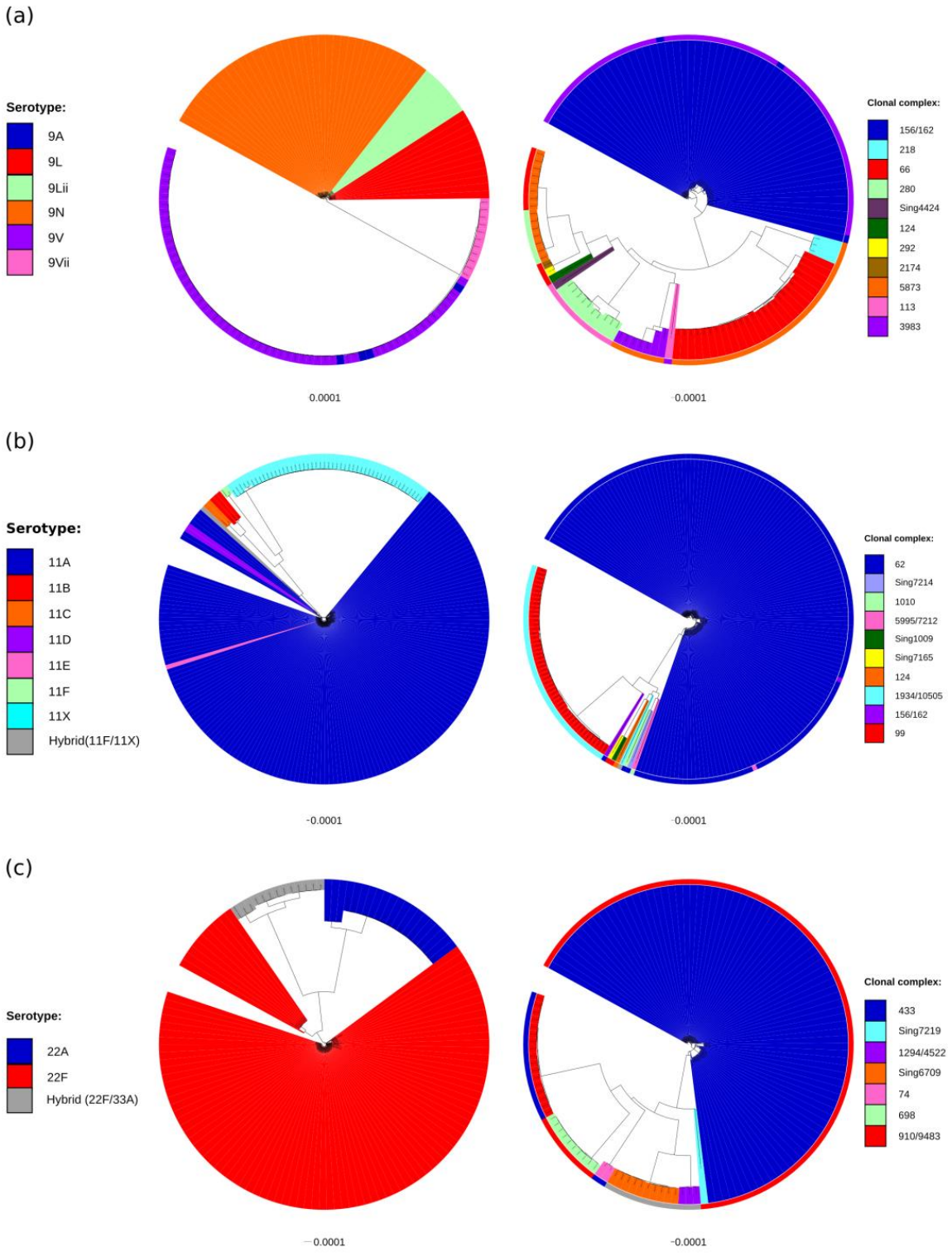


Figure 5.3 (a) Phylogenetic trees generated based on 12 *cps* locus genes amongst 126 serogroup 9 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,259 full-length coding loci found in all 126 serogroup 9 genomes (right). (b) Phylogenetic trees generated based on 10 *cps* locus genes amongst 222 serogroup 11 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,204 full-length coding loci found in all 222 serogroup 11 genomes (right). (c) Phylogenetic trees generated based on 16 *cps* locus genes amongst 130 serogroup 22 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,390 full-length coding loci found in all 130 serogroup 22 genomes (right). The outer rings of the trees on the right indicate the serotype by colour as detailed in the corresponding trees on the left.

the UK). Also of note was that the serotype 9V reference sequence differed at one amino acid in *wcjB* from all but one serotype 9V pneumococcus. The *cps* locus of serotype 9A is similar to 9V apart from a disrupted *wcjE* gene [15], but serotype 9A was uncommon in this study (n=3).

Serotype 9N pneumococci were also frequently identified (n=36), were recovered from 1938-2014 in five countries (Table 5.2), and were mainly of two genetic lineages (CC66, n=26; CC3983, n=7). The serotype 9N *cps* locus was generally highly conserved apart from isolated amino acid changes in *wze*, *wchA*, *wchO* and *wcjC* (the latter among Thailand pneumococci only; Supplementary File 5.5). Most serotype 9L pneumococci (n=18) were identified within the Thai dataset (n=15; CC5873) except for three historical isolates isolated from 1941-1968. No variable amino acids were observed among 10 of 13 serotype 9L *cps* locus genes although a distinct *wzg* allele was observed among seven Thai pneumococci (serotype 9Lii; Supplementary File 5.5).

Serogroup 11

Serogroup 11 consists of six known serotypes but most were uncommon in this study. Serotypes 11C, 11D, 11E and 11F were represented by single pneumococci from 1957, 1986, 2012 and 1952, respectively, and only two serotype 11B pneumococci from 1940 and 2010 were identified (Table 5.2; Figure 5.3). 75% of the 222 serogroup 11 isolates were serotype 11A and 92% of those were in CC62. Excluding unusual variants (see below), seven of 16 genes within the *cps* locus of the serotype 11A pneumococci were identical to the serotype 11A reference, whilst five genes were differentiated from the reference by various single amino acid changes (Supplementary File 5.6). Most serotype 11A pneumococci consistently differed from the reference in genes *wchA* (glycosyltransferase), *wcwC* and *wcjE* (acetyl transferases).

A putative novel *cps* locus (11X) was identified among 49 pneumococci (CC99) from Thailand (Figure 5.3). The sequence was most similar to serotype 11F, although the nucleotide sequence similarity for each gene ranged from 85.5-99.6% (Supplementary File 5.6). Notably, the novel *cps* locus had distinctive *wcwC* and *wcrL* (glycosyltransferase) genes (91.3% and 85.5% similarity to serotype 11A) that did not match any other serotype sequences in GenBank. The novel locus also had intact *gct* (CDP-glycerol pathway) and *wcjE* genes, similar to serotype 11A. Interrogation of the PubMLST database revealed that pneumococci within CC99 were also identified in South Korea and China, but whether they possess the novel *cps* locus is unknown. Finally, a single hybrid combination of serotype 11A/11X was identified in one pneumococcus from Massachusetts (Figure 5.3 and Supplementary File 5.6).

Serogroup 22

130 serogroup 22 pneumococci were analysed and serotypes 22F, 22A and a hybrid were identified (Figure 5.3). Serotype 22F pneumococci (n=98) were recovered from 1940-2014 in four countries (Table 2), and 89% were in CC433. Overall, one major and one minor cluster of the serotype 22F *cps* locus were observed: all 19 genes differed from the reference by ≤2 amino acids (Figure 5.3 and Supplementary File 5.7). Apart from two historical pneumococci from 1939 (CC74), serotype 22A pneumococci (CC910/9483) were from Thailand (n=17). The serotype 22A *cps* locus was also conserved: 16 of 19 genes were identical to the serotype 22A reference. Finally, a serotype 22F/33A hybrid *cps* locus was identified among 13 Thai pneumococci (Figure 5.3 and Supplementary File 5.7).

Serogroups/types with high diversity in the *cps* locus

Serogroup 6

In Chapter 4, 974 serogroup 6 pneumococci were analysed which demonstrated a high level of *cps* locus diversity within the serogroup, and revealed that the vast majority of isolates that had previously been phenotypically typed as serotype 6B were in fact serotype 6E, which were ~7% divergent at a nucleotide level in the *cps* locus [253]. Subsequent to that study, other investigators reported that the serotype 6B and 6E polysaccharide capsule structures were identical [129]. Consequently, in this chapter serotype 6E sequences will be referred to as serotype 6Bii (serotype 6B class 2), as originally described by Mavroidi and colleagues, whilst serotype 6B class 1 sequences will be annotated as serotype 6Bi [128]. In this chapter the original 974 genomes and an additional 208 new genomes were analysed, so that serogroup 6 *cps* locus diversity could be compared to other high prevalence serogroups, in particular serogroups 19 and 23.

Analyses of the 1,182 serogroup 6 *cps* locus sequences revealed seven serotypes and nine different hybrid serotypes (Figure 5.4). Serotype 6Bii pneumococci were the most prevalent (n=420), were collected from 1981-2014 in 16 countries (Table 5.3), and four major lineages were represented: CC90 (n=200); CC315 (n=104); CC4405 (n=54); and CC273 (n=21). The serotype 6Bii *cps* locus sequence was conserved (Supplementary File 5.8). Serotype 6A pneumococci (n=283) were also frequently recovered. They were members of 16 different CCs but the two major lineages were CC460 (n=109) and CC490 (n=74). There were five major clusters of serotype 6A *cps* loci and sequence variation occurred across the *cps* locus (Figure 5.4 and Supplementary File 5.8).

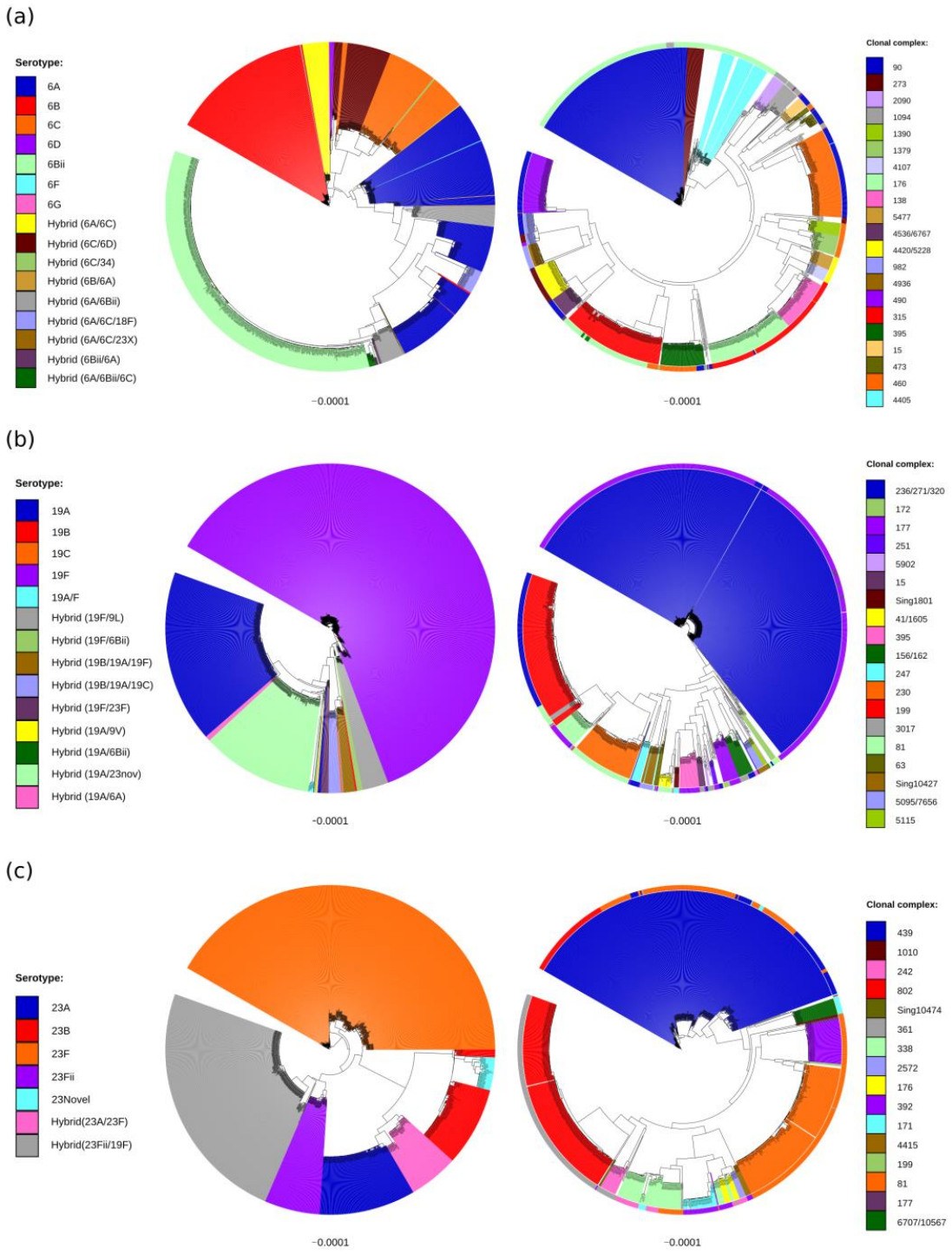


Figure 5.4. (a) Phylogenetic trees generated based on 13 *cps* locus genes amongst 1,182 serogroup 6 pneumococcal genomes (left) and constructed using the concatenated sequence of 470 full-length coding loci found in all 1,182 serogroup 6 genomes (right). CC's with ≥10 isolates were coloured as listed in the key. (b) Phylogenetic trees generated based on 13 *cps* locus genes amongst 1,185 serogroup 19 pneumococcal genomes (left) and constructed using the concatenated sequence of 505 full-length coding loci found in all 1,185 serogroup 19 genomes (right). CC's with ≥5 isolates were coloured as listed in the key. (c) Phylogenetic trees generated based on 17 *cps* locus genes amongst 874 serogroup 23 pneumococcal genomes (left) and constructed using the concatenated sequence of 628 full-length coding loci found in all 874 serogroup 23 genomes (right). The outer rings of the trees on the right indicate the serotype by colour as detailed in the corresponding trees on the left.

171 serotype 6Bi pneumococci collected from 1939-2014 in four countries were characterised and these were mainly of CC176 (n=103) or CC138 (n=62; Figure 5.4). The serotype 6Bi reference was not representative of the majority of serotype 6Bi pneumococci and was likely a capsular switch from serotype 6A to 6Bi. If the 6Bi reference sequence was excluded from the alignment then the *cps* locus was generally conserved among serotype 6Bi pneumococci (Supplementary File 5.8). 106 serotype 6C pneumococci were revealed and were collected from 2001-2014 in three countries. Eight CCs were represented, but four were the most common: CC395 (n=44); CC1379 (n=24); CC1390 (n=160); and CC315 (n=13). Four distinct variants were observed in the *cps* loci sequence alignment and the variation occurred across the *cps* locus (Supplementary File 5.8).

Serotypes 6D and 6F were rare and serotypes 6G and 6H were not identified; however, hybrid serogroup 6 *cps* loci were common. Combinations of serogroup 6 pneumococci were most often observed, although the import of *cps* locus sequence from other serotypes was also revealed (Figure 5.4 and Supplementary File 5.8).

Serogroup 19

1,185 pneumococci within serogroup 19 were analysed, the majority of which were unequivocal serotypes 19F (n=745) and 19A (n=210). 88% (n=658) of the serotype 19F pneumococci were recovered in 9 countries from 1995-2014 and were members of CC236/271/320. Three other major serotype 19F lineages included CC81 (n=29), CC395 (n=24) and CC177 (n=13). Four different clusters of serotype 19F *cps* locus sequences were identified and amino acid variation occurred across the *cps* locus (Table 5.3; Figure 5.4 and Supplementary File 9).

Serotype 19A pneumococci were recovered in 5 countries from 1939-2014 (Table 5.3). 78% (n=164) were in CC199, a widely-distributed lineage (http://pubmlst.org/spneumoniae/; Figure 5.4). Two other major lineages were CC236/271/320 (n=17) and CC247 (n=11), the latter being the earliest reported vaccine escape pneumococci [168]. Alignments of the unequivocal serotype 19A *cps* loci to the 19A reference sequence demonstrated identical sequences between the 19A reference and three historical 19A pneumococci from 1939, 1952 and 1968. However, modern serotype 19A pneumococci differed considerably from the 19A reference across the *cps* locus (e.g. the nucleotide sequences of *rmlC*, *rmlB* and *rmlD* diverged from the reference sequence by up to 8%, 15% and 1%, respectively; Supplementary File 5.9). The *cps* loci of modern 19A pneumococci were highly similar. Serotype 19B and 19C pneumococci were rare (n=1 each).

The remaining 222 serogroup 19 pneumococci revealed nine different types of hybrid *cps* loci (Figure 5.4 and Supplementary File 5.9). Hybrids containing 19A *cps* locus sequences (n=145) were most common and 93% of these pneumococci were a combination of 19A/23X (novel serotype 23, see below) that resulted in serotype 23-like changes in *rmlC*, *rmlB* and *rmlD*. Three major CCs described 82% of the 19A/23X hybrids: CC230 (n=76); CC199 (n=21); and CC41/1605 (n=14).

47 pneumococci contained a hybrid *cps* locus containing serotypes 19F/9L (n=31), 19F/23F (n=9) or 19F/6Bii (n=7) and six pneumococci revealed a combination 19A/19F *cps* locus that has been reported previously (Supplementary File 5.9; [357-359]). 30 pneumococci from Thailand possessed a *cps* locus that was a hybrid of either 19B/19A/19C (n=16) or 19B/19A/19F (n=14), each with divergent *wzg*, *wzh*, *wzd and wze* genes and evidence for multiple instances of recombination, while the rest of the *cps* locus of these pneumococci was similar to serotype 19B.

Serogroup 23

874 pneumococci within serogroup 23 were characterised and 43% (n=375) were unequivocal serotype 23F pneumococci recovered in 22 countries from 1940-2014 (Table 5.3). The major genetic lineages were CC439 (n=160), CC81 (n=142) and CC338 (n=20). There were two major serotype 23F *cps* locus clusters (Figure 5.4) and sequence variation occurred across the *cps* locus, but was mainly concentrated in *wzg*, *wchA*, *wchF*, *rmlA*, *rmlC*, *rmlB* and *rmlD* (Supplementary File 5.10). A variant serotype 23F (23Fii; n=49) *cps* locus was also revealed, which differed significantly at *wzg* and *wzh* from the serotype 23F reference. Pneumococci with this locus were recovered in Thailand and the USA from 1983-2010 and were mainly of CC171 (n=30). The serotype 23Fii variant was also detected as part of a hybrid *cps* locus with 23Fii and 19F sequences (Figure 5.4). 217 pneumococci collected from 1997-2010 in four countries possessed this hybrid *cps* locus and two major lineages were revealed, CC802 (n=193) and CC242 (n=15). Serotype 19F-like sequence was identified in *wzg* and part of *wzh* but otherwise the *cps* locus was highly similar to serotype 23F.

84 pneumococci from four countries (1945-2014) typed as serotype 23A (Table 5.3). All were members of CC439 and the *cps* locus sequence was conserved (Figure 5.4 and Supplementary File 5.10). There were also 44 pneumococci from three countries (2008-2011) with a hybrid serotype 23A/23F *cps* locus and 80% of these pneumococci were in CC338. The *cps* locus changes were mainly in *wzg*, although *rmlA* and *rmlD* also included some variable amino acids that matched serotype 23B sequence.

There were 76 serotype 23B pneumococci collected from 1941-2014 (Table 5.3). All were members of CC439 except one pneumococcus from 1941 and the *cps* loci were highly conserved (Figure 5.4 and Supplementary File 5.10). In addition, 29 pneumococci

possessed some serotype 23B sequence, but the nucleotide sequences of eight genes, *wchX* through *rmlD*, differed by 14% (*rmlB*) to 22% (*rmlD*) from the serotype 23B reference (Supplementary File 5.10). Unusually, *rmlD* was inverted, which suggested a genome rearrangement (Supplementary File 5.10). A search of all available *cps* locus reference sequences failed to identify the possible source of this sequence. This putative novel serotype 23X was present in 29 pneumococci collected from 1996-2014 in five countries. 15 pneumococci were from a Thai lineage and the others were mainly in CC338 (n=7) and CC439 (n=5).

Serotype 14

Four major *cps* locus clusters were clearly identified among a collection of 258 serotype 14 pneumococci collected from 1939-2013 (Figure 5.5). 58 pneumococci possessed *cps* loci that were identical or nearly identical to the serotype 14 reference sequence and all but two were in CC124. 49 pneumococci, predominantly of CC15, possessed *cps* loci with a documented frameshift in *wciY* that is not believed to disrupt capsule production [126, 360]. Pneumococci with the frameshifted *wciY* also possessed a divergent *Lrp*, a large repetitive protein which varies in size across the entire serotype 14 collection. CC124 and CC15 are among the most common serotype 14 lineages circulating worldwide (http://pubmlst.org/spneumoniae/).

88 pneumococci from Thailand possessed a serotype 14 variant (14ii) *cps* locus with significant amino acid differences across seven *cps* locus genes and these pneumococci were all members of CC63, a widely-distributed lineage

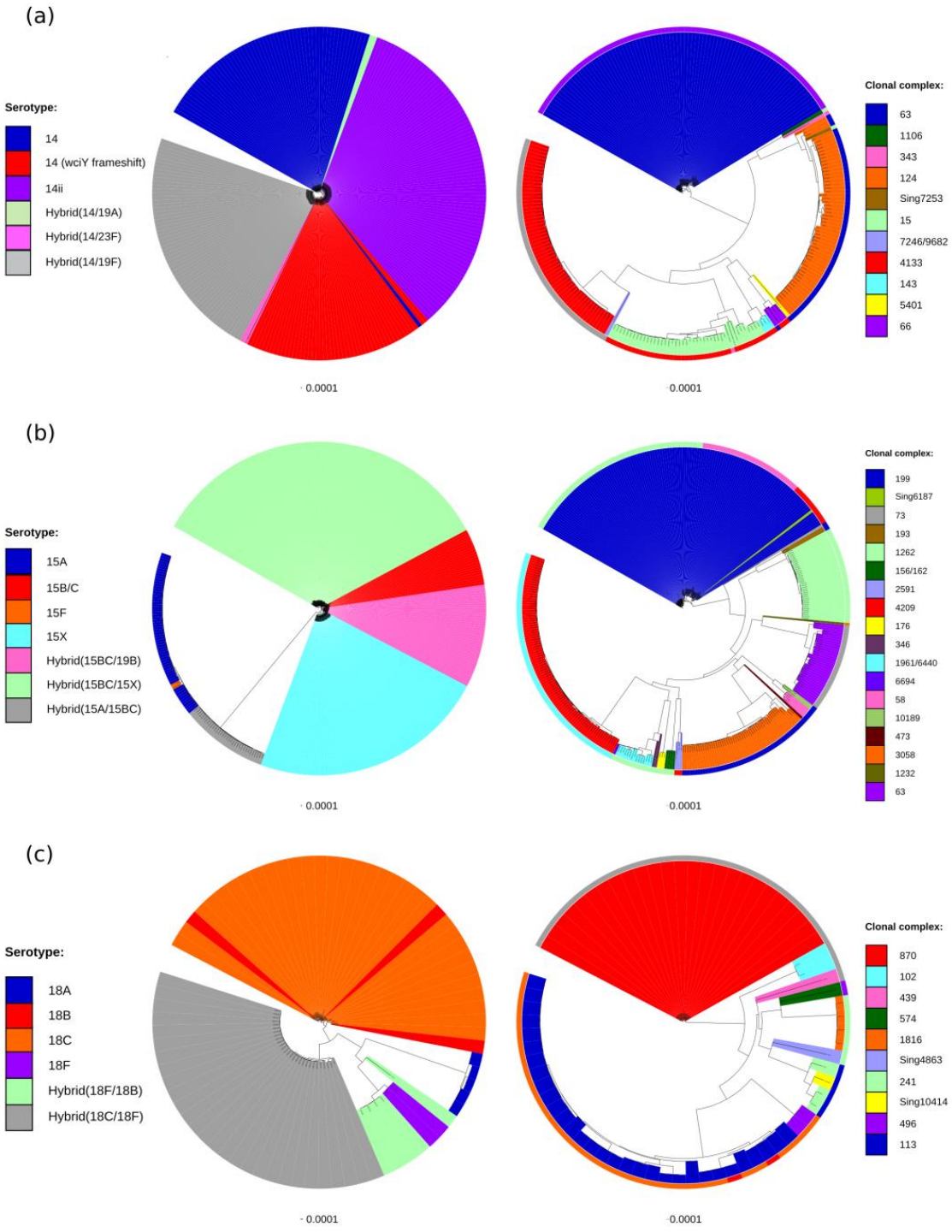


Figure 5.5 (a) Phylogenetic trees generated based on 12 *cps* locus genes amongst 258 serotype 14 pneumococcal genomes (left) and constructed using the concatenated sequence of 911 full-length coding loci found in all 258 serotype 14 genomes (right). (b) Phylogenetic trees generated based on 13 *cps* locus genes amongst 385 serogroup 15 pneumococcal genomes (left) and constructed using the concatenated sequence of 893 full-length coding loci found in all 385 serogroup 15 genomes (right). (c) Phylogenetic trees generated based on 17 *cps* locus genes amongst 71 serogroup 18 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,302 full-length coding loci found in all 71 serogroup 18 genomes (right). The outer rings of the trees on the right indicate the serotype by colour as detailed in the corresponding trees on the left.

(pubmlst.org/spneumoniae/; Figure 5.5). In addition, three different hybrid *cps* loci were identified, which were combinations of serotypes: 14/19F (n=60); 14/23F (n=2); and 14/19A (n=2; Figure 5.5 and Supplementary File 5.11). The serotype 14/19F hybrid *cps* locus was found in one South African pneumococcus from 1988 and in 59 recent pneumococci from Thailand (CC4133).

Serogroup 15

Serotype clusters within serogroup 15 pneumococci (n=385) were clearly differentiated (Figure 5.5). 62 serotype 15A pneumococci were recovered in 5 countries from 1939-2014 (Table 5.3) and 84% were members of CC3058. The *cps* loci of serotype 15A pneumococci were similar to the 15A reference sequence: relatively few amino acids varied across the 16 *cps* locus genes and the variation was predominantly among 51 serotype 15A pneumococci from Thailand (Supplementary File 5.12). A single example of serotype 15F was collected in 1963.

The serotype 15B/C *cps* loci were diverse. Only 20 pneumococci possessed a *cps* locus similar to the serotype 15B and 15C references and 80% were of CC199. Initially typed as serotype 15B/C, a putative novel *cps* locus (15X) was found in 92 pneumococci of CC4209 isolated in Thailand. Gene order was identical to that of the 15A and 15B/C *cps* loci; however, there were distinct differences in the amino acid sequences of *wzg*, *wzh* and *wzy* (polymerase), and other genes displayed evidence of recombination with serotype 15A (*wchA*, *gtp1* to *gtp3*, sugar biosynthesis; Supplementary File 5.12).

Finally, three large groups of hybrid *cps* loci were identified (Figure 5.5 and Supplementary File 5.12). 114 pneumococci were members of CC199 and possessed a hybrid *cps* locus combination of either 15BC/15X (n=74) or 15BC/19B (n=40). The 15BC/15X hybrid was also identified among pneumococci from six other genetic

lineages. A hybrid combination of 15A/15BC was also revealed among 34 pneumococci in CC63.

Serogroup 18

71 pneumococci of serogroup 18 were investigated and the predominant serotype was 18C (n=31). Nearly all 18C pneumococci were in CC113, a widespread 18C lineage (http://pubmlst.org/spneumoniae/), and the serotype 18C *cps* locus was conserved (Figure 5.5 and Supplementary File 5.13). Serotypes 18A and 18F were uncommon (n=4 and n=1, respectively) and only found in Thailand. One Icelandic pneumococcus had a serotype 18B *cps* locus, which differs from serotype 18C by a frameshift in *wciX* (acetyl transferase).

Two hybrid *cps* loci were revealed: 18C/18F (n=28) and 18F/18B (n=5). The 18C/18F hybrid pneumococci (CC870) were mainly from Thailand, although three were from Iceland (CC102) and the USA (CC439); these pneumococci possessed variable amino acids at 12 *cps* locus genes, including four genes involved in sugar biosynthesis (Supplementary File 5.13).

Serogroup 33

115 pneumococcal genomes were analysed and this serogroup was unusual in that only five *cps* locus genes shared ≥70% sequence identity across the serotypes within serogroup 33, even though the genes that comprised each *cps* locus were broadly similar (Appendix 5.3; [15]). Serotype clusters were clearly delineated within the tree except for serotypes 33A and 33F, which differ only by a frameshift in *wcjE* (Figure 5.6). Serotype 33F (n=40) pneumococci were frequently identified in three countries from 1999-2014, and

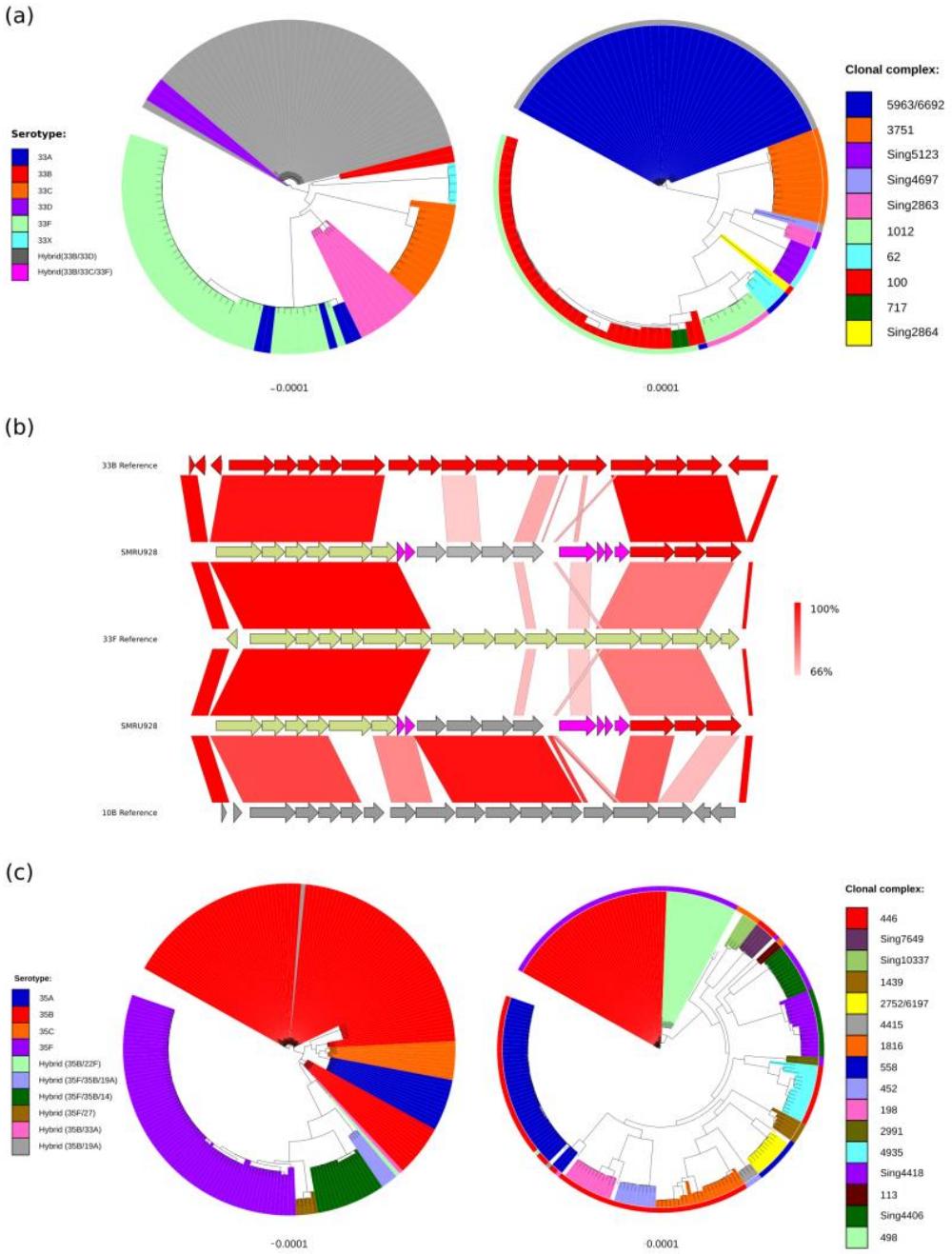


Figure 5.6 (a) Phylogenetic trees generated based on 5 *cps* locus genes amongst 115 serogroup 33 pneumococcal genomes (left) and constructed using the concatenated sequence of 1,218 full-length coding loci found in all 115 serogroup 33 genomes (right). (b) Comparison of the putative serogroup 33X *cps* locus sequence to the serotype 33B, 33F and 10B reference sequences. The results of pairwise BLAST nucleotide sequence comparisons are shown: darker red colour highlights greater sequence conservation between the pair of sequences. In order to generate pair-wise comparisons between all three reference sequences the putative serogroup 33X *cps* locus sequence was included twice. Coding loci in the putative serogroup 33X *cps* locus sequence were coloured according to the level of sequence similarity to coding loci from one of the reference sequences: light green for serotype 33F; grey for serotype 10B; red for serotype 33B; and disrupted loci were highlighted in pink. (c) Phylogenetic trees generated based on 7 *cps* locus genes amongst 228 serogroup 35 pneumococcal genomes (left) and constructed using the concatenated sequence of 993 full-length coding loci found in all 228 serogroup 35 genomes (right). The outer rings of the trees on the right indicate the serotype by colour as detailed in the corresponding trees on the left.

all but two were in CC100. Three of four serotype 33A pneumococci were historical (1937-1946, CC62 and CC100), two serotype 33D were from 1979, and one historical serotype 33B from 1962 was detected. Eleven serotype 33C pneumococci from Thailand were identified, all of CC3751. Sequence alignments of serotype 33F, 33A and 33C *cps* loci revealed conserved sequences when compared to each respective reference sequence (Supplementary File 5.14).

With the exception of two pneumococci from 1979, all capsular sequences extracted from pneumococci initially typed as serotype 33D revealed evidence of recombination with serotype 33B, which likely altered ≥6 *cps* locus genes (n=44; Figure 5.6 and Supplementary File 5.14). All but one of these hybrid pneumococci were from Thailand. A second hybrid *cps* locus comprised of sequence from serotypes 33B/33C/33F was revealed in eight pneumococci of CC62 from the UK and Iceland. Finally, five Thai pneumococci possessed a novel 33X *cps* locus, which consisted of a complex arrangement of nucleotide sequences from serotypes 33B, 33F and 10A/B *cps* locus genes (Figure 5.5).

Serogroup 35

228 pneumococci within serogroup 35 were analysed and four known serotypes were identified, in addition to several hybrid *cps* loci (Figure 5.6). Two distinct clusters of serotype 35B (n=108) pneumococci were identified. The pneumococci were recovered in five countries from 1939-2014 and 5 major and 3 minor CCs were recognised (Table 5.3; Figure 5.6). 81 serotype 35B *cps* locus sequences were conserved relative to the 35B reference, with the exception of some scattered variation (Supplementary File 5.15). The remaining 35B *cps* loci possessed many variable amino acids relative to the reference

sequence and these were of two major groups from Iceland (n=10) and Thailand (n=17). 22 of these were members of CC1816 and the rest were a unique Thai lineage.

Single clusters of 35F (n=73), 35A (n=11) and 35C (n=8) were also revealed (Figure 5.6). Serotype 35F pneumococci were identified in five countries from 1939-2014 and three major lineages were detected: CC446 (n=41; Iceland, UK); CC498 (n=17; USA); and CC4406 (n=12; Thailand). Serotype 35A and 35C sequences were mainly identified among Thai pneumococci, apart from historical isolates from 1939 (35A) and 1941-1943 (35C), and these pneumococci were members of one or one predominant CC, respectively (Figure 5.5). The *cps* loci of each serotype were generally highly conserved (Supplementary File 5.15).

Finally, three hybrid *cps* loci of different serotype combinations were identified in Thailand (35F/14/35B, n=16; 35F/27, n=5; and 35F/19A/35B, n=4). Three additional hybrid combinations were identified in the USA (35B/22F, 35B/33A and 35B/19A, n=1 each; Supplementary File 5.15).

5.5 Discussion

Genomics has revolutionised microbiological research: thousands of bacterial genomes are now in the public domain and the deluge of new data continues. Maximising the utility of the vast quantities of genome sequence data from well-sampled studies is advancing our understanding of fundamental aspects of microbiology, epidemiology, population biology and microbial evolution. Given the significance of the polysaccharide capsule to pneumococcal biology it is essential to understand the variation (or lack thereof) that occurs within and between different serotypes. In this study the level of sequence diversity at the *cps* locus for each serogroup/type was assessed and hybrid or putatively novel serotypes were identified within each serogroup.

This study revealed serogroups/types that were largely conserved and others that were rather diverse. Diversity among high prevalence serogroups like 6, 19 and 23 was expected, but variation among serotypes like 14 was more surprising. To what extent the variation is biologically informative remains to be deciphered, but is the essential next step in gaining a deeper understanding of pneumococcal capsular biology. Furthermore, unusual variation at the level of the *cps* locus and the circulating genetic lineages was frequently noted among the pneumococci from Thailand. Whether this reflects an unusual collection of pneumococci or whether it is simply that collections of pneumococci from resource-poor countries differ significantly from other geographical regions remains to be determined. Finally, the selection of reference strains of pneumococci must be made carefully, since some *cps* locus reference strains [15] are not particularly good representatives of modern circulating pneumococci. Sequencing the 90 *cps* loci some years ago was a big step forward at a time when sequencing costs were still very high and limited *cps* loci data were available, but some revisions to the selections of reference pneumococci would now be useful.

Investigating serotype-specific sequence variation is crucial to the design of sequence-based serotyping methodology. Sequence diversity within regions targeted by PCR or microarray could lead to erroneous test results and subsequently, a misinformed view about the circulating serotypes. Since many countries are now using PCVs, pre- and post-PCV introduction surveillance is essential and correctly identifying the serotypes of circulating pneumococci is central to surveillance.

Understanding *cps* locus diversity is also fundamental to assessing PCV effectiveness. Longer-term effectiveness depends (at least in part) on the PCV-specific polysaccharides representing the circulating pneumococci in order to elicit an appropriate vaccine-mediated antigenic response, and on a lack of subsequent capsular polysaccharide

changes that could lead to a reduction in the protective capacity and a potential for vaccine failure. The emergence of *cps* locus variants appears to be part of the normal biological processes of pneumococci and is unlikely to be driven primarily by the introduction of vaccines [253, 348], but the selective pressure of vaccine use may result in a population-level increase in the variants that do emerge and this could impact overall vaccine effectiveness.

It is known that small changes to genes within the *cps* locus can lead to changes in the polysaccharide structure, *e.g.* single amino acid changes in the *wciP* of serogroup 6 lead to the different serotype 6A and 6B polysaccharides [128], but equally the large genetic differences between serotypes 6Bi (6B) and 6Bii (6E) apparently do not confer a difference in the polysaccharide [129]. Future work should investigate precisely which sequence-based changes revealed in this work result in a biologically relevant change, particularly among high prevalence serotypes and those included in current PCVs. Equally, any future higher-valency formulations of PCVs that utilise the polysaccharide capsule to elicit a protective immune response will undoubtedly benefit from an assessment of *cps* locus diversity for each vaccine serotype, to assess the pre-PCV introduction capsular diversity and the potential for reduced vaccine effectiveness.

This study also revealed many hybrid *cps* loci and putatively novel serotypes. The hybrid *cps* loci are unsurprising, given the genome-wide level of recombination that occurs among pneumococci, including at the *cps* locus. Confirmation of putatively novel serotypes requires a biochemical assessment of the polysaccharide capsule to confirm novelty, which is not a trivial task. One might take the view that rare serotype variants are not important enough to warrant complicated and costly biochemical exploration; however, putatively novel serotypes that are common in a population arguably are worth further assessment.

This work provides a major advance in our understanding of pneumococcal serotypes and sequence diversity at the *cps* locus. Genomic analyses of a very large and diverse dataset have allowed for a fine-scale investigation of pneumococcal *cps* diversity and the data are publicly available in order that they might be used to maximal effect within the research community.

5.6 Appendixes

Appendix 5.1. PMEN clones included in this study

Reference	Other name	Serotype	ST	Accession number	Year	Reference(s)
PMEN1	Spain^{23F}-1	23F	81	ATCC 700669; ENA	1984	[18, 361, 362]
PMEN2	Spain^{6B}-2	6Bii	90	ATCC 700670; ENA	1998	[363]
PMEN3	Spain^{9V}-3	9V	156	ATCC 700671; ENA	1993	[361, 364]
PMEN4	Tennessee^{23F}-4	23F	37	ATCC 51916; ENA	1991	[365, 366]
PMEN5	Spain^{14}-5	14	18	ATCC 700902; ENA	1990	[333]
PMEN6	Hungary^{19A}-6	19A	268	ATCC 700673; ENA	1989	[363, 367]
PMEN8	S.Africa^{6B}-8	6Bii	185	ATCC 700675	1990	[334]
PMEN12	Finland^{6B}-12	6Bii	238	ATCC 700903	1987	[335]
PMEN13	S.Africa^{19A}-13	19A	41	ATCC 700904; ENA	1988	[334]
PMEN14	Taiwan^{19F}-14	19F	236	ATCC 700905; ENA	1997	[368]
PMEN15	Taiwan^{23F}-15	23F	242	ATCC 700906; ENA	1997	[368]
PMEN17	Maryland^{6B}-17	6Bii	384	ATCC BAA-342; ENA	1997	[336]
PMEN18	Tennessee^{14}-18	14	67	ATCC BAA-340; ENA	1997	[336]
PMEN22	Greece^{6B}-22	6Bii	273	ATCC BAA-658; ENA	1996	[369, 370]
PMEN27	Sweden^{1}-27	1	217	PJ351/1; ENA	1992	[371, 372]
PMEN28	Sweden^{1}-28	1	306	PJ755/1; ENA	1997	[372, 373]
PMEN29	USA^{1}-29	1	615	NCTC7465; ENA	1948	[372]
PMEN35	Netherlands^{14}-35	14	124	M269-14; ENA	1980	[374]
PMEN36	Netherlands^{18C}-36	18C	113	M241-18C; ENA	1980	[374]
PMEN39	Netherlands^{7F}-39	7F	191	M248-7F; ENA	1984	[374]
PMEN40	Sweden^{1}-40	1	304	PJ537-1; ENA	Unknown	[375]

Appendix 5.2. *cps* reference sequences included in the study

Reference	Other name	ST[a]	Accession number	Year	Country of isolation	Reference
1	519/43	5316	CR931632	1943	Denmark	[15]
3	524/62	7211	CR931634	1962	Denmark	[15]
6A	34351 Rodriques	7188	CR931638	1952	USA	[15]
6B	2616/39	7199	CR931639	1939	Denmark	[15]
6C	CHPA388	--	EF538714	1999-2002	USA	[257]
6D	MNZ21	282	HM171374	2008	South Korea	[337]
6E	HK02-14	90	Multiple	2008	Hong Kong	[259]
6F	MNZ21135 PS6657	--	KC832411	1992-2012	Germany	[110]
6G	MNZ1135 PS16864	--	KC832410	1992-2012	Germany	[110]
7A	2040/37	191	CR931640	1937	Denmark	[15]
7B	Johnson	7180	CR931641	1952	USA	[15]
7C	Sutcliff	7171	CR931642	1971	USA	[15]
7F	554/62	191	CR931643	1962	Denmark	[15]
9A	Wilder	312	CR931645	1962	USA	[15]
9L	T9233/128/68	7240	CR931646	1968	Denmark	[15]
9N	533/62	71	CR931647	1962	Denmark	[15]
9V	980/68	123	CR931648	1968	Denmark	[12]
11A	1813/39	7214	CR931653	1939	Denmark	[12]
11B	8087/40	7165	CR931654	1940	Denmark	[12]
11C	Eddy nr. 53	7201	CR931655	1957	USA	[12]
11D	70/86	62	CR931656	1986	Denmark	[12]
11F	34356	7212	CR931657	1952	Denmark	[12]
14	34359	124	CR931662	1952	USA	[12]
15A	389/39	7166	CR931663	1939	Denmark	[12]
15B	7904/39	1	CR931664	1939	Denmark	[12]
15C	553/62	7172	CR931665	1972	Denmark	[12]
15F	688/63	7177	CR931666	1963	Denmark	[12]
18A	8609/43	241	CR931671	1952	Denmark	[12]
18B	1033/41	4706	CR931672	1941	Denmark	[12]
18C	4593/40	4706	CR931673	1940	Denmark	[12]
18F	Gethens	7182	CR931674	1961	USA	[12]
19A	nr. 141/68	7226	CR931675	1968	Denmark	[12]
19B	nr. 4594	199	CR931676	1971	Germany	[12]
19C	7588/39	3785	CR931677	1939	Denmark	[12]
19F	485/61	425	CR931678	1961	Denmark	[12]
22A	3405/39	7181	CR931681	1939	Denmark	[12]
22F	1772/40	7219	CR931682	1940	Denmark	[12]
23A	1196/45	439	CR931683	1945	Denmark	[12]
23B	1039/41	7168	CR931684	1941	Denmark	[12]
23F	Dr Melchior	515	CR931685	1996	Denmark	[12]
33A	Biehl	1012	CR931698	1946	USA	[12]
33B	E294	2864	CR931699	1962	Denmark	[12]
33C	7098/41	2865	CR931700	1941	Denmark	[12]
33D	CSF/79	2863	CR931701	1979	India	[12]
33F	3084/37	7192	CR931702	1937	Denmark	[12]
35A	1936/39	7242	CR931704	1939	Denmark	[12]
35B	4356/39	7234	CR931705	1939	Denmark	[12]
35C	7765/43	5989	CR931706	1943	Denmark	[12]
35F	361/39	7164	CR931707	1939	Denmark	[12]

Appendix 5.3. Serogroup *cps* loci included in the study

Serogroup/type	Genes
1	*aliB, wzg, wzh, wzd, wze, wchB, wchC, wchD, wzy, wzx, gla, ugd, rmlA, rmlC, rmlB*
3	*wzd, ugd, wchE, galU*
6	*wzg, wzh, wzd, wze, wchA,wciO, wciP, wzy, wzx, rmlA, rmlC, rmlB, rmlD*
7	*wzg, wzh, wzd, wze, wchA, wchF, rmlA, rmlC, rmlB, rmlD*
9	*wzg, wzh, wzd, wze, wchA, wchO, wcjA, mnaA, wzy, wcjB, wzx, ugd*
11	*wzg, wzh, wzd, wze, wchA, wchI, wchK, wcwT, wcwU, wzx*
14	*wzg, wzh, wzd, wze, wchA, wchJ, wchK, wzy, wchL, wchM, wchN, wzx*
15	*wzg, wzh, wzd, wze, wchA, wchJ, wchL, wchM, wchN, wchX, gtp1, gtp2, gtp3*
18	*wzg, wzh, wzd, wze, wchA, wchF, wciU, wciV, wciW, wzx, wzy, wciY, gct, rmlA, rmlC, rmlB, rmlD*
19	*wzg, wzh, wzd, wze, wchA, wchO, wchP, wchQ, mnaA, rmlA, rmlC, rmlB, rmlD*
22	*wzg, wzh, wzd, wze, wchA, wchF, ugd, wcwV, whaB, wzy, wcwX, wzx, rmlA, rmlC, rmlB, rmlD*
23	*wzg, wzh, wzd, wze, wchA, wchF, wchV, wchW, wzx, wchX, gtp1, gtp2, gtp3, rmlA, rmlC, rmlB, rmlD*
33	*wzg, wzh, wzd, wze, wzx*
35	*wzg, wzh, wzd, wze, wciB, wzx, wciG*

5.7 Supplementary information (see attached USB drive)

Supplementary File 5.1. Study datasets

Supplementary File 5.2. Serotype 1

Supplementary File 5.3. Serotype 3

Supplementary File 5.4. Serotype 7

Supplementary File 5.5. Serotype 9

Supplementary File 5.6. Serotype 11

Supplementary File 5.7. Serotype 22

Supplementary File 5.8. Serotype 6

Supplementary File 5.9. Serotype 19

Supplementary File 5.10. Serotype 23

Supplementary File 5.11. Serotype 14

Supplementary File 5.12. Serotype 15

Supplementary File 5.13. Serotype 18

Supplementary File 5.14. Serotype 33

Supplementary File 5.15. Serotype 35

6. A comparison of the estimated core genomes of pneumococcal carriage datasets reveals differences between pneumococcal populations from different regions

Genomes for the three published pneumococcal datasets (UK, USA and Thailand), as well as the 1,000 non-pneumococcal *Streptococcus* species, were obtained from the rMLST database (see Chapters 4 and 5). Genomes from the VICE study were downloaded from the Brueggemann BIGSdb database (see Chapter 4).

6.1 Abstract

The availability of large well-characterised pneumococcal carriage datasets from different geographical locations allows for the comparison of estimated core genomes calculated using the Bayesian Decision model described in Chapter 3. The expectation and hypothesis tested was that the estimated core genomes of carriage pneumococci from each of these four geographical locations would be broadly similar in terms of the total number and the composition of the core genes. The aim of this study was to calculate the estimated core genome for each of four pneumococcal carriage datasets and, investigate the differences and similarities observed in the core genomes between datasets. A dataset of 3,121 genomes was assembled from three published studies and the VICE study. The estimated core genome was calculated for each dataset and the four estimates were compared by clustering the nucleotide sequences for each gene in each core genome. A putative function was assigned to each gene and the core genes common to all four datasets (the shared core genome) were used to construct a phylogenetic tree depicting the relationships between all the genomes included in the study. The pan-genome (core plus accessory genes) from each dataset was combined together and clustered to examine how each of the pan-genomes varied between the datasets. The analyses revealed a very

small shared core genome and showed that the core and accessory genomes of the Thailand dataset differed significantly from the other, more similar datasets. Examination of the genes unique to the Thailand dataset revealed 14 large genomic regions that included transposons and bacteriophages. Taken together these findings showed that the Thailand dataset had a much higher level of diversity than the other dataset, which were very similar to each other. This work suggested that there is significant genome-wide heterogeneity amongst pneumococcal populations and this has important implications for making broader conclusions about pneumococcal biology based on a single dataset.

6.2 Introduction

In Chapter 3, a Bayesian decision model method was described for estimating the core genome for datasets comprised of draft bacterial genomes. The analyses included the calculation of estimated core genomes for two different pneumococcal datasets including a published dataset of carriage genomes from Boston, Massachusetts (USA) [99]. Since that study was undertaken, two additional datasets of pneumococci recovered from healthy children in Southampton (UK), and the Maela refugee camp (Thailand) were sequenced and published [100, 247]. The genomes from the ongoing VICE study were also completed since the development of the Bayesian decision model. Thus, there were four pneumococcal carriage datasets for which estimated core genomes could be calculated, allowing for the comparison of distinct pneumococcal populations from four different geographical locations.

The pneumococci in the four carriage datasets included in this study were all collected from children under the age of seven, but in addition, the Thailand dataset also included pneumococci collected from the mothers of the children being sampled. The number of

children under the age of five in each of the four study locations varied from ~5,000 in the Maela refugee camp (located on the Thailand-Myanmar border) to ~40,000 in Boston. Southampton, and Reykjavik and the surrounding areas, where the majority of the pneumococci included in the VICE study were collected, had similar populations of 20,000 to 25,000 children under the age of five.

The hypothesis was that the estimated core genomes of carriage pneumococci from each of these four geographical locations would be broadly similar in terms of the total number and the composition of the core genes. Therefore the aims of this study were to: i) estimate the core genome for each of four carriage datasets; ii) compare the four sets of core genes; iii) identify and characterise the genes common to all four core genomes (the shared core genome); and iv) compare the genes that comprise the pan-genome of each dataset.

6.3 Methods

Datasets selected for analyses

The datasets analysed in this study are described in Table 6.1. Due to the large size of the Thailand dataset (n = 3,085; [100]), R (https://www.R-project.org/) was used to randomly sample 1000 genomes to include in this analyses. Metadata for the published genomes were manually extracted from the original publications. Complete lists of the pneumococcal genome data included in this study, with accession numbers and available metadata are listed in Supplementary File 1.

Table 6.1. Summary of study datasets.

Dataset	No. of genomes	Years of isolation	No. of STs	No. of CCs	Serotypes	Source of data
Iceland	987	2009-2014	99	48	31	[236], VICE study
USA	616	2001-2007	139	63	31	[99]
UK	518	2006-2011	128	45	43	[247]
Thailand	1000	2007-2010	211	115	61	[100]
Non-pneumococcal *Streptococcus* species	1000	Unknown	N/A	N/A	N/A	rMLST database

Sequence types (STs), clonal complexes (CCs) and serotypes

Multilocus sequence type (MLST) data were automatically extracted from each genome using BIGSdb [20] and STs were clustered into clonal complexes (CCs) using Phyloviz [332]. Serotypes for each of the pneumococci included in the study were assigned based on the genome sequence, using seqSerotyper.R.

Core genome analyses

Prokka [262] was used to predict coding sequences (CDS) in each genome and add annotation using a bespoke pneumococcal sequence database compiled for this study, which was based upon the available gene annotation data from all pneumococcal genomes in GenBank [271]. The resulting annotation files in gff format were then input into Roary and clustered using sequence identity thresholds of 70% and 90% [264]. Core genome thresholds were calculated for each dataset using the Bayesian core genome

method described in Chapter 3 (Table 6.2). Putative paralogues were removed from the analysis and the resulting core genes were extracted and aligned using MAFFT [376]. A custom script was used to construct a Venn diagram showing the distribution of core genes amongst the four datasets.

Example sequences of each core gene were extracted to create a set of core genes for each dataset. The four sets of core genes were then compared against each other and clustered in cd-hit using a similarity threshold of 90% [265]. Cluster of Orthologous Groups (COG) functional groups were assigned to each gene in the four core gene datasets using eggnog-mapper.py [377].

Sequence alignments for genes in the shared core genome were concatenated to create a core gene alignment that was then used to create a phylogenetic tree using FastTreeMP [267]. The tree was reconstructed to account for recombination using ClonalFrameML [268] and hierBAPS [378] was run on the core gene alignment to construct sequence clusters that were depicted on the final phylogenetic tree using iTOL [270].

Pneumococcal essential genes

In order to assess how many genes in the shared core genome were potentially essential to cell growth and survival, the names of genes previously identified as being essential in pneumococcus using Tn-seq were obtained from the original publication [379]. This data was used to extract the relevant amino acid sequences from the TIGR4 genome and cd-hit was used to compare the set of essential genes against the amino acid sequences of the genes in the shared core genome using a sequence identity threshold of 70%.

Pan-genome analyses

Nucleotide sequences for the genes in the pan-genome of each dataset were combined and clustered in cd-hit using a similarity threshold of 70% and an alignment threshold of 90%. The distribution of the shared and unique genes in the pan-genomes of each study dataset was represented by a Venn diagram that was constructed using a custom script.

1,000 genomes from 65 non-pneumococcal Streptococcal species were also selected for comparative analyses (Table 6.1; Supplementary File 6.5). Genome Comparator [20] was used to extract the rMLST gene sequences from the 1,000 Streptococcal and 3,121 pneumococcal genomes; these sequences were aligned using MAFFT and concatenated to create an alignment that was used to build a phylogeny using FastTreeMP. ClonalFrameML was used to reconstruct the tree and the final tree was annotated using iTOL.

Since the analyses revealed that the Thailand dataset appeared to be unusual (see below), gene names and genomic positions for the genes unique to the Thailand pan-genome were extracted and manually inspected to identify large genomic regions (>25 Kb) unique to this dataset. Nucleotide sequences for each of these large genome regions were extracted using Artemis [380] and compared against both GenBank [271] and the set of 1,000 non-pneumococcal genomes in rMLST to find homologous regions. Regions putatively identified as transposons were annotated using the *CONJscan* module of the mobile.pasteur.fr website. EasyFig [381] was used to create comparison diagrams between the extracted large regions and the homologous regions identified in either Genbank or rMLST.

6.4 Results

Description of the datasets used in the analyses

The Icelandic VICE dataset consisted of 987 genomes for pneumococci of 31 serotypes and 48 CCs (99 STs) collected between 2009 and 2014. PCV10 was introduced into Iceland in April 2010 and thus the dataset is comprised of pneumococci from both the pre- and post-vaccine period. The pneumococci were recovered from nasopharyngeal swabs taken from children 1-6 years old attending day-care centres in Reykjavik and the surrounding areas.

PCV7 was introduced in the United States in 2000. 616 pneumococci were collected and sequenced from nasopharyngeal swabs taken from children under the age of seven attending primary care facilities in Boston, Massachusetts and the surrounding area between 2001 and 2007. 31 serotypes were identified and MLST data revealed 63CCs (139 STs).

Pneumococci in the UK dataset (n = 518) were isolated from nasopharyngeal swabs from children or children of siblings ≤4 years old attending the Southampton General Hospital outpatient department between 2006 and 2011 and sequenced. PCV7 was introduced in the UK in 2006 and was superseded by PCV13 in April 2010; thus the UK dataset contains genomes from both pre-and post PCV-7 and pre- and post-PCV13 introduction. 43 serotypes and 45 CCs (128 STs) were identified.

Nasopharyngeal swabs were collected from a cohort of over 500 infants and 242 of their mothers between 2007 and 2010 in Maela, a refugee community close to the border of Thailand and Myanmar as part of a pre-vaccine longitudinal carriage study. 3,085 pneumococci were randomly selected for sequencing. From this dataset, 1,000 genomes

were randomly sampled for inclusion in this study. Characterisation of the 1,000 genomes revealed 115 CCs (211 STs) and 61 serotypes.

Estimated core genomes for each dataset

The number of genes in the pan-genomes of the UK, USA and Iceland datasets were very similar varying from 7,277 in the UK dataset to 7,425 in the USA dataset (Table 6.2). The Thailand dataset was unusual as there were approximately twice as many genes (n = 15,751) in the pan-genome as there were in each of the other three datasets. The percentage of genomes within each dataset that possessed each core gene ranged from ≥99.70% to ≥99.90%, consistent with the values calculated for other bacterial datasets in

Table 6.2 Results of the estimated core genome analyses.

Dataset	Pan-genome (n)	% of genomes that possess core gene	Putative paralogues removed (n)	Estimated core genes (n)
Iceland	7,301	≥99.85	8	1,059
USA	7,425	≥99.80	7	1,029
UK	7,277	≥99.70	6	1,052
Thailand	15,751	≥99.90	5	394

Chapter 3. The number of putative paralogues identified in the initial Roary analyses was minimal and varied from five in the Thailand dataset to eight in the Iceland dataset; these genes were removed from further analyses. As well as having a much larger pan-genome, the estimated core genome for the Thailand dataset was much smaller (n =394)

than the core genomes calculated for the USA (n =1,029), UK (n = 1,052), and Iceland (n = 1,059) datasets. The large pan-genome and smaller estimated core genome of the Thailand dataset was unexpected and suggestive of a greater level of diversity in this dataset compared to the others.

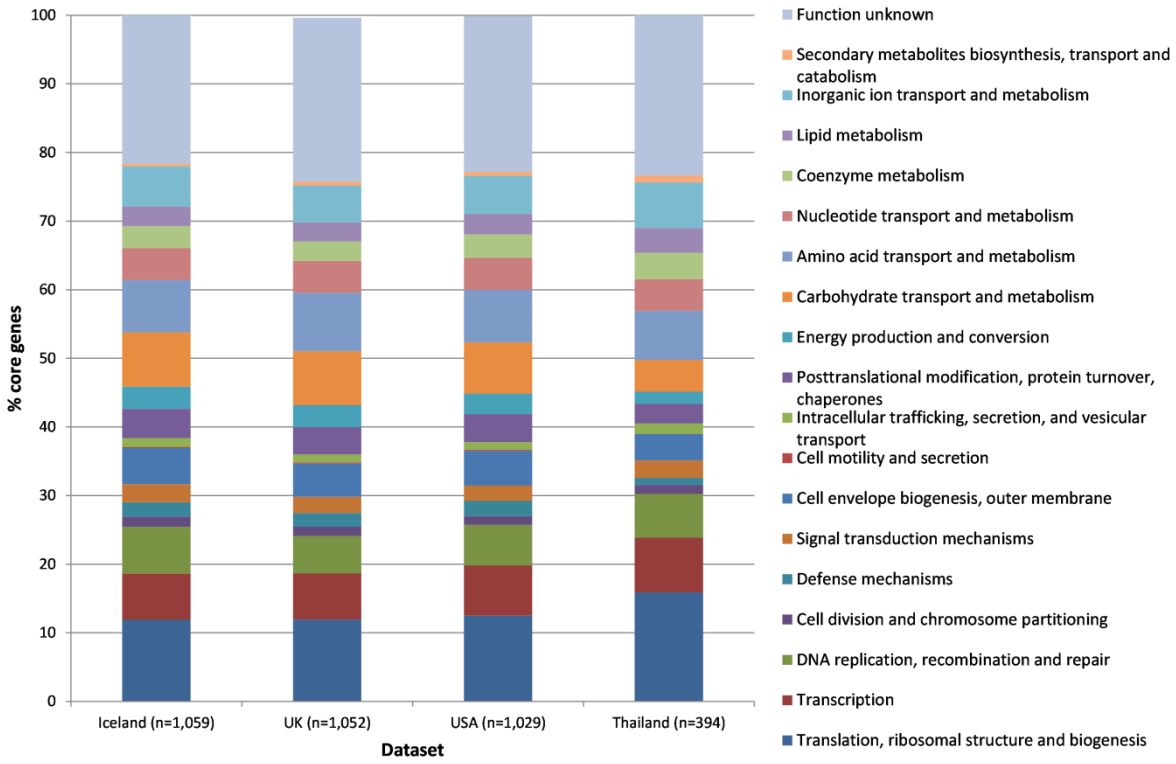
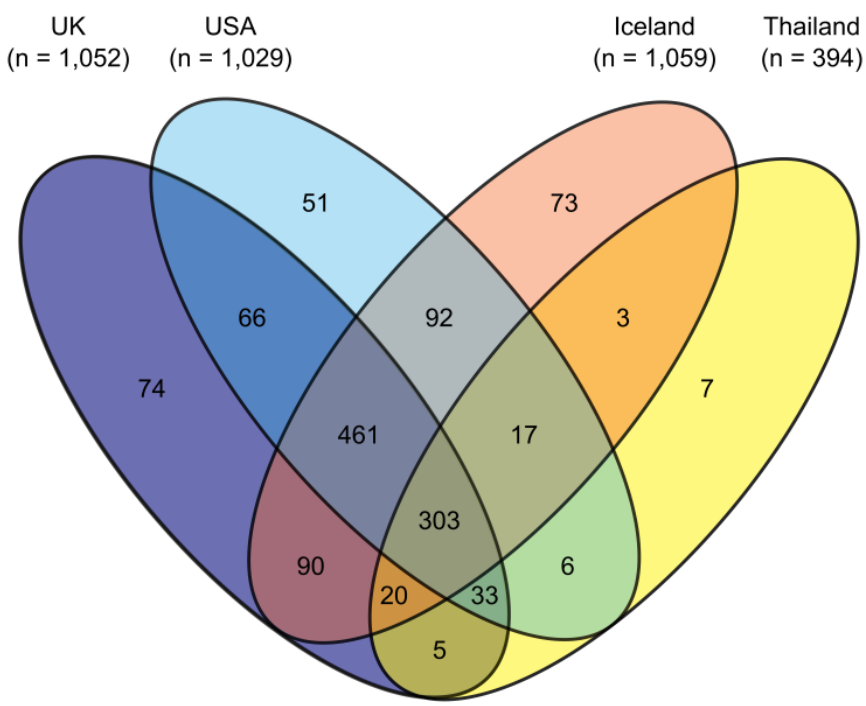


Figure 6.1 Cluster of Orthologous Groups (COG) functional categories for each set of estimated core genes.

Despite the differences observed in the number of genes in the estimated core genomes of the four datasets, the distribution of COG functional categories across the estimated core genomes for the four datasets was broadly similar (Figure 6.1). In each case the largest proportion of estimated core genes belonged to the function unknown category (21.7% - 24.1%) followed by the translation, ribosomal structure and biogenesis (11.9% - 15.7%), amino acid transport and metabolism (7.1% - 8.6%) and transcription (6.7% - 7.9%) categories (Supplementary File 6.3).

a)



b)

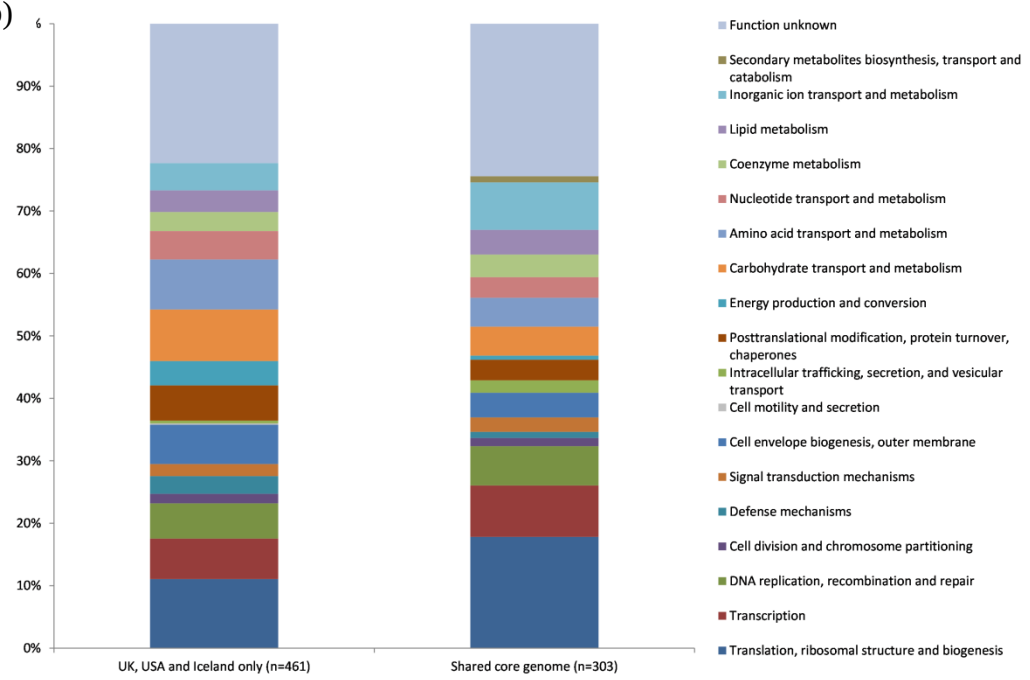


Figure 6.2 a) Venn diagram showing the comparison of the genes in the estimated core genomes calculated for each of the four datasets; b) COG functional categories for the shared core genome (n = 303) and the estimated core genes shared by the UK, USA and Iceland datasets only (n = 461).

Comparison of core genomes

The comparison of the genes in each of the four estimated core genomes revealed 303 genes that were common to all four datasets (defined as the "shared core genome"; Figure 2a). A further 461 genes were common to the UK, USA and Iceland datasets but were not present in the Thailand dataset *i.e.* not part of the shared core genome, but these genes would have formed part of a core genome calculated using just the datasets from the UK, Iceland and USA. The number of core genes unique to a given dataset varied from 7 (Thailand) to 74 (UK) and the largest number of core genes shared by two datasets was 92 (Iceland and USA).

The five most prevalent COG functional categories amongst the shared core genes were hypothetical proteins of unknown function (24.4%), translation, ribosomal structure and biogenesis (17.8%), transcription (8.3%), inorganic ion transport and metabolism (7.6%) and DNA replication, recombination and repair (6.3%; Figure 6.2b and Supplementary File 6.2).

Examination of the 461 genes that were common to the UK, USA and Iceland datasets but not present in the Thailand dataset revealed that the distribution of COG functional categories resembled that of the shared core genome. There was a higher proportion of genes belonging to the translation, ribosomal structure and biogenesis category in the shared core genome (16.8%) compared to the set of 461 genes (11.7%) whilst the proportion of carbohydrate transport and metabolism genes (8.2% vs 4.6%) and amino acid transport and metabolism genes (8.0% vs 4.6%) was higher in the set of 461 genes (Figure 6.2b).

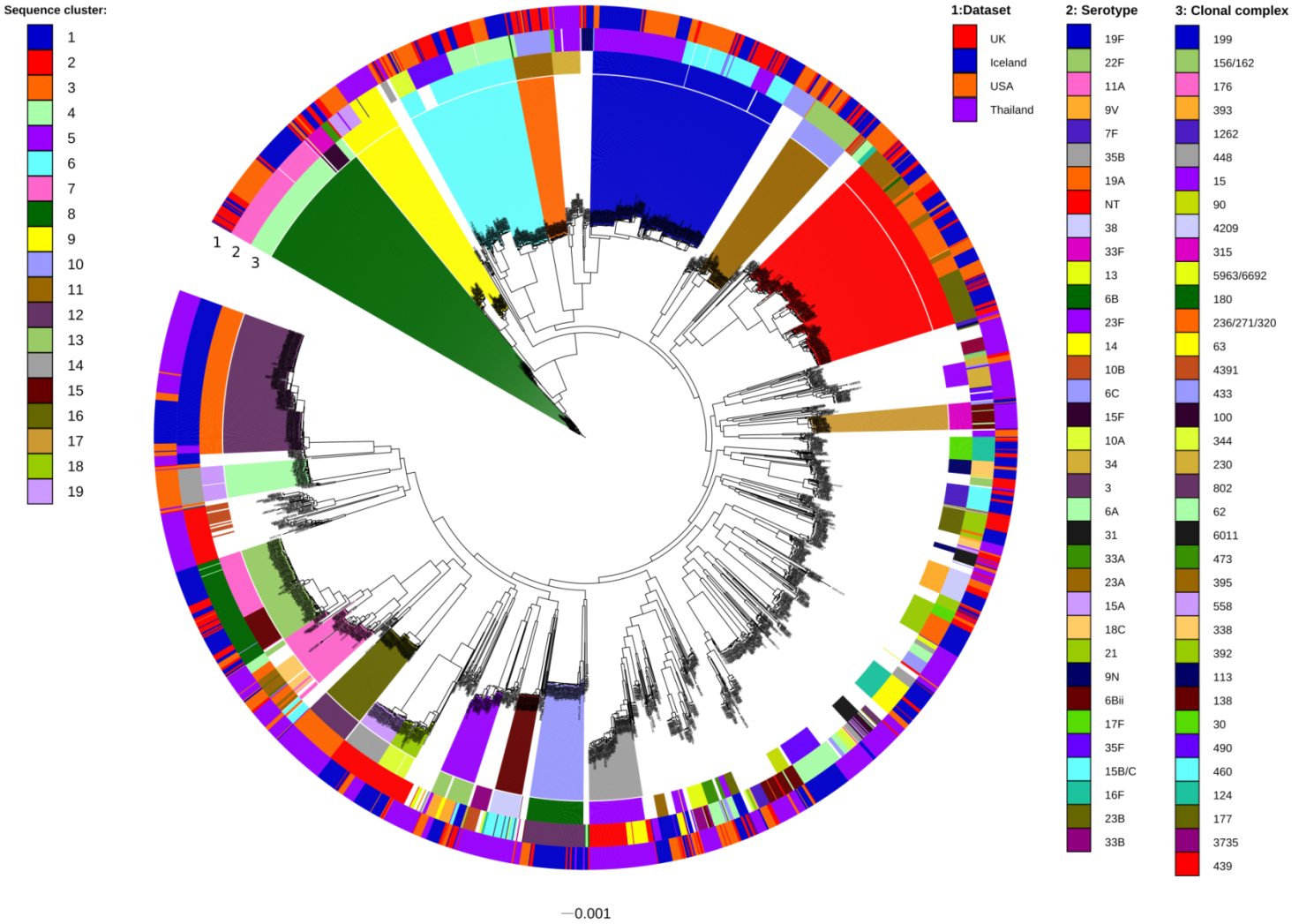


Figure 6.3 Phylogeny based on an alignment of 303 shared core genes. Clades are coloured according to SC. Ring 1 shows the dataset the genomes come from; ring 2, the serotype of each genome; and ring 3 the CC of each genome. The colours that represent each SC, dataset, serotype and CC are shown in the legends.

Shared core genome phylogeny

A sequence alignment of the 303 shared core genes and Bayesian population clustering was used to generate a phylogenetic tree of the genomes included in this study (Figure 6.3). This revealed 19 monophyletic sequence clusters (SCs) which corresponded to clades in the phylogenetic tree. These SCs ranged in size from 36 (SC17) to 263 (SC1) genomes. The population clustering also identified a twentieth group of genomes (not annotated on the phylogenetic tree) which made up a polyphyletic cluster of the remaining low-frequency genotypes (n=1,177). All of the monophyletic SCs were concordant with clonal complex (Figure 6.3). Nine of the 19 SCs contained no genomes from Thailand (SC1, SC2, SC3, SC4, SC6, SC8, SC10, SC11, SC13 and SC18; Figure 6.4) while three consisted solely of Thai genomes (SC14, SC15 and SC16) and; SC4 (CC558[35B]) was only found amongst genomes from the USA.

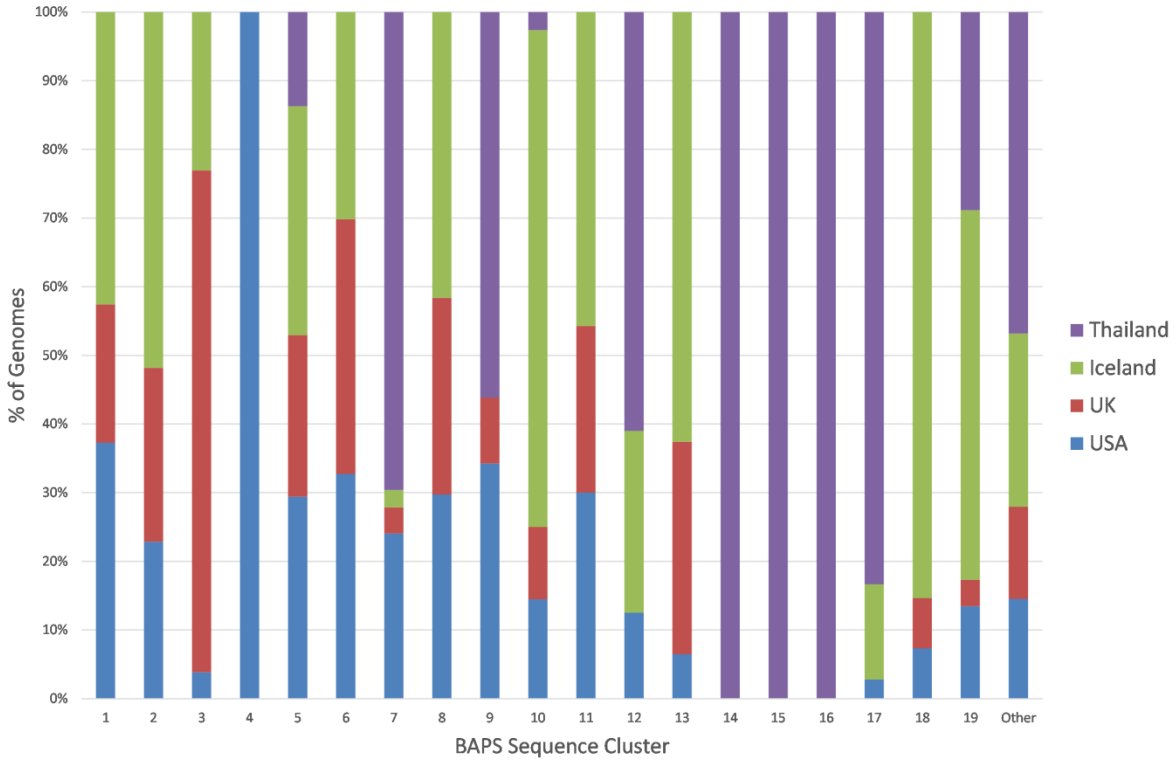


Figure 6.4 Distribution of the SCs amongst the four datasets

Examination of the serotype data for the three SCs unique to Thailand revealed that each cluster described pneumococci with either a novel or hybrid serotype (see Chapter 5): SC14 (CC15) genomes were either NT or hybrid serotype 14/19F; SC15 (CC4209) genomes had a putative novel serotype 15X capsule; and genomes from SC16 (CC802) were hybrid serotype 19F/23F. The most prevalent cluster, SC1 (n=263) was concordant with CC199 (PMEN37), a widely-distributed global lineage (http://pubmlst.org/spneumoniae/). Two other PMEN lineages, PMEN31 (SC10) and PMEN42 (SC18), were also amongst the sequence clusters not found in the Thailand dataset. 49.3% of the 1,177 genomes that did not form SCs were from Thailand and 26.6% were from Iceland.

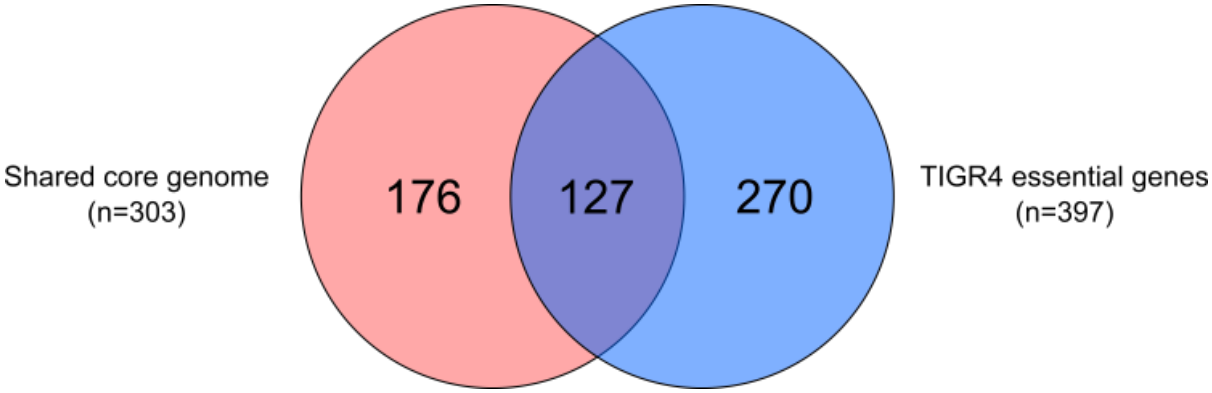


Figure 6.5 Venn diagram showing the comparison of the 303 genes in the shared core genome with 397 essential genes identified in TIGR4.

Pneumococcal essential genes

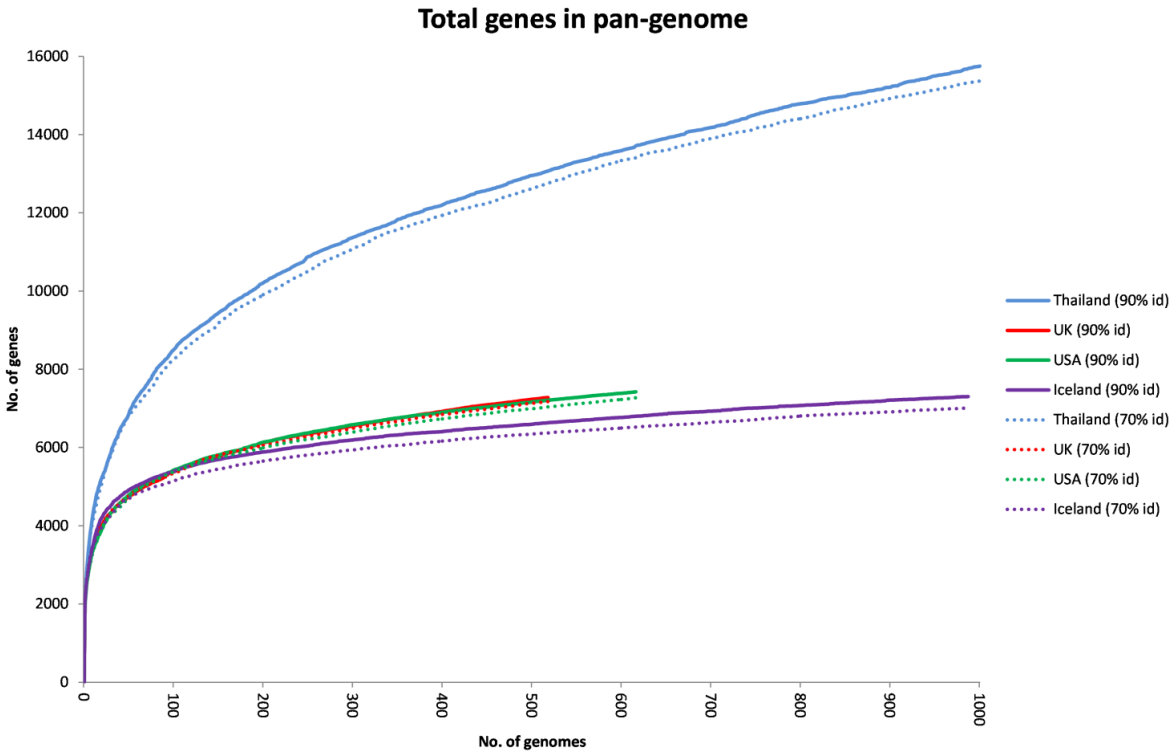
A previously published essentiality study used Tn-seq to identify essential pneumococcal genes [379]. In order to identify which of the shared genes were putatively essential, a comparison was made between the 303 shared core genes and 397 genes identified as

possibly or likely essential in the Tn-seq (Figure 6.5). Only 127 of the 397 genes identified as essential in TIGR4 were present amongst the shared core genome suggesting that the number of genes shared by all four datasets that could be considered essential for the survival of the pneumococcus is very small and that, potentially, the estimate for TIGR4 is inflated. A large proportion of the 127 genes were involved in basic cell functions such as DNA replication, ribosomal proteins, RNA transcription and central carbon metabolism (Supplementary File 6.2).

Comparison of the pan-genomes

The much larger number of genes in the Thailand pan-genome, and thus accessory genome, as well as the very small estimated core genome calculated for this dataset suggested that the Thailand dataset was unusual as compared to the other three datasets. The first attempt to understand the unusual nature of the Thailand dataset was to re-estimate the pan-genome using a lower blast identity threshold of 70%, which should have accounted for larger nucleotide differences between the same gene in a population (it has previously been shown that genes such as *pbp2x* may differ by as much as 25% at a nucleotide sequence identity level [382, 383]. However, there were no significant differences in the size of the pan-genomes calculated for any of the four datasets when using either threshold (Figure 6.6a) and in particular, the Thailand pan-genome was still twice the size of the other datasets. The pan-genome of the UK, USA and Iceland datasets began to plateau at approximately 7,000 genes whilst the Thailand pan-genome continued to increase sharply as more genomes were added and did not plateau even when the number of genes in the pan-genome approached 16,000. All four pan-genomes were open, *i.e.* the pan-genome continued to increase as more genomes were added to the analysis, a previously reported observation in pneumococcal pan-genome analyses [233].

a)



b)

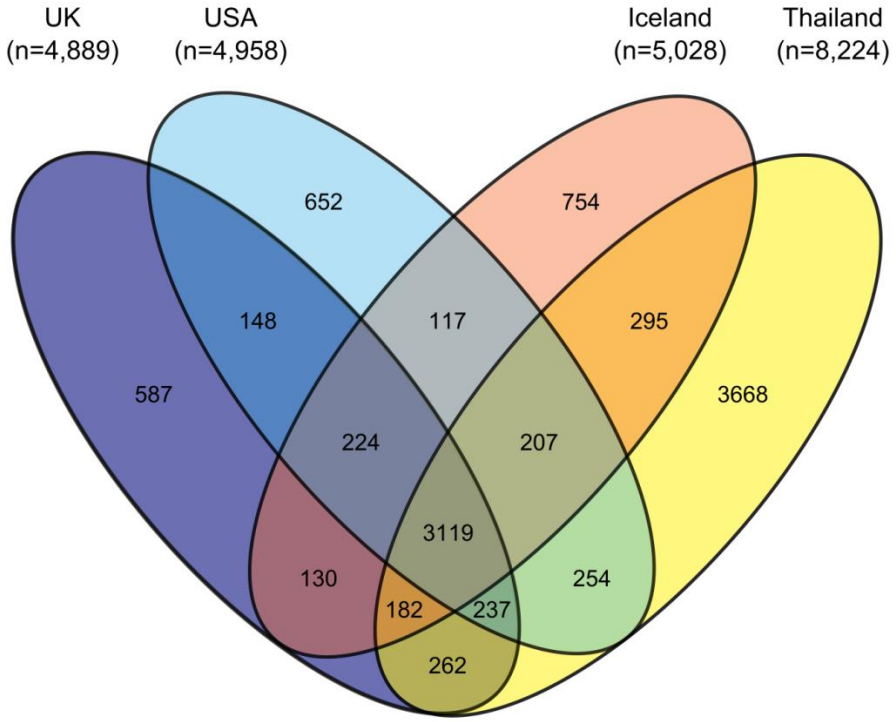


Figure 6.6 a) Results of Roary pan-genome analyses at 70% and 90% sequence identity for all four datasets; b) Venn diagram showing the comparison of the genes in each pan-genome of each dataset.

A gene-specific comparison of the pan-genomes of the four datasets was then performed. Clustering based on the nucleotide sequences of 37,754 genes from the four pan-genomes showed that there were 10,836 gene clusters (hereinafter referred to as genes) present amongst the four pan-genomes using thresholds of 70% identity and 90% alignment. 3,119 of these genes were shared by all four datasets (Figure 6.6b) and included the genes found in the shared core genome as well as the genes shared amongst the accessory genome of each dataset. The number of genes unique to each dataset differed, with smaller numbers observed in the UK (n = 587), USA (n = 682) and Iceland (n = 754) datasets compared to 3,668 unique genes in the Thailand dataset.

Like the pattern seen in the distribution of COG functional categories for the estimated core genomes, the distribution of COG functional categories across the four sets of genes unique to each dataset was similar with the most prevalent categories being function unknown (42.1% - 48.6%), DNA replication, recombination and repair (11.6 – 14.7%), cell envelope biogenesis, outer membrane (6.3% - 10.9%), carbohydrate transport and metabolism (4.8% - 6.7%) and defence mechanisms (4.6% - 6.0%; Figure 6.7 and Supplementary File 6.4).

The finding that the Thailand dataset had many more unique genes when compared to the other datasets confirmed the unexpected diversity observed in this dataset. As it was strange that a dataset collected in a closed refugee camp with a population of ~50,000 people showed so much diversity other possible explanations for this diversity were explored. These included the quality of the genome sequences in the Thailand dataset, an error in the sampling strategy when choosing 1,000 genomes out of the total dataset of 3,085 sequences and the inclusion of other non-pneumococcal Streptococcal genomes in the dataset.

Genome quality and sampling strategy of the Thailand dataset

All of the genomes for the Thailand dataset were downloaded from the ENA and assembled using Velvet. Genome sequences were uploaded to the rMLST database only if they passed the initial quality control steps, which considered total genome length and

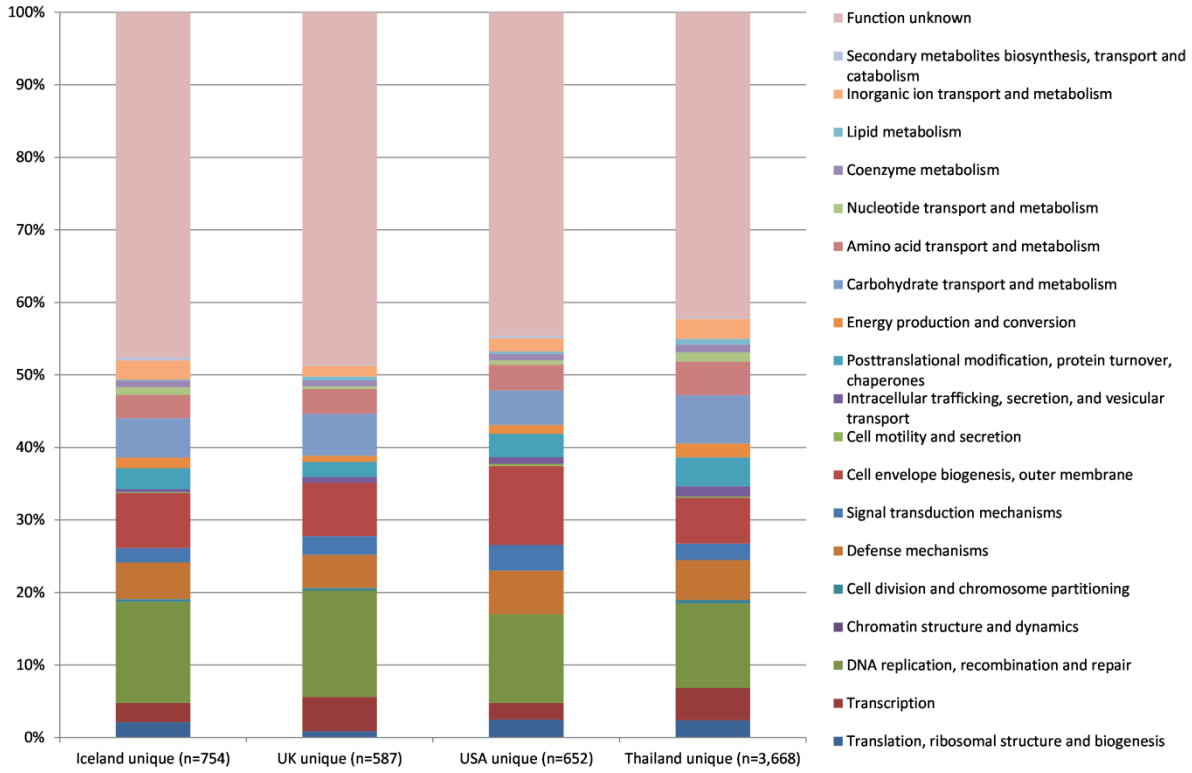


Figure 6.7 COG functional categories for the set of genes unique to the pan-genome of each dataset.

contig number and a scan of rMLST loci to assign the likely species of the genomes. 80 of the 3,085 genomes in the original dataset failed quality control and were discarded leaving 3,005 genomes to analyse. An examination of the sequence assembly metrics for the 1,000 genomes included in the study described in this chapter revealed one genome that was poorly assembled (~1,600 contigs as compared to an average of 222 contigs for the other genomes in the dataset). To test whether or not this genome significantly skewed the Roary pan-genome analyses, it was replaced by another genome of the same ST, serotype and rMLST type, and the analyses were re-run. The results were not

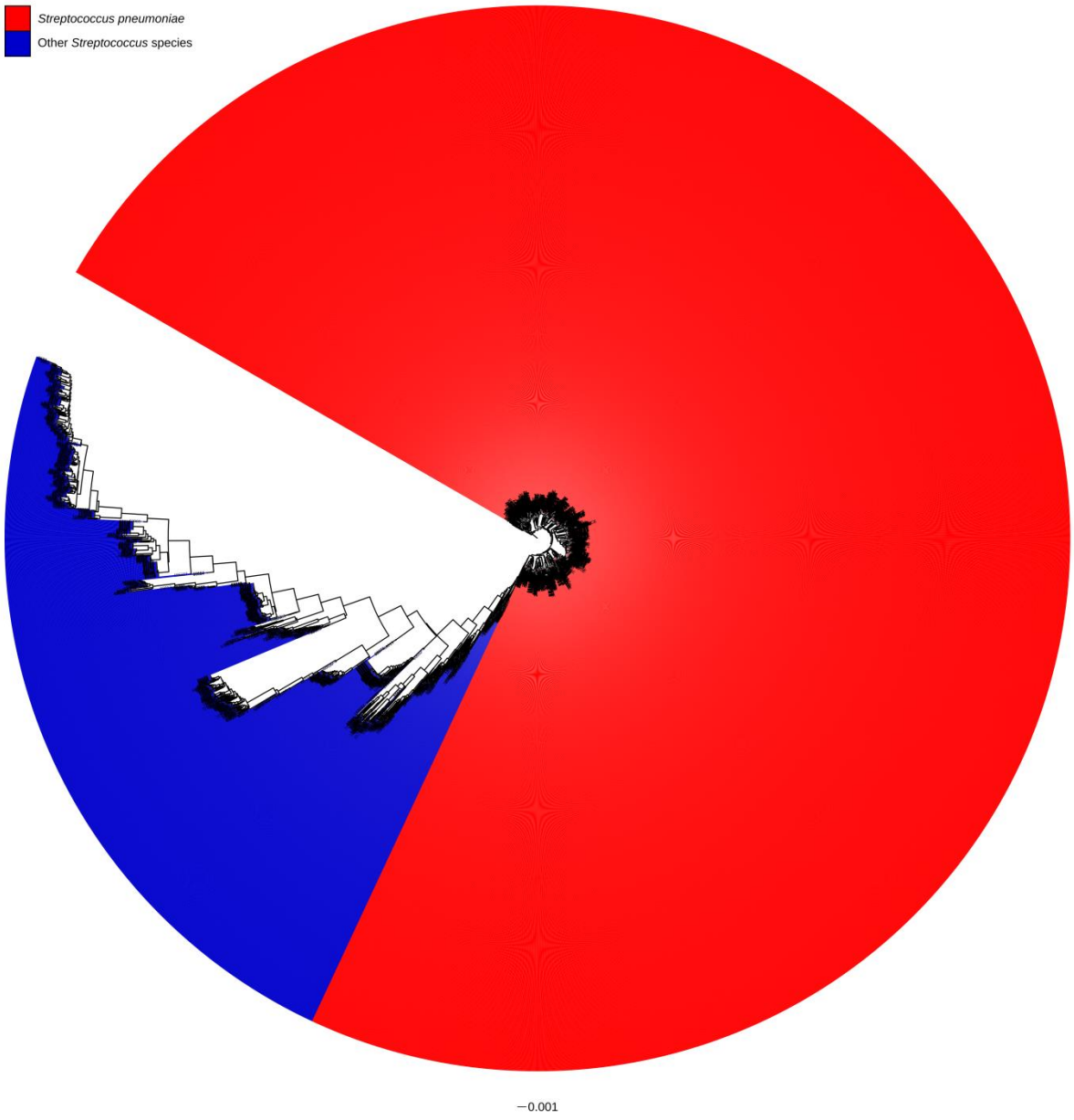


Figure 6.8 Phylogeny of 4,121 pneumococcal and non-pneumococcal *Streptococcus* genomes based on the concatenated sequences of 53 rMLST genes.

significantly affected: the core-genome increased by 4 genes and the pan-genome decreased by 64 genes (data not shown).

In order to check whether the sampling strategy used was appropriate, a further 1,000 genomes from the remaining ~2000 genomes were sampled in the same manner. The

number of STs and serotypes for this sample were nearly identical to the original sample dataset (data not shown) suggesting that the observed epidemiological diversity was similar to that of the other Thailand genomes not included here. Thus, the differences observed between the Thailand dataset and the other datasets could not be explained by poor quality sequence data or by a biased sampling strategy.

Calculation of the Thailand pan-genome after NT genomes were excluded

One possible explanation for the large pan-genome and small core genome of the Thailand dataset was the inclusion of a much higher proportion of NT genomes when compared to the other datasets (512/3,085 (16.6%) genomes in the original publication [ref] were NT whilst the proportion of NT pneumococci in the other three datasets varied between 1.3% and 6.6%). To investigate the possible contribution of excess NT genomes to the much larger Thailand pan-genome, the NT genomes were excluded from the original dataset and a random sample of 1,000 genomes was selected from the remaining ~2,500 genomes. Roary was then run on this dataset and a core genome was calculated. This analysis showed that excluding the NT genomes had relatively little effect on the size of the Thailand pan- and core genomes, as the size of pan-genome decreased from 15,751 to 14,537 genes and the size of the core genome increased from 394 to 441 genes.

Comparison of pneumococcal genomes to other Streptococcal genomes

Another possible explanation for the differences observed in both the size of the Thailand pan-genome and the much greater number of unique genes in the Thailand dataset was that the Thailand dataset contained genomes from non-pneumococcal *Streptococcus* species. To investigate this possibility, a phylogenetic tree was constructed based on the 53 rMLST loci sequences [250] extracted from the 3,121 pneumococcal genomes that comprised the four study datasets and those of a dataset of 1,000 genomes spanning 65

different *Streptococcus* species. All 3,121 pneumococcal genomes clustered together separately from the other *Streptococcus* species, demonstrating that the observed differences among the Thailand dataset were unlikely to be explained by the inclusion of non-pneumococcal genomes (Figure 6.8).

Characterisation of genes unique to the Thailand pan-genome

Based on the above findings, the inclusion of genomes from other Streptococcus species in the Thailand dataset was not responsible for the differences between this dataset and the other three datasets in the study. However, the size of the Thailand accessory genome was suggestive of the possibility that there were significant regions in at least a proportion of the Thailand genomes that may be non-pneumococcal in origin. To further explore this possibility, the 3,668 genes unique to the Thailand pan-genome were manually inspected using the gene identifier numbers assigned by Prokka and the gene frequency information provided by the Roary output. Briefly, the list of genes was sorted based on the gene identifier; runs of successive gene identifier numbers were indicative of larger genomic regions that were not found in the genomes from the other three datasets. This analysis revealed that there were 14 regions that were >25 Kb in length, the largest of which was ~66 Kb. The number of genomes that these 14 large regions were found in varied from one to 46. The nucleotide sequences of the regions were extracted and compared to GenBank and the dataset of 1,000 non-pneumococcal *Streptococcus* genomes in an attempt to find a possible match so as to better characterise these regions (Table 6.3).

Four different examples of Tn1549-like integrative and conjugative elements (ICEs), three of which also included Tn916 with Tet(M) which mediates tetracycline resistance, were found in 21 genomes (Figure 6.9). The nearest matches to these variable Tn1549-

like regions were predominately from other Streptococci but the best match to the Tn1549 region found in SMRU1170 was from *Filofactor alocis* ATCC35896, a gram-positive anaerobe commonly implicated in periodontal disease [384]. Tn916 was also found on its own in one genome and most closely matched a version of the ICE from *Staphylococcus aureus* 2395 USA500. Another type of ICE, TnGBS2, which uses a DDE transposase instead of a phage-like integrase for mobility and is commonly found in other Streptococcal species such as *S. mitis* and *S. oralis* but previously thought to be uncommon in pneumococcus [385], was found in a single genome, SMRU148.

No significant matches were identified in either Genbank or the set of 1,000 Streptococcal genomes for four different bacteriophage sequences and an unknown transposon fragment. Another focus of the Brueggemann group is the identification and characterization of pneumococcal bacteriophages. The four bacteriophage sequences identified amongst the Thailand dataset were compared to a database of 67 pneumococcal bacteriophage sequences assembled from a collection of 482 pneumococcal genomes, 336 of which were described in Chapter 3. This comparison failed to reveal any closely related sequences, likely reflecting the diversity observed amongst pneumococcal bacteriophages [386]. The bacteriophage regions varied in size from ~29 Kb to ~39Kb although only the two sequences from SMRU1351 and SMRU392 were full-length bacteriophage sequences. 46 genomes from multiple lineages possessed a region approximately 30 Kb in size that contained a number of genes involved in carbohydrate metabolism, including PTS lactose and ascorbate transporters as well as genes that constituted the pentose and glucuronate interconversion pathway. This pathway is an alternative to glycolysis for the metabolism of glucose [82]. Searching for this region in Genbank revealed a hit to *S. pneumoniae* 70585 ($ST289^5$), a disease causing pneumococcus from Bangladesh.

Table 6.3. Large genomic regions unique to the Thailand dataset. The best overall significant matches are highlighted in bold.

Representative genome	Genomes (n)	Length of region (bp)	Genbank best match (% identity)	rMLST best match (% identity)	Fragment type
SMRU1398	7	66142	***S. agalactiae* ILRI112 (98.0)**	*S. dysgalactiae* 66090 (96.4)	Tn1549 and Tn916 with Tet(M)
SMRU1170	2	59943 (1 gap)	***Filofactor alocis* ATCC35896 (99.0)**	*S. dysgalactiae* 65857 (89.8)	Tn1549 and Tn916 with Tet(M)
SMRU1457	10	51873	No significant match	***S. dysgalactiae* 65857 (97.4)**	Tn1549 and Tn916 with Tet(M)
SMRU2268	2	41961	***S. anginosus* C238 (93.0)**	*S. constellatus* 63991 (90.7)	Tn1549 no Tn916
SMRU1351	11	39520	No significant match	No significant match	Bacteriophage
SMRU2725	19	38392	No significant match	No significant match	Unknown transposon fragment
SMRU392	1	34256	No significant match	No significant match	Bacteriophage
SMRU158	1	32902 (2 gaps)	*Streptococcus* sp. VT162 (94.0)	***S. oralis* 63998 (95.6)**	Unknown transposon fragment
SMRU1017	2	32098	No significant match	No significant match	Partial bacteriophage sequence
SMRU128	46	30967	***S. pneumoniae* 70585 (99.0)**	*S. suis* 66662 (87.6)	Pentose and glucoronate interconversion region
SMRU148	1	30628	No significant match	***S. oralis* ATCC42996 (94.9)**	TnGBS2
SMRU1266	1	28998	No significant match	No significant match	Partial bacteriophage sequence
SMRU1770	3	26625 (3 gaps)	*Streptococcus* sp. VT162 (85.0)	***S. pseudopneumoniae* 110329 (98.5)**	Unknown transposon fragment
SMRU602	1	25230	***Staphylococcus aureus* 2395 USA500 (99.8)**	*S. dysgalactiae* 66058 (99.0)	Tn916 with Tet(M)

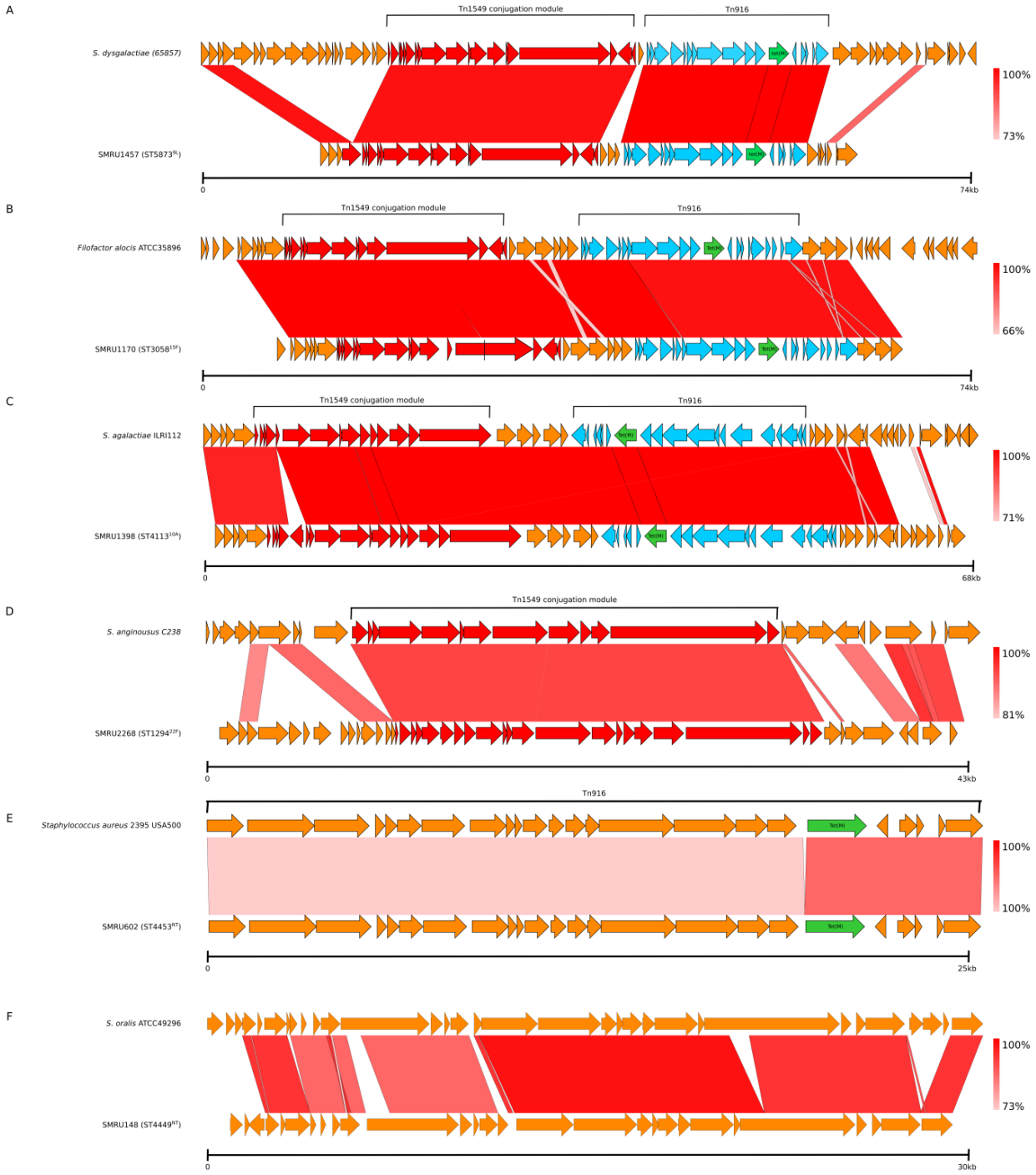


Figure 6.9 ICEs identified amongst the Thailand genomes**.** A-C) Tn1549-like ICEs with Tn916; D)Tn1549-like ICE without Tn916; E) Tn916; and F) TnGBS2

6.5 Discussion

This study began by utilising our Bayesian decision core genome model to estimate the core genomes for four pneumococcal carriage datasets. Comparing and contrasting these estimated core genomes revealed differences between each of the datasets and in particular the Thailand dataset. This study also revealed a very small shared core genome amongst the four datasets.

The expectation at the beginning of this study and the resulting hypothesis that was tested was that the estimated core genomes of pneumococcal populations from around the world would be broadly similar. The results of this study have shown that this not the case with one dataset in particular, from Thailand, showing considerable differences to the other three datasets included in the analyses. Whilst there was some variation in the size of the datasets included in this study (518 to 1000), the Thailand dataset was unusual as it had an unexpectedly small estimated core genome (n=394) even when compared to the estimated core genome of the Iceland dataset (n=1,059) which had an almost identical number of genomes (1000 vs 987).

The simplest explanation for this observed difference would be that the diversity of the pneumococci circulating in the Maela refugee camp is much greater than that in either large cities (Boston and Southampton) or what is effectively a national dataset from Iceland. There is certainly evidence from the epidemiological data from Thailand that shows that there are approximately twice as many STs and serotypes as there are in Iceland, which on this basis could be considered the least diverse of the four datasets. However, the explanation that there is simply a greater level of diversity amongst the pneumococci in Maela, a small refugee camp of only 50,000 inhabitants, where the movement of people in and out of the camp is restricted, seems at odds with what might

be expected. The expectation would be, that with very little flow of people in and out of the camp, there would be a similar bottleneck in the flow of pneumococci into the camp and thus the same lineages would have begun to prevail over time leading to a decrease in diversity as these lineages began to outcompete the others. The majority of epidemiological studies that have been conducted in developed or resource-rich countries have shown that the serotypes and sequence types (STs) that circulate in both carriage and disease are broadly similar across different populations [30, 242, 387, 388]. The fact that nine of the 19 SCs identified in this study were only found in the three Western datasets is consistent with the previous findings. It is possible that these lineages were simply not included by chance either during the initial collection of samples using nasopharyngeal swabs or else because of the sampling methodology used in this study to choose 1,000 genomes out of the original set of ~3000 genomes but this seems unlikely given the number of lineages concerned. More recent epidemiological studies that have examined the pneumococci circulating in less well characterised regions such as Kenya and Nepal revealed that, whilst the most common serotypes tend to be the same as those in the developed world, the diversity of circulating genetic lineages, based on the number of novel sequence types, may well be greater [57, 389]. However, given that these two studies were based solely on MLST analysis, it is impossible to draw any conclusions about the pan-genomes of these populations.

Interestingly, all of the SCs unique to Thailand had putative novel or hybrid serotypes previously identified in Chapter 5. Whilst some of these hybrid serotypes were found in genomes outside of Thailand, the analyses in Chapter 5 showed that almost all the putative novel and hybrid serotypes identified in the 14 serotypes and/or serogroups included in that analysis were from the Thailand dataset. This would suggest, that as well as there being more diversity amongst the Thailand pneumococci, there are almost

certainly lineages of pneumococci present in this population with novel genomic regions such as the capsular locus and the large regions described above that haven't been identified before in any other large genome datasets.

The difference between the Thailand dataset and the other three was further highlighted when the pan-genomes of the datasets were compared. The large number of genes unique to the Thailand pan-genome was of interest as there were approximately five times as many unique genes as there were in the other datasets. The focus of this part of the study was to find possible explanations for why the Thailand dataset was so different. Examination of the genes unique to the pan-genome of the Thailand dataset revealed large genomic regions that were obvious first choices for interrogation. The large regions that were identified were predominately from other Streptococci but in two cases the best match to these ICEs were from unrelated species (*Staphylococcus aureus* and *Filofactor alocis*); however, it is very likely that there are other genomic regions of interest to be found in the list of ~3,700 genes unique to the Thailand pan-genome so a more in-depth study of these regions as well as any other large regions identified in the pan-genomes of the other three datasets should be undertaken in the future.

This study has shown very clearly that population level datasets differ and may not be representative of the global pneumococcal population. Whilst there were definite similarities between the three datasets from the developed world, the Thailand dataset was clearly distinct. Considering the sampling strategy used to select genomes from Thailand for inclusion in this study and the possibility that non-pneumococcal *Streptococcus* species were included in the dataset could not explain the observed differences, suggesting that the differences were in fact real. Taken together, these findings suggest that the global pneumococcal population may be less homogenous than expected and that the current understanding of the level of diversity amongst

pneumococci has failed to take into account how much the pneumococcal populations found in regions that have tended to be underrepresented in studies before now may differ. Therefore, care should be taken when inferring broader conclusions about the pneumococcus based on the findings of studies on pneumococci collected in a relatively small geographical area as it is clear that not all pneumococcal populations are the same. The availability of large numbers of genomes means that meta-analyses of large datasets can now be undertaken in order to better understand the diversity of the entire global pneumococcal population. More studies like this one will need to be performed, incorporating genomes from less well-studied regions around the world, to better understand the true diversity of the pneumococcal genome.

6.6 Supplementary information (see attached USB drive)

Supplementary File 6.1. Datasets included in the study

Supplementary File 6.2. List of core genes for each dataset

Supplementary File 6.3. List of shared core genes

Supplementary File 6.4. Genes unique to each dataset's pan-genome

Supplementary File 6.5. Other Streptococcal genomes included in the analyses

7. A genome wide association study identified gene variants significantly associated with invasive disease in pneumococcus.

John Lees (Wellcome Trust Sanger Institute), the lead author of SEER, provided advice regarding the analyses performed in this chapter and helped resolve errors that occurred during the running of the SEER program. The genome sequences from the VICE study, Maela and Massachusetts were described in earlier chapters. My co-supervisor, Dr Danny Wilson, provided advice and guidance for the analyses conducted in this chapter.

7.1 Abstract

The exact mechanisms behind the ability of pneumococci to infect sterile sites outside of their ecological niche, the human nasopharynx, are not well understood. It may be possible to use a genome wide association study (GWAS) approach to identify putative virulence factors among pneumococci. Applying GWAS to bacterial genome datasets is a new approach and the first studies adapted programs designed for human GWAS. Programs specifically designed for performing bacterial GWAS have recently become available and one of these programs, SEER, was used to analyse the VICE study genomes to identify genes significantly associated with disease phenotypes (invasive disease, otitis media and pneumonia). The first step of the analysis was to validate SEER by identifying genes associated with penicillin non-susceptibility in the VICE study genomes. The results of this analysis were assessed in the context of the well-documented prior knowledge of the mechanisms behind pneumococcal penicillin non-susceptibility. Subsequently, three separate SEER analyses were performed to identify k-mers significantly associated with invasive disease, otitis media and pneumonia respectively. No k-mers were significantly associated with otitis media and pneumonia; however, 89 genes were identified as being significantly associated with invasive disease. The genes

with the most significant p-values were a cysteine desulfurase (*nifS*) and a tRNA 4-thiouridine(8) synthase (*thiI*). Other genes that were previously identified as being associated with pneumococcal virulence were also found among the list of candidate loci associated with invasive disease. The GWAS approach provides a powerful tool for identifying potential genes associated with invasive disease and these findings may be used to design appropriate experimental studies to confirm the associations.

7.2 Introduction

Genome-wide association studies (GWAS) were developed to test large numbers of genetic variants, such as single nucleotide polymorphisms (SNPs), k-mers (regular sized genome fragments) or indels for statistically significant associations with a given phenotype. The technique was originally developed to search for variants associated with disease in human genomes and has been used successfully for a number of years. The technique is now being extended to the analysis of bacterial genomes and the first bacterial GWAS was published in 2013 [238]. The first GWAS performed on pneumococcal genomes identified a number of genes associated with antibiotic non-susceptibility in two different pneumococcal genome datasets from Maela and Massachusetts [240]. These included genes associated with penicillin non-susceptibility such as those that form part of the peptidoglycan pathway (*pbp2x*, *pbp2b*, and *pbp1a*), as well as other genes associated with cell division (*ftsL* and *gpsB*) and recombination (*recU*). The results of this study were consistent with the role of altered penicillin-binding proteins in penicillin non-susceptibility, which has been extensively investigated for a number of decades [182].

Since performing a GWAS on bacterial genomes is a comparatively recent development, most studies have adapted programs developed for human GWASs to analyse the data.

Programs specifically designed for bacterial GWASs are now publicly available and the first of these is SEER (sequence element enrichment) [390]. SEER is a reference-independent pipeline able to find bacterial sequence elements in the form of k-mers (whole genome sequences are broken down into "words" with a length of 31 base pairs) associated with a particular phenotype whilst controlling for clonal population structure by applying metric multi-dimensional scaling (a type of principal component analysis) to a random subsample of k-mers [390]. Failure to correct for population structure (populations are considered to be structured if they contain sub-populations that are, on average, more closely related to each other than they are to members of the wider population i.e. the distribution of alleles within the population is non-random [239, 391, 392]) in GWASs may lead to false associations. These false associations are due to genetic relatedness rather than causality, when cases (samples with the phenotype being tested) are more closely related to each other than to the controls (samples with the control phenotype) [393]. The main advantages of using k-mers as opposed to SNPs (generated by comparing a set of genomes to a reference and identifying variant nucleotides) are: 1) k-mers allow for the core and accessory genomes to be considered simultaneously i.e. variation can be identified in regions not present in the reference genome but present in other genomes in the dataset; and 2) different types of variation including SNPs, indels and differences in gene content can be captured by using k-mers [239, 390].

The exact mechanisms behind the ability of pneumococcus to invade a sterile site such as blood or CSF are not well understood, although several virulence factors are known. These include the pneumococcal polysaccharide capsule (certain serotypes are significantly associated with IPD), pneumococcal surface proteins A (*pspA*) and C (*pspC*), and immunoglobulin A1 metalloprotease (*zmpA*) [394-396]. It may be possible

to use the GWAS approach to identify putative virulence factors among pneumococci. The VICE study dataset is the first population-level genome dataset to include both carriage and disease (invasive and non-invasive) pneumococci and includes over 1,900 genomes, both pre- and post-vaccine. This composition makes the VICE study collection of genomes a powerful dataset for looking for genetic variants associated with a phenotype such as invasive disease or otitis media, as it includes large numbers of both controls (carriage genomes) and case samples (e.g. invasive disease genomes).

The first aim of this study was to assess the performance of SEER. This was done by running SEER on the VICE study dataset to look for k-mers significantly associated with penicillin non-susceptibility. As the mechanisms for penicillin non-susceptibility in pneumococcus are well understood and primarily involve variants of the penicillin-binding proteins *pbp2x* and *pbp2b* [182], the SEER methodology was considered to be valid if it was able to successfully identify the genes previously identified as being implicated in penicillin non-susceptibility (significant k-mers were mapped back to a reference genome to identify these genes or candidate loci). The second aim of this study, following the successful validation of SEER, was to use SEER to search for k-mers, and thus candidate loci, that were significantly associated with three distinct phenotypes present in the VICE study dataset: 1) invasive disease; 2) otitis media; and 3) pneumonia.

7.3 Methods

Dataset used in this study

A total of 1,916 pneumococci were sequenced in the VICE study (Table 7.1 and Supplementary File 7.1). The dataset included 987 carriage genomes, 183 invasive disease genomes and 746 non-invasive disease genomes recovered from the lower

respiratory tract or middle-ear fluid. Invasive pneumococci were isolated from blood, cerebrospinal fluid (CSF), pleural fluid or joint fluid. Pneumococcal pneumonia isolates were those recovered from lung aspirates or sputum samples taken from patients with suspected pneumonia. Pneumococci associated with otitis media were cultured from swabs of discharging middle ear fluid from patients with suspected infections.

Table 7.1 Breakdown of the VICE study dataset

Carriage/disease	Source	Genomes (n)
Carriage	Nasopharynx	987
Invasive disease	Blood	172
	CSF	7
	Joint fluid	2
	Pleural fluid	2
Pneumonia	Lung aspirate	45
	Sputum	238
Otitis media	Middle ear fluid	463
Total		**1916**

Penicillin susceptibility values for all the Icelandic isolates were calculated as detailed in the General Methods chapter. Thirteen isolates were not tested for penicillin susceptibility and were removed from these analyses.

Phylogenetic analyses of VICE study pneumococci

Genes for each of the 1,916 genomes were predicted using Prokka [262] and the resulting output in gff format was used as input for Roary [264] which was run at a BLAST sequence identity threshold of 90%. Our Bayesian core genome model (see Chapter 3) was applied to the Roary results. Nucleotide sequences for each of the core genes were

extracted and aligned using MAFFT [376] before each of the alignments was concatenated to form a core gene alignment. FastTreeMP [267] was run on this core gene alignment using a nucleotide general time-reversible model to create an initial phylogenetic tree. ClonalFrameML [268] was then run using the initial phylogenetic tree and core gene alignment to reconstruct the phylogeny taking recombination into account. Hierarchical population clustering was performed using hierBAPS [378] and the resulting sequence clusters were annotated on the final phylogeny using iTOL [270]. The genomes were also annotated based on whether they came from carriage or disease as well as their penicillin susceptibility.

K-mer extraction, counting and filtering

DSK (disk streaming of k-mers) [397] was used to extract k-mers, 31 base pairs in length, from each genome in the dataset. K-mers for each genome were combined and counted using the combineKmers program included with SEER; non-informative k-mers and k-mers occurring in less than two samples were removed. To facilitate parallel processing, using 8 cores, the combined k-mer file was then split into 16 pieces.

SEER

In order to correct for the clonal population structure of bacterial populations, a program included as part of the SEER pipeline (kmds) builds a distance matrix from a random subsample of the k-mers on which multi-dimensional scaling is performed. This process was performed on all 16 files containing k-mers and a final distance matrix was created by reading in the 16 separate distance matrices to create the final distance matrix that was used as an input for SEER. Default settings were used except that the number of principal components used to correct for population structure was increased from the default of 3 to 20: the large number of genomes in the VICE study dataset required a

higher number of principal components to adequately capture the underlying structure of the population.

SEER also required a phenotype file containing the genome names and a binary phenotype represented by 0 for a control sample or 1 for a case sample. SEER was run with default values: a Minor Allele Frequency (MAF) >1% to filter out low frequency k-mers, an unadjusted p-value threshold of <10e-5 and an adjusted p-value of <10e-8. The SEER results were filtered a second time using filter_SEER to exclude k-mers that were significant but found in less than 5% of the genomes and those k-mers not associated with the phenotype being tested.

Downstream analyses

The significant k-mers identified in each analysis needed to be mapped back to a reference pneumococcal genome. Selection of the most appropriate reference genome was determined by mapping the filtered significant k-mers for each SEER analysis back to each genome in the VICE study dataset using Bowtie2 [398]. The genome to which the most significant k-mers mapped was used as the reference for that analysis. The significant k-mers were mapped to the reference sequence using BLAT [261] with a step size of 2 and a minimum match size of 15 which allowed for inexact but close matches to the reference (small differences in k-mers compared to the reference sequence would not prevent them being mapped). A script included with SEER, blast_top_hits.pl, was used to filter the results of the BLAT output to remove significant k-mers with low mapping scores. The coding regions for each reference sequence were predicted using Prokka and BEDTOOLS was used to merge the filtered BLAT output file with the annotated reference file to identify the coding regions with the most mapped significant k-mers [399]. Mapping significant k-mers back to an annotated reference file identified coding

regions significantly associated with the phenotype being tested (defined as candidate loci). Amino acid sequences for each of the candidate loci identified in this study were extracted and COG functional categories and KEGG pathways were assigned to these sequences using eggnog_mapper.pl [377].

A custom script was used to combine the Bowtie2 and SEER significant k-mer outputs to create a file of significant k-mers, their genome positions relative to the reference sequence and their associated p-values. The resulting output was used to create a Manhattan plot showing the results of mapping the significant kmers to the reference sequence along with the associated p-values in R using the *gwasplot* function of the *qqman* library [400]. The final figures were edited and annotated using Inkscape [401].

Comparison of previously published candidate loci associated with penicillin non-susceptibility to the candidate loci identified in this study

Amino acid sequences for candidate loci associated with penicillin non-susceptibility in from two datasets, Maela (n = 3,085) and Massachusetts (n = 616) were extracted [240]. Since the previously published analyses were performed using a SNP-based methodology, a comparison of the results generated by the two different GWAS methodologies was possible. In order to identify overlapping candidate loci between the three datasets, the amino acid sequences of candidate loci identified in this study were combined with the amino acid sequences of the candidate loci from the previous studies and a clustering analyses was performed using cd-hit [265] at a sequence identity threshold of 70%.

7.4 Results

Phylogenetic analyses of VICE study pneumococci

Our Bayesian core genome model estimated 946 core genes. A concatenated sequence alignment of these 946 core genes and Bayesian population clustering was used to generate a phylogenetic tree of the VICE genomes included in this study (Figure 7.1). This revealed 19 monophyletic sequence clusters (SCs) that corresponded to clades in the phylogenetic tree. These SCs ranged in size from 23 (SC2) to 295 (SC16) genomes. The population clustering also identified a twentieth group of genomes (not annotated on the phylogenetic tree) which made up a polyphyletic cluster of the remaining low-frequency genotypes (n = 255). Penicillin non-susceptibility was associated with three lineages, SC2, SC16 and SC19 which represented the globally-distributed MDR clones PMEN2 (Spain^{6B}-2), PMEN14 (Taiwan^{19F}-14) and PMEN42 (Norway^{NT}-42) respectively. SC16 was predominately composed of pneumococci isolated from the middle ear fluid and these genomes formed the majority of the case samples in the SEER analyses that looked for k-mers significantly associated with otitis media.

Identification of loci associated with beta-lactam non-susceptibility

A total of 1,903 genomes were analysed as part of the penicillin non-susceptibility SEER analysis: 448 case (non-susceptible) samples and 1455 control (susceptible) samples. After filtering, the results of the SEER analysis revealed that were 8,122 k-mers significantly associated with penicillin non-susceptibility (Figure 7.2). Mapping back the significant k-mers to the genomes in the dataset revealed that the genome with the highest number of mapped significant k-mers was VICE0279 (CC230^{19A}, penicillin non-

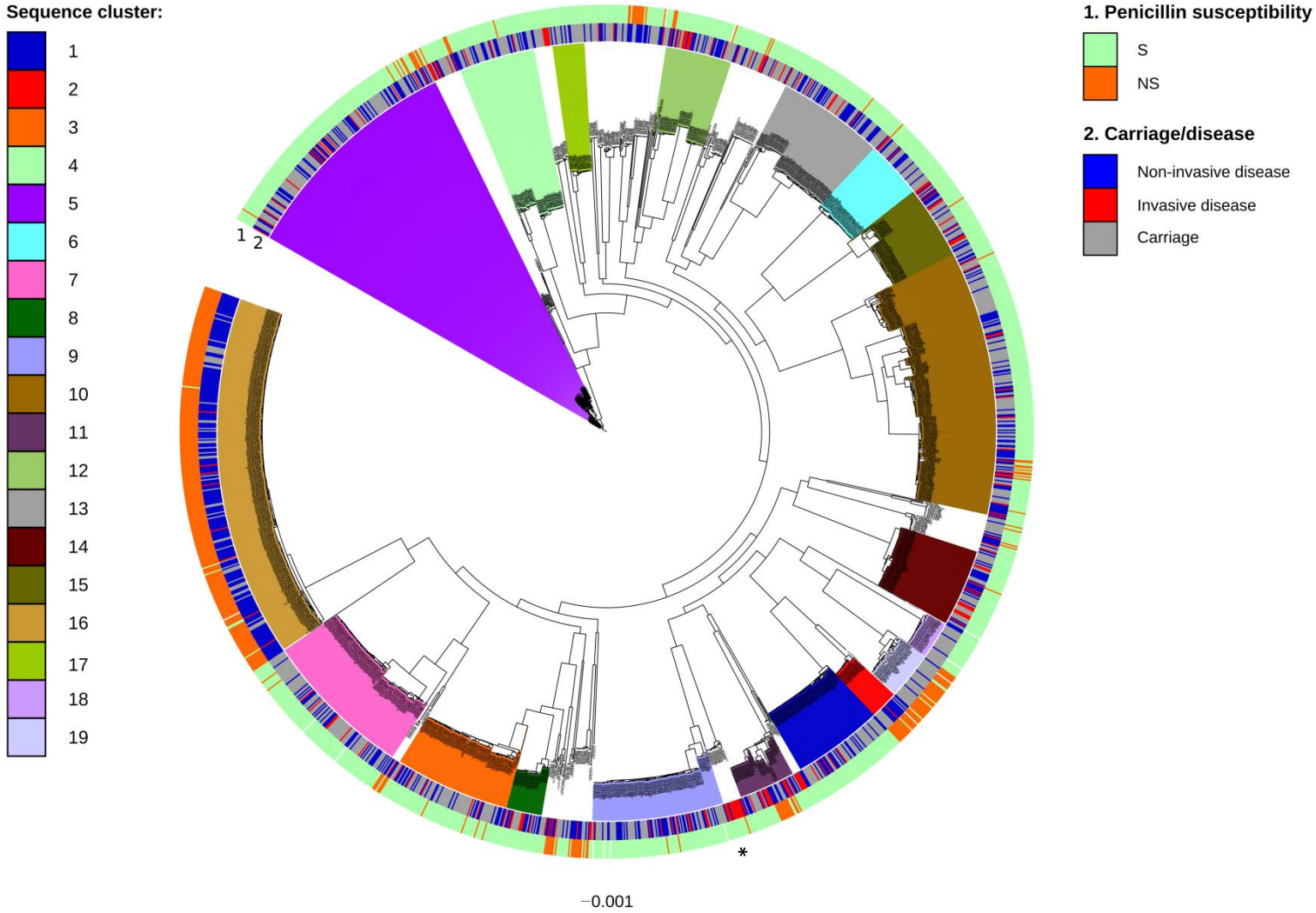


Figure 7.1 Phylogenetic tree of 1916 VICE study genomes based on a core gene alignment of 946 genes. Clades are coloured according to sequence cluster. Ring 1 shows the penicillin susceptibility of each genome (S = susceptible; NS = non-susceptible) and ring 2 shows whether the genome was isolated from carriage or from disease. * marks the CC191[7F] lineage.

susceptible; 85.7% of significant k-mers mapped) so this genome was used as the reference. The significant k-mers mapped back to 433 candidate loci in the reference genome and 51 of these candidate loci had at least one significant k-mer that was over the significance threshold of $-\log_{10}(p) = 8$ (Supplementary File 7.2). The top five candidate loci amongst the list of 51 genes with mapped significant k-mers above the significance threshold of $-\log_{10}(p) = 8$ were *pbp2X*, a hypothetical protein, a membrane protein, *ribE* and *pbp2B*.

The penicillin binding proteins, *pbp2X* and *pbp2B* form part of the peptidoglycan synthesis pathway (cell wall biosynthesis) and many previous studies have shown that amino acid alterations to these genes increase pneumococcal resistance to β-lactams such as penicillin [172]. Alterations to a third gene, *pbp1A* confer high level resistance to penicillin if the pneumococci already have altered *pbp2B* and *pbp2X* genes [177], but there were no significant k-mers associated with this gene in this study. Two other candidate loci adjacent to *pbp2X* and involved in cell division and cell wall biosynthesis, *ftsL* and *mraW* were also significantly associated with penicillin non-susceptibility. Thirteen of the candidate loci, including two of the five genes with the highest significant p-values were hypothetical or membrane proteins with an unknown function (Supplementary File 7.2). The remainder of the candidate loci were distributed across the reference genome and included genes involved in recombination (*recJ* and *recO*), riboflavin metabolism (*ribE*) and DNA replication (*ssb* and *dnaD*).

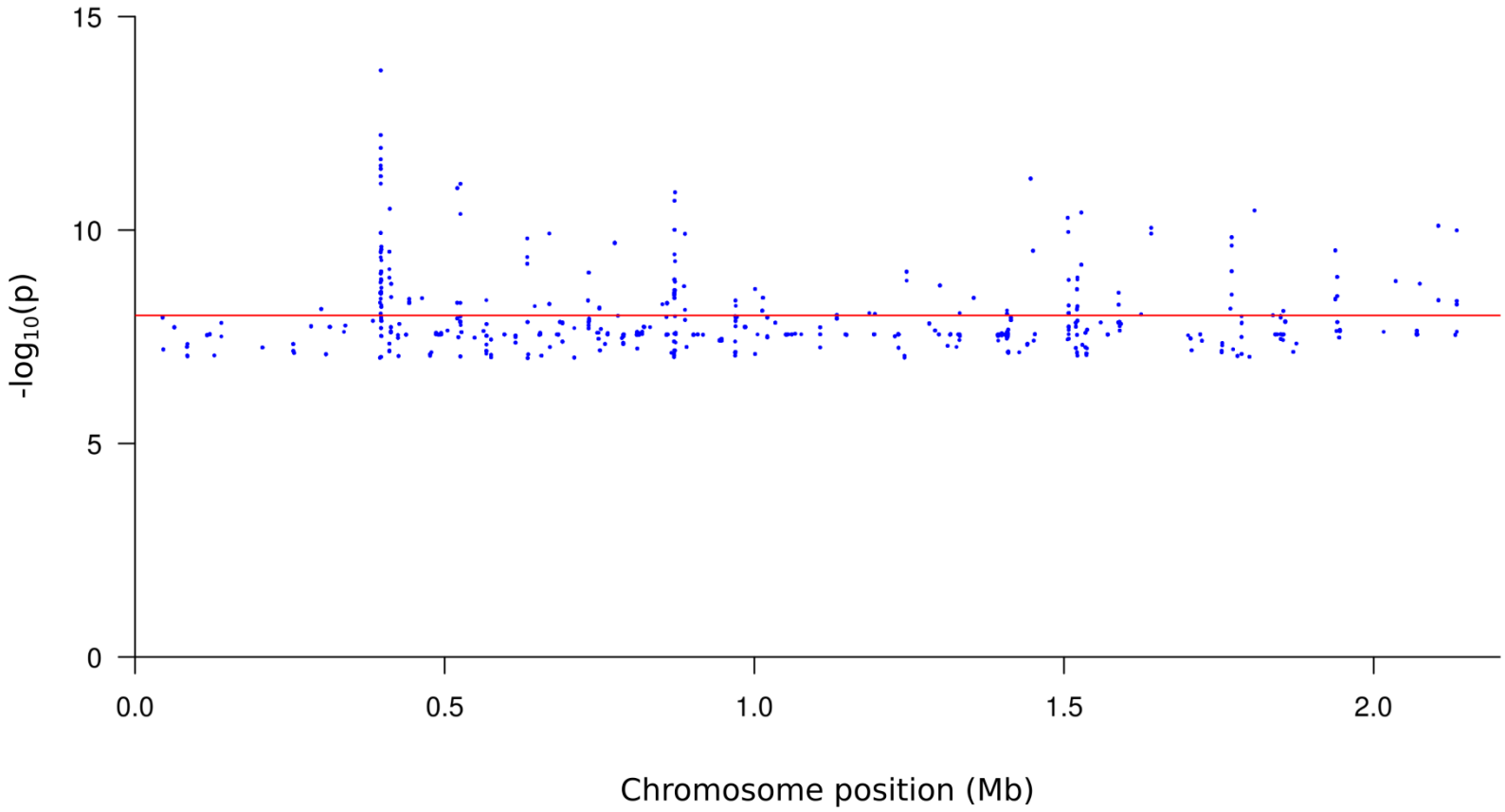


Figure 7.2: Manhattan plot showing the statistical significance of association (y-axis) for each significant k-mer associated with penicillin non-susceptibility arranged in order in the genome (x-axis). The horizontal red line indicates a significance cut-off after Bonferroni correction of p = 10e-8.

Comparison of penicillin non-susceptibility candidate loci

The amino acid sequences of the candidate loci identified in the VICE, Maela and Massachusetts datasets were extracted and compared and only four candidate loci were found to be shared by all three datasets: *pbp2B*, *pbp2X*, *mraW* and *ftsL*. This suggested that penicillin non-susceptibility in pneumococcus may largely be due to alterations in

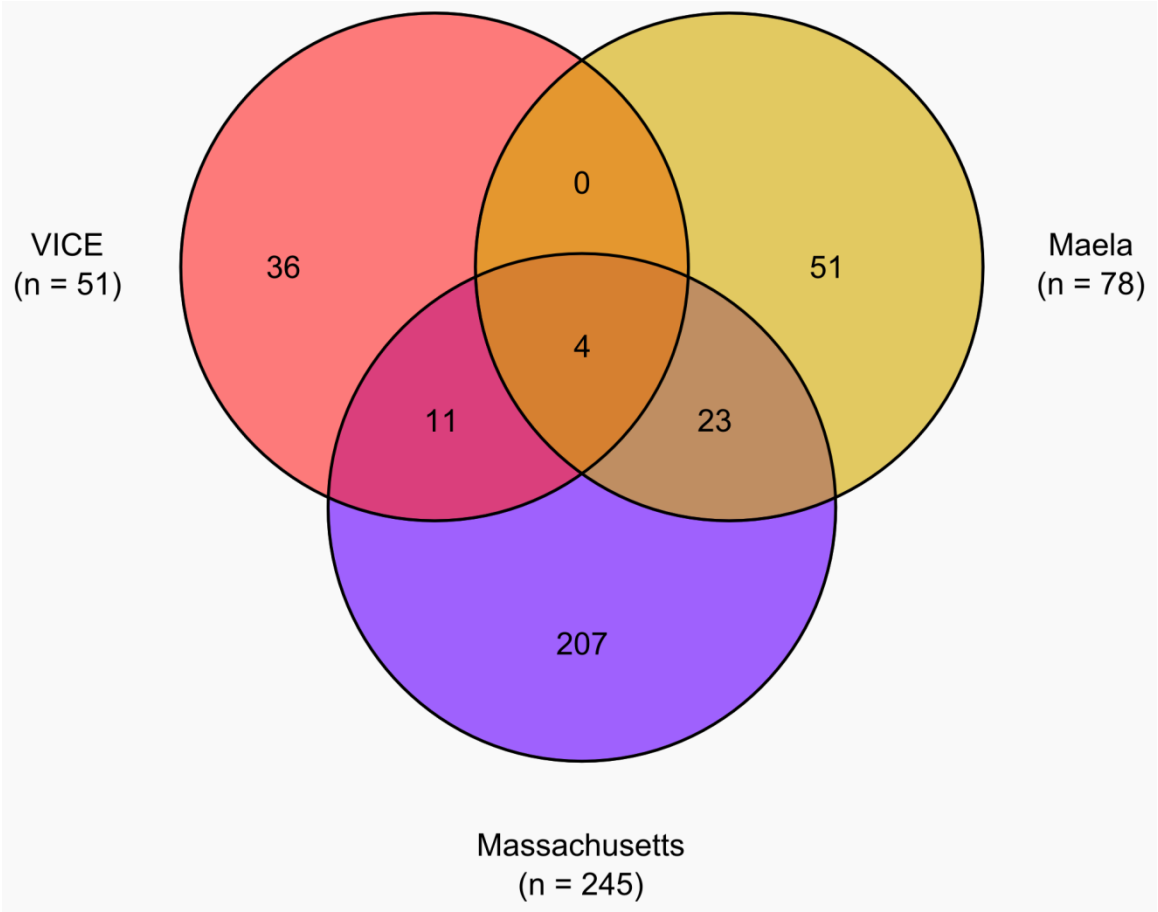


Figure 7.3 Venn diagram showing the comparison of candidate loci associated with penicillin non-susceptibility identified in this study (VICE) and previous studies (Maela and Massachusetts)

these four genes, whereas the remainder of the candidate loci were false positives or "hitchhiker" genes that were in linkage disequilibrium with the common four genes. Unusually, the majority of the candidate loci in the Massachusetts dataset (207/245) were

unique to that dataset and in their original publication, the authors attributed this finding to a higher number of non-causative SNPs or hitchhikers in the Massachusetts dataset when compared to the Maela dataset as well as the possibility that there was also an imperfect control for population structure in the Massachusetts dataset analysis [240].

SEER analyses

SEER was used to search for k-mers significantly associated with three distinct phenotypes: invasive disease, otitis media and pneumonia in the VICE study dataset (Table 7.2). The only test that identified k-mers significantly associated with the given phenotype was the invasive disease (carriage versus invasive disease genomes) analysis. No significant k-mers were identified for the pneumonia (carriage versus lung aspirate and sputum genomes) and otitis media (carriage versus middle ear fluid genomes) analyses.

Table 7.2. Results of SEER analyses

SEER analysis	Control samples (n)	Case samples (n)	Significant kmers (n)	Filtered significant kmers (n)	% filtered significant kmers mapped to reference
Penicillin NS[a] vs penicillin S[b]	1455	448	29931	8122	87.1
Invasive vs carriage	987	183	8503	5605	93.9
Otitis media vs carriage	987	463	0	0	N/A
Pneumonia vs carriage	987	282	0	0	N/A

a) NS = Non-susceptible (Intermediate and resistant); b) S = Susceptible

Identification of loci significantly associated with invasive disease

A total of 1,170 genomes (987 controls and 183 cases) were analysed to search for genetic variants associated with IPD. 5,605 significant k-mers were associated with invasive disease after filtering and VICE0274 ($CC191^{7F}$; 93.9% of significant kmers mapped) was chosen as the reference genome. The significant k-mers mapped to 370 candidate loci and 89 of these candidate loci were above the significance threshold of $-\log_{10}(p) = 8$ (Figure 7.4 and Supplementary File 7.3). The top five candidate loci were *nifS*, *thiI*, two hypothetical proteins and an integral membrane protein.

There is no published evidence for the function of the adjacent genes *nifS* (cysteine desulfurase) and *thiI* (tRNA 4-thiouridine(8) synthase) in the pneumococcus. In other bacterial species such as *Salmonella* Typhi, and *Escherichia coli*, ThiI acts as a sulfurtransferase receiving the sulfur donated by the cysteine desulfurase before transferring the sulfur to the target molecule or other sulfur carrier proteins [402]. ThiI is required for thiazole synthesis in the thiamine biosynthesis pathway [403] and is also involved in the tRNA thionucleoside modification, 4-thiouridine, which acts as a photosensor of near-UV radiation [404]. In *Salmonella* Typhi and *E. coli*, ThiI proteins contain three domains: a THUMP domain which binds tRNA, a PP-loop pyrophosphotase which is involved in the formation of the tRNA adenylated-uridine intermediate, and the rhodanese (Rhd) domain which provides the sulfurtransferase site Cys456 [402]. In Firmicutes (which include pneumococci), however, the Rhd domain is missing and the role of this domain is played by the upstream cysteine desulfurase. Gene inactivation experiments in *Bacillus subtilis* showed that *nifS* and *thiI* were required for the biosynthesis of 4-thiouridine but not for thiamine biosynthesis [402]. Closer examination of the position of significant mapped k-mers relative to *nifS* and *thiI* showed that the significant k-mers were broadly distributed across the length of *nifS* whilst the majority of

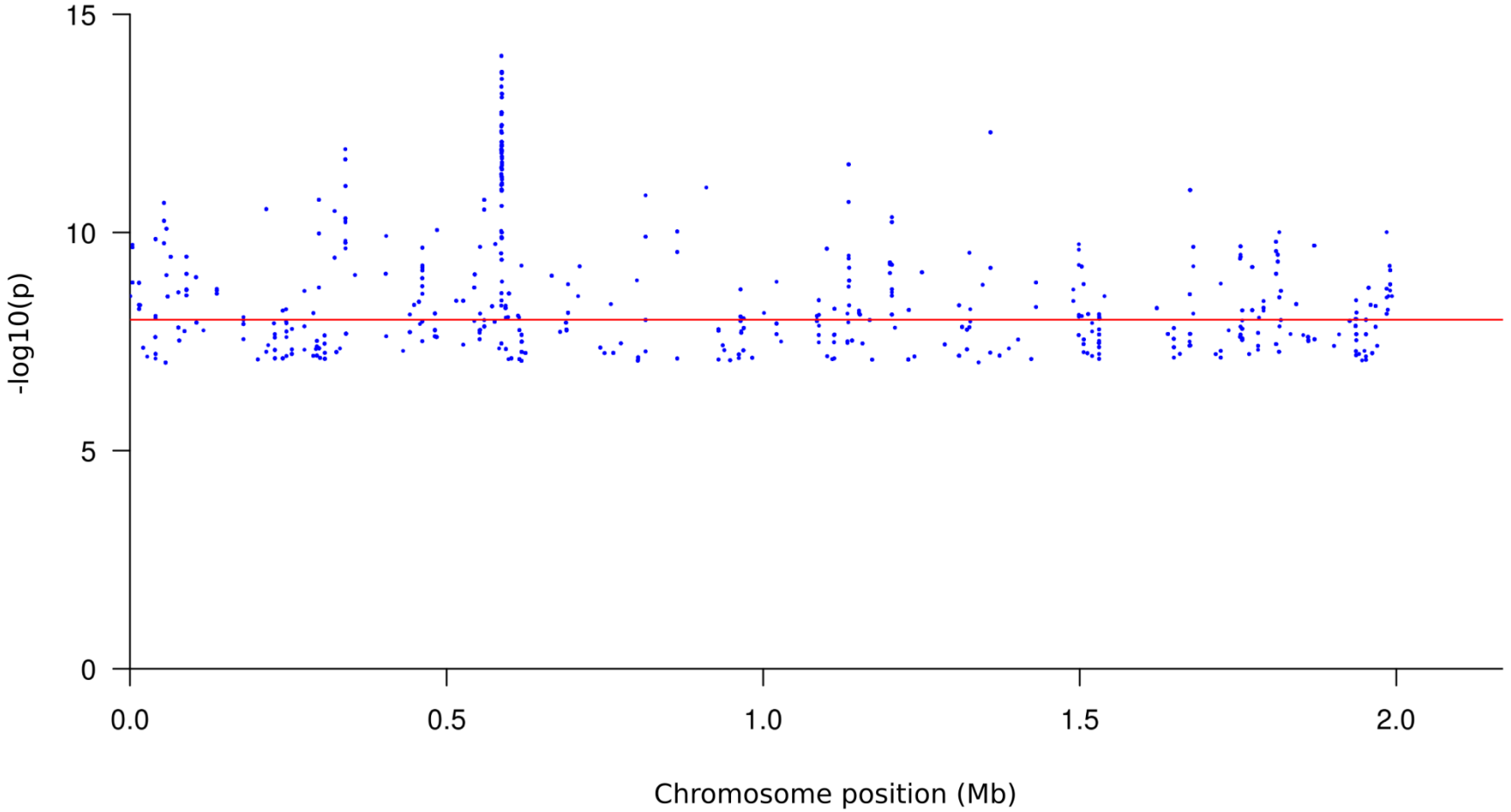


Figure 7.4. Manhattan plot showing the statistical significance of association (y-axis) for each significant k-mer associated with invasive disease arranged in order in the genome (x-axis). The horizontal red line indicates a significance cut-off after Bonferroni correction of p = 10e-8.

the k-mers that mapped to *thiI* were at the 5' end of the gene and in particular, the THUMP domain (Figure 7.5a). This domain plays a role in RNA modification and it is unknown what effect alterations to the THUMP domain may have.

A number of the candidate loci have previously been identified as having a role in pneumococcal virulence or pathogenesis: *pspA*; choline-binding proteins *cbpE* and *cbpG*; type I restriction modification system genes *hsdM, hsdR* and *hsdS*; glucose-1-phosphate uridylyltransferase *galU*; a branched-chain amino acid aminotransferase *ilvE*; and three conserved *cps* locus genes, *wzg*, *wzd* and *wze*.

PspA (pneumococcal surface protein A) is a known pneumococcal virulence factor that has been shown to have two roles: it prevents the binding of complement component C3 to the pneumococcal cell surface, thus inhibiting complement-mediated opsonophagocytosis [394], and it also binds lactoferrin which reduces the bacteriocidal effects of apolactoferrin [395]. *cbpE* (choline-binding protein E) is a receptor for human plasminogen (Plg)/plasmin and has been shown to facilitate the migration of pneumococcus through tissue barriers such as the epithelium and endothelium layers [405]. Variants of *cbpG* (choline-binding protein G) have shown to increase adherence to eukaryote cells and vaccination with *cbpG* provided a small but statistically significant degree of protection against colonisation and substantial protection against systemic infection in mice [406].

hsdM, hsdR, hsdS are co-transcribed together as part of the SpnD39III Type 1 restriction-modification (RM) system and genetic rearrangements within this RM system, *hsdS* in particular, have been shown to be the underlying mechanism for phase variation in pneumococcus [125]. The same study showed that allelic and, thus phase variants, of SpnD39III demonstrated differences in virulence: some alleles were more susceptible to

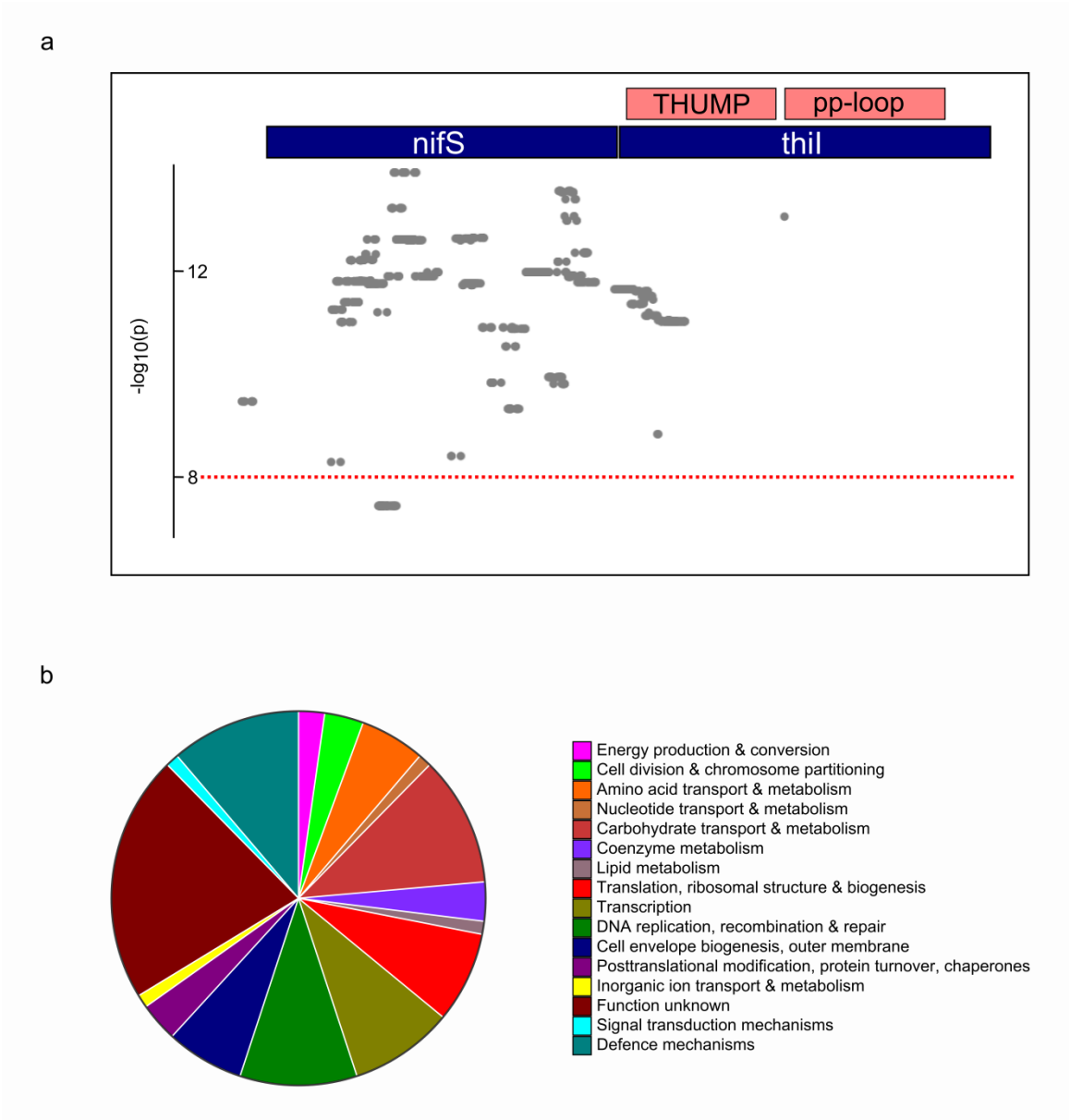


Figure 7.5 a) Zoomed-in view of the significant k-mers (grey) that mapped to *nifS* and *thiI* (in blue). The domains present in *thiI* are shown above the gene in pink; b) Pie-chart showing the COG functional categories of the 89 invasive disease candidate loci.

phagocytosis by macrophages and poor at colonisation, whilst other alleles were better able to colonize the murine nasopharynx but did not show any changes in virulence when compared to the wild-type alleles [125].

galU is involved in sugar metabolism but has also been shown to affect capsular biosynthesis as pneumococcal mutants with disruptions to this gene produced little or no capsule and exhibited growth defects [407, 408]. The pneumococcal capsule protects the cell against phagocytosis so variation in *galU* may have an effect on the amount of capsule produced and thus an effect on the success of pneumococcus at invading sites outside the nasopharynx. Upon exposure to human lung epithelial cells for 0.5 hour and 1 hour time periods, *ilvE* was among several genes that were upregulated and this suggested a role in pathogenesis [409]. The presence of the three conserved capsular genes that are found in all *cps* loci apart from serotype 3 in the list of candidate loci was likely due to the association between serotype and invasive disease. In this study, certain lineages such as CC191^{7F} are almost exclusively found in invasive disease and lineage-specific variation in these genes may have contributed to the significant p-values associated with these genes.

The 89 invasive disease candidate loci data identified by seer were re-analysed to look for associations between particular serotypes and the candidate loci (Supplementary File 7.3). For the majority of the candidate loci, the five most commonly associated serotypes per locus were also among the most prevalent of the 33 different serotypes identified among the 183 invasive disease pneumococci (Supplementary File 7.3). Therefore based on this analysis, there did not appear to be a strong relationship between the invasive disease candidate loci and serotype, over and above the well-recognised association between certain serotypes (e.g. 14, 7F, 9V) and invasive disease.

Further work to closely examine allelic variants for each candidate locus might reveal an association of certain alleles with serotype, although this analysis is not trivial as the allelic variants for each of the candidate loci would have to be investigated separately. However, the list of candidate loci associated with invasive disease might be reduced by reanalysing the invasive disease dataset with another GWAS program (see Discussion and Chapter 8). The results of this reanalysis could be compared with the original seer analysis to identify genes that were common to both sets of analyses. This would potentially narrow the list of candidate loci associated with invasive disease, making any downstream analyses such as considering the association between these genes and particular serotypes more straightforward.

Assigning COG functional categories to all 89 invasive disease candidate loci revealed that the largest category was function unknown (21.3%), followed by carbohydrate transport and metabolism (11.2%), defence mechanisms (11.2%), DNA replication, recombination and repair (10.1%), and transcription (9.0%; Figure 7.5b).

7.5 Discussion

This is the first time that a population-level pneumococcal genome dataset composed of both carriage and disease genomes, has been used for a GWAS. The first aim of this chapter was to validate SEER, a new program for performing GWAS on bacterial genome datasets. This was done by searching for genes or candidate loci associated with penicillin non-susceptibility in the VICE dataset and indeed, SEER predicted a significant association between penicillin non-susceptibility and genes previously demonstrated experimentally to confer penicillin non-susceptibility. The results of the SEER analysis were compared against loci previously identified as being associated with penicillin non-susceptibility using a different GWAS methodology. This comparison confirmed the role

of *pbp2B*, *pbp2X*, *mraW* and *ftsL* in penicillin non-susceptibility, but also suggested that there were potentially a large number of false positive associations found using both GWAS methodologies. The second aim of this thesis was to use SEER to search for k-mers significantly associated with three distinct phenotypes. Two of these analyses (k-mers significantly associated with otitis media and pneumonia) revealed no significant k-mers. The third analysis, which attempted to find k-mers significantly associated with invasive disease, found 5,605 significant k-mers that mapped back to 89 candidate loci.

The two candidate loci with the most significant p-values, *thiI* and *nifS*, have not been studied in pneumococcus. However, as both genes are conserved across the Firmicutes, it is possible that they perform the same function in pneumococcus. The reasons for the significant association of these two genes with invasive disease are unclear but one may speculate. It has been shown that these two genes are involved in the bacterial stress response to UV exposure, which is generally hazardous to bacteria, and thiamine biosynthesis [404]. The pneumococcus has adapted to survive in its primary ecological niche, the human nasopharynx, by developing mechanisms to obtain essential nutrients such as carbohydrates and by developing mechanisms for evading the host immune response (the polysaccharide capsule). It could be hypothesised that, upon entering a site outside the nasopharynx such as blood or CSF where the normally available nutrients are no longer available, a stress response is induced and alternative pathways for obtaining nutrients essential for survival are activated.

There are two possible explanations for the large number of candidate loci (n = 89) identified in the invasive disease analysis: 1) there was insufficient control of population structure leading to an excess of "hitchhiker" loci *i.e.* false positives; and 2) there isn't a single "disease" gene in pneumococcus but virulence is due to a number of different loci

that form a virulence profile, and in combination with the state of the host's immune system at the time of infection may be the most important factor in IPD.

SEER has been designed to specifically adjust for population structure in order to reduce the number of significant associations due to genetic relatedness rather than causality. However, large numbers of likely false positives were identified by SEER in both the penicillin non-susceptibility and invasive disease analyses. These results suggested that there was still insufficient control for population structure despite using a high number of principal components (n = 20) which should have provided a greater level of control for population structure in both analyses. A more stringent approach, such as increasing the number of principal components, may reduce the number of false positives but the most suitable approach would be to reanalyse the invasive disease dataset using an alternative GWAS methodology such as bugwas (see Chapter 8) [410]. The results could then be compared, as was done for the results of the penicillin susceptibility analysis, and any candidate loci that were common to both analyses could be considered have a much higher likelihood of being associated with invasive disease. This would potentially reduce the number of candidates for any functional work to confirm the role of these candidate loci in invasive disease. One example of this functional work would be to knock out these candidate loci in a systematic manner and use a murine model of pathogenesis to test whether or not knocking out any of the candidate loci prevents IPD.

The second possible explanation was that certain pneumococci have a virulence profile, that under the right conditions in the host *e.g.* reduced immune efficiency, leads to disease. Based on this hypothesis, host genetics and health have a much larger role to play than simply the presence or absence of virulence genes in the pneumococcus *i.e.* a pneumococcus with a particular virulence profile may cause disease in one child but be cleared from carriage with no adverse effects in another child. From the point of view of

host genetics, it is likely that polymorphisms in host defence function affect the susceptibility to disease caused by different pneumococcal lineages [411] but this is very difficult to prove without at least some knowledge about the host's genome. In the future, studies which aim to sequence both the pneumococci implicated in IPD and the host to allow for true host-pathogen association studies may be able to begin to reveal the true role of host susceptibility in IPD.

A closer examination of the available metadata for the hosts associated with the invasive disease pneumococci included in the analysis in this chapter showed, that at least for Iceland, there was a wide distribution of ages amongst the hosts with only 12 children under the age of 10 included along with 62 adults over the age of 65 in the 183 cases of invasive disease (Supplementary File 7.1). This suggests that infection was largely opportunistic and there was not a hyper virulent lineage responsible for causing invasive disease in Iceland. The available metadata for the VICE study is limited but it is possible that by examining patient records, risk factors for pneumococcal disease such as co-infection with influenza or else some kind of immunocompromised state could be identified for the patients with IPD in the study. This may help the analyses by using this information as covariates in the GWAS which might reduce the number of possible false positives.

The introduction of pneumococcal conjugate vaccines has led to a significant decrease in cases of IPD [154, 155] but IPD has not completely been eradicated due to the rise in IPD in the post-PCV era caused by non-vaccine serotypes. This makes the development of alternative approaches, like protein-based vaccines, that target all pneumococci regardless of serotype increasingly important. One application of the results of the invasive disease analysis would be to use the list of candidate loci as a starting point for identifying potential candidates for inclusion in protein-based pneumococcal vaccines (at least one of

the candidate loci, *pspA*, has already been used as the basis of a protein based vaccine [412]). To date, the identification of potential candidates for inclusion in any protein-based vaccine has been based on identifying predicted surface proteins based on the presence of signal peptides, choline-binding domains and sortase motifs [413, 414], signature-tagged mutagenesis [415], anti-genomics [415] or *in vivo* transcriptional analysis [416]. These studies only considered small numbers of pneumococci so the results of the SEER analysis, which was performed on nearly 2,000 genomes, may provide an improved level of resolution when considering targets for novel protein-based vaccines.

Performing a GWAS on bacterial genome datasets offers a powerful means of identifying genes that may be associated with invasive disease and thus a suitable methodology for identifying targets for further investigation in the laboratory. However, care needs to be taken when interpreting the results as the analyses performed in this chapter showed that the false positive rate is likely high. This is not a new problem and is not unique to the application of GWAS to bacterial datasets; the results from many of the early association studies performed on human genome data were proven to be incorrect and in many cases, the results could not be replicated. Any genes identified as being associated with invasive disease using a GWAS approach as was done in this study will need to be confirmed with functional work or analysis using a different methodology. As other suitable datasets similar in composition to the VICE study dataset become available, we could increase confidence in our ability to successfully detect genes associated with invasive disease by comparing any lists of candidate loci from other studies to the list of candidate loci identified in this study. Whilst functional work and analysis using different methodologies may help provide increased confidence in the results, it is likely that the

key to truly understanding IPD will be to include an analysis of host genetics along with the analysis of the bacterial genomes.

7.6 Supplementary information (see attached USB drive)

Supplementary File 7.1 VICE study dataset

Supplementary File 7.2 Penicillin non-susceptibility candidate loci

Supplementary File 7.3 Invasive disease candidate loci

8. Summary and future work

8.1 Summary

The pneumococcus is an important global pathogen responsible for a significant disease burden with considerable associated health and economic costs. The introduction of effective pneumococcal conjugate vaccines has led to a dramatic reduction in the number of deaths due to invasive pneumococcal disease, as well as a significant reduction in the number of cases of non-invasive pneumococcal disease. Despite this success, the rates of carriage of pneumococcus have not changed due to an increase in the prevalence of non-vaccine serotypes in the pneumococcal population, showing that the pneumococcus is capable of adapting to vaccine-driven evolutionary pressure. This has long term implications for the continued efficacy of vaccination programmes around the world making the continued monitoring of global pneumococcal populations after the introduction of effective vaccines important.

The increasing number of pneumococcal genomes from large well-characterised genomes means that more detailed studies of the processes driving pneumococcal genomic evolution can now be conducted. The data presented in this thesis showed how WGS can be used to better our understanding of key aspects of pneumococcal biology and the findings can be summarised into three principle areas of interest: the pneumococcal pan-genome, *cps* locus diversity and the application of a GWAS approach for identifying genes associated with invasive pneumococcal disease.

The pneumococcal pan-genome

The availability of large genome datasets comprised mostly of draft genomes meant that a core genome method that attempted to take into account the fragmented nature of draft

genomes, where regions of the genome and thus genes may be missing due to issues with sequencing, was required. In Chapter 3, I described a Bayesian decision model for estimating bacterial core genomes for datasets principally comprised of draft genomes, showed that it was scalable to any number of genomes and used the model to identify several areas of interest amongst the five bacterial datasets that were worthy of further study. I then applied it to four different pneumococcal carriage datasets from around the world (UK, USA, Iceland and Thailand). A comparison of the estimated core genomes for the four datasets revealed a shared core genome of only 303 genes and a similar comparison of the pan-genomes of the four datasets highlighted differences amongst the datasets and in particular the Thailand dataset which was significantly different from the other three datasets. An examination of the genes unique to the Thailand dataset revealed 14 large genomic regions, which included transposons and bacteriophages. Attempting to identify the source of these large regions revealed that many were likely to have been imported from other streptococci.

***cps* locus diversity**

A detailed analysis of 974 serogroup 6 pneumococci in Chapter 4 revealed that nearly half of the genomes considered were serotype 6E (a variant of serotype 6B), identified a novel serotype 6C/6E hybrid from Thailand and showed that vaccine-induced serotype 6B antibodies were able to kill serotype 6E pneumococci. Further analyses of 5,405 genomes from 14 serotypes/serogroups in Chapter 5 identified several putative novel and hybrid serotypes, based on the divergence of the nucleotide sequences of these *cps* loci from the published reference sequences. There were also considerable differences in the diversity of *cps* loci amongst the various serotypes/serogroups considered: serogroups such as 9 and 11 had low *cps* locus diversity and serogroups like 6 and 23 had high *cps* locus diversity.

Using a GWAS approach to identify genes associated with invasive disease

The application of the GWAS approach to bacterial datasets is a recent innovation and tools specifically designed for these analyses have begun to be made available. One of these new tools, SEER, was validated by identifying genes in the VICE study dataset previously known to be associated with penicillin non-susceptibility in pneumococcus. Applying SEER to the VICE study dataset to look for genes significantly associated with IPD led to the identification of 89 candidate loci (see Chapter 7).

8.2 Future work

The pneumococcal pan-genome

There is a desire amongst some researchers in the pneumococcal community to develop a core genome scheme to improve on the pneumococcal MLST and rMLST schemes as has been done for other bacterial species such as *Neisseria meningitidis* and *Campylobacter jejuni* [417]. Whilst the pneumococcal MLST scheme has proved itself to be robust even in the age of genomics, defining a clonal complex has become increasingly subjective as the networks of related sequence types have increased in both complexity and connectedness. The work in Chapter 6 showed that, whilst the core genome varies considerably from region to region, there is still a set of 303 shared genes amongst pneumococci from different regions of the world that could potentially be used as the basis for a pneumococcal core genome scheme. The obvious first step would be to take the set of 303 shared core genes identified in this study and assess the relative frequency of these genes amongst genomes from other parts of the world not included in this study. In this way the number of genes that are truly common to all pneumococci, and thus worthy of inclusion in a core genome scheme, could be further refined. This pneumococcal core genome scheme could then be integrated into a website such as

pubMLST. This would allow the pneumococcal core genome scheme to be applied to the ~8,500 genomes already in the pubMLST database and would enable researchers from around the world to type their newly sequenced pneumococcal genomes using the scheme. Pneumococcal lineages could be assigned based on the scheme and downstream analyses, such as the swift construction of informative phylogenetic trees based on the nucleotide sequences extracted from each genome for each of the genes included in the core genome scheme, could be undertaken.

Following the analyses in Chapter 6, it is clear that there is considerable plasticity in the accessory genomes both between and within pneumococcal populations from different geographical locations. In particular, the dataset from the Maela refugee camp had a large accessory genome and a very small core genome. Explanations for these differences were sought by examining the genes and, in particular, the large genomic regions (>25 Kb) unique to the Thailand dataset. This was merely a first step as the analysis only considered a relatively small proportion of the ~3,700 genes unique to the Thailand dataset. The analysis could be extended to consider the other genes not present in large genomic regions. In a similar fashion to the analysis conducted in Chapter 6, the likely origin of these genes could be sought and this may begin to provide further explanations for the unexpected diversity in the Thailand dataset. Alongside this analysis, the genes unique to the other three datasets in the analysis in Chapter 6, which varied in number from 587 (UK) to 754 (Iceland), could also be explored. For example, data not shown in this thesis from an earlier analysis, similar to that conducted in Chapter 6, identified an ICE from *Streptococcus intermedius* in a single genome from the Icelandic dataset. On this basis, and from published work that has looked at MGEs in the USA dataset [418], exploring the regions unique to each of the four datasets included in Chapter 6 would likely reveal novel bacteriophages and other MGEs. This would help

expand our understanding of these diverse, and to date, poorly understood regions of the pneumococcal genome.

cps locus diversity

An important finding of this thesis was the identification of a number of hybrid and putative novel serotypes in Chapter 5. The question of whether or not these capsular variants encode novel polysaccharide capsules or induce a different immune response is one that needs addressing as it may have important implications for vaccine efficacy as well as our understanding of the overall diversity of pneumococci globally.

The first step in answering this question would be to elucidate the biochemical structure of the polysaccharide capsules from pneumococci with hybrid and putatively novel serotypes. Determining polysaccharide structures has traditionally been technically challenging and many of the early polysaccharide structures were incomplete or contained errors [120]. There are also difficulties related to the purity of the capsular polysaccharide: pure capsular polysaccharide is required for structural analyses but preparations of capsular polysaccharide are often contaminated with cell wall polysaccharide which may skew any analyses [120]. In recent years, with the development of analytical technologies such as nuclear magnetic resonance (NMR), mass spectrometry (MS) and gas-liquid chromatography, it has become much easier to deduce polysaccharide structures. This means that it should now be possible to work out the polysaccharide capsule structures for each of the putative novel serotypes described in Chapter 5.

The second step would be to assess the efficacy of vaccine-induced antibodies against the novel and hybrid serotypes, at least for the examples of hybrid and novel serotypes from serogroups included in any of the pneumococcal conjugate vaccines. This sort of

analysis, in the form of opsonophagocytosis assays, was performed on serotype 6E pneumococci in Chapter 4 and revealed that vaccine-induced antibodies to serotype 6B pneumococci were able to kill serotype 6E pneumococci. The finding that there were no differences in the antigenic responses generated by these two capsular loci was unexpected given the >5% nucleotide sequence divergence between the two, but was explained by the subsequent publication of the polysaccharide structure of serotype 6E [129], which revealed that there were no differences in the structure of the polysaccharide structures between serotypes 6B and 6E. It's likely that similar results will be observed for at least some of the putative novel and hybrid serotypes identified in Chapters 4 and 5 but, given the presence of many of these capsular variants in the Thailand dataset, a proportion of them will be revealed to be truly novel and can be added to the already extensive list of known serotypes. This will improve the routine epidemiological monitoring of pneumococcus that is being undertaken around the world and will help in the design of future vaccines and treatments.

Using a GWAS approach to identify genes associated with invasive disease

The GWAS approach used in Chapter 7 identified 89 genes significantly associated with invasive disease in pneumococcus. The role of these candidate loci in IPD will have to be confirmed by further functional work such as gene knockouts and animal models of disease. Both of these are expensive methodologies and, given the number of candidate loci to be considered, before undertaking laboratory based analyses, the results from the invasive disease analysis in Chapter 7 could first be confirmed using alternative computational methods. These include using a different GWAS program, a non-GWAS computational methodology such as a random forest analysis or a combination of the two approaches. The results of this second computational analysis could then be used to

refine the list of genes that would require experimental validation with respect to their association with invasive disease.

Another GWAS program specifically designed for analysing genomes, bugwas [410], became available during the course of the work described in Chapter 7. The analyses could be repeated using bugwas and the results from both programs could then be compared. The main difference between SEER and bugwas is the choice of model when controlling for population structure amongst the genomes included in the analysis: population structure is controlled for by SEER using metric multi-dimensional scaling (MDS) whilst bugwas uses a linear mixed model (LMM) to control for bacterial population structure.

A random forest is a machine learning technique that is beginning to be used in the analysis of large biological datasets. The random forest algorithm is an ensemble method which combines the information of multiple binary classification trees built around a group of predictor variables (in the context of genome analysis, these predictors could be allelic variants of genes) towards a response variable, such as a particular phenotype. The output of a random forest analysis on a dataset of genes would be composed of both the classification success rates of the response variable (phenotype) and a ranking of the predictor variables (genes) quantifying their relative role in the classification process. For the example of identifying genes associated with penicillin susceptibility or invasive disease, genomes would be mapped back to a reference and an allele number would be assigned to each example of a gene based on how the nucleotide sequence of that gene varied from the reference sequence. These allele numbers would then be used as the predictor variables and the response variable would be penicillin susceptibility or invasive disease. In this way, each gene would be assigned a score based on whether or not they were predictive of the phenotype.

Using a variety of approaches such as different GWAS methodologies and a random forest approach will improve our confidence in the list of genes associated with invasive disease that were identified in Chapter 7. This will better our understanding of the mechanisms behind invasive disease in pneumococcus and this list of genes will also have potential applications such as identifying potential candidates for inclusion in protein-based pneumococcal vaccines (see Chapter 7).

It is clear that the large scale application of genomics has revolutionised bacterial research and this thesis has made extensive use of the large numbers of pneumococcal genomes that have become available in the past few years to investigate important aspects of pneumococcal biology. As well as revealing important findings about the pneumococcal pan-genome, *cps* locus diversity and genes associated with IPD, this thesis has also shown that our understanding of the true diversity of pneumococci collected around the globe is still a work in progress. Further work will be required, using genome sequences extracted from pneumococci collected in thus-far understudied regions of the world, before the biological and evolutionary processes that drive pneumococcal evolution are fully understood.

9. References

1. Sternberg GM: **A Fatal Form of Septicaemia in the Rabbit Produced by the Subcutaneous Injection of Human Saliva: An Experimental Research**: John Murphy & Company; 1881.
2. Pasteur L: **Note sur la maladie nouvelle provoquee par la salive d'un enfant mort de la rage**. *Bull Acad Med* 1881, **10**:94-103.
3. Facklam R: **What happened to the streptococci: overview of taxonomic and nomenclature changes**. *Clin Microbiol Rev* 2002, **15**(4):613-630.
4. Chandler LJ, Reisner BS, Woods GL, Jafri AK: **Comparison of four methods for identifying *Streptococcus pneumoniae***. *Diagn Microbiol Infect Dis* 2000, **37**(4):285-287.
5. Wannamaker LW, Matsen JM: **Streptococci and streptococcal diseases: recognition, understanding, and management**; 1972.
6. Buchanan R, Gibbons N: **Bergey's manual of determinative bacteriology**. *Williams and Wilkins, Baltimore* 1974:1-1246.
7. Kawamura Y, Hou XG, Sultana F, Miura H, Ezaki T: **Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus***. *Int J Syst Bacteriol* 1995, **45**(2):406-408.
8. Chi F, Nolte O, Bergmann C, Ip M, Hakenbeck R: **Crossing the barrier: Evolution and spread of a major class of mosaic *pbp2x* in *Streptococcus pneumoniae, S. mitis* and *S. oralis***. *International Journal of Medical Microbiology* 2007, **297**(7-8):503-512.
9. Kilian M, Poulsen K, Blomqvist T, Havarstein LS, Bek-Thomsen M, Tettelin H, Sorensen UB: **Evolution of *Streptococcus pneumoniae* and its close commensal relatives**. *PLoS One* 2008, **3**(7):e2683.
10. Tuomanen EI, Mitchell TJ, Morrison DA, Spratt BG: **The pneumococcus**: ASM press; 2004.
11. Schleifer KH, Kilpper-Bälz R: **Transfer of *Streptococcus faecalis* and *Streptococcus faecium* to the Genus *Enterococcus* nom. rev. as *Enterococcus faecalis* comb. nov. and *Enterococcus faecium* comb. nov**. *International Journal of Systematic and Evolutionary Microbiology* 1984, **34**(1):31-34.

12. Neufeld F: **Ueber die Agglutination der Pneumokokken und über die Theorieen der Agglutination**. *Medical Microbiology and Immunology* 1902, **40**(1):54-72.
13. Lafong AC, Crothers E: **Simple latex agglutination method for typing pneumococci**. *J Clin Pathol* 1988, **41**(2):230-231.
14. Slotved HC, Kaltoft M, Skovsted IC, Kerrn MB, Espersen F: **Simple, rapid latex agglutination test for serotyping of pneumococci (Pneumotest-Latex)**. *J Clin Microbiol* 2004, **42**(6):2518-2522.
15. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA *et al*: **Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes**. *PLoS Genet* 2006, **2**(3):e31.
16. Pai R, Gertz RE, Beall B: **Sequential multiplex PCR approach for determining capsular serotypes of *Streptococcus pneumoniae* isolates**. *J Clin Microbiol* 2006, **44**(1):124-131.
17. Figueiredo AM, Austrian R, Urbaskova P, Teixeira LA, Tomasz A: **Novel penicillin-resistant clones of *Streptococcus pneumoniae* in the Czech Republic and in Slovakia**. *Microb Drug Resist* 1995, **1**(1):71-78.
18. McDougal LK, Facklam R, Reeves M, Hunter S, Swenson JM, Hill BC, Tenover FC: **Analysis of multiply antimicrobial-resistant isolates of *Streptococcus pneumoniae* from the United States**. *Antimicrob Agents Chemother* 1992, **36**(10):2176-2184.
19. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA *et al*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms**. *Proc Natl Acad Sci U S A* 1998, **95**(6):3140-3145.
20. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level**. *BMC Bioinformatics* 2010, **11**:595.
21. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG: **eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data**. *J Bacteriol* 2004, **186**(5):1518-1530.

22. Francisco AP, Bugalho M, Ramirez M, Carrico JA: **Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach**. *BMC Bioinformatics* 2009, **10**:152.
23. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG: **Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential**. *J Infect Dis* 2003, **187**(9):1424-1432.
24. Granat SM, Mia Z, Ollgren J, Herva E, Das M, Piirainen L, Auranen K, Makela PH: **Longitudinal study on pneumococcal carriage during the first year of life in Bangladesh**. *Pediatr Infect Dis J* 2007, **26**(4):319-324.
25. Gratten M, Montgomery J, Gerega G, Gratten H, Siwi H, Poli A, Koki G: **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* 1989, **20**(4):501-509.
26. Hill PC, Akisanya A, Sankareh K, Cheung YB, Saaka M, Lahai G, Greenwood BM, Adegbola RA: **Nasopharyngeal carriage of *Streptococcus pneumoniae* in Gambian villagers**. *Clin Infect Dis* 2006, **43**(6):673-679.
27. Leach AJ, Boswell JB, Asche V, Nienhuys TG, Mathews JD: **Bacterial colonization of the nasopharynx predicts very early onset and persistence of otitis media in Australian aboriginal infants**. *Pediatr Infect Dis J* 1994, **13**(11):983-989.
28. Siber GR, Klugman KP, Mäkelä PH: **Pneumococcal vaccines: the impact of conjugate vaccine**. *Pneumococcal vaccines: the impact of conjugate vaccine* 2008.
29. Aniansson G, Alm B, Andersson B, Larsson P, Nylen O, Peterson H, Rigner P, Svanborg M, Svanborg C: **Nasopharyngeal colonization during the first year of life**. *J Infect Dis* 1992, **165 Suppl 1**:S38-42.
30. Bogaert D, Engelen MN, Timmers-Reker AJ, Elzenaar KP, Peerbooms PG, Coutinho RA, de Groot R, Hermans PW: **Pneumococcal carriage in children in The Netherlands: a molecular epidemiological study**. *J Clin Microbiol* 2001, **39**(9):3316-3320.
31. Gray BM, Converse GM, 3rd, Dillon HC, Jr.: **Epidemiologic studies of *Streptococcus pneumoniae* in infants: acquisition, carriage, and infection during the first 24 months of life**. *J Infect Dis* 1980, **142**(6):923-933.

32. Hendley JO, Sande MA, Stewart PM, Gwaltney JM, Jr.: **Spread of *Streptococcus pneumoniae* in families. I. Carriage rates and distribution of types**. *J Infect Dis* 1975, **132**(1):55-61.

33. Hammitt LL, Bruden DL, Butler JC, Baggett HC, Hurlburt DA, Reasonover A, Hennessy TW: **Indirect effect of conjugate vaccine on adult carriage of *Streptococcus pneumoniae*: an explanation of trends in invasive pneumococcal disease**. *J Infect Dis* 2006, **193**(11):1487-1494.

34. Leino T, Auranen K, Jokinen J, Leinonen M, Tervonen P, Takala AK: **Pneumococcal carriage in children during their first two years: important role of family exposure**. *Pediatr Infect Dis J* 2001, **20**(11):1022-1027.

35. Givon-Lavi N, Fraser D, Porat N, Dagan R: **Spread of *Streptococcus pneumoniae* and antibiotic-resistant *S. pneumoniae* from day-care center attendees to their younger siblings**. *J Infect Dis* 2002, **186**(11):1608-1614.

36. Bogaert D, van Belkum A, Sluijter M, Luijendijk A, de Groot R, Rumke HC, Verbrugh HA, Hermans PW: **Colonisation by *Streptococcus pneumoniae* and *Staphylococcus aureus* in healthy children**. *Lancet* 2004, **363**(9424):1871-1872.

37. Millar EV, O'Brien KL, Watt JP, Bronsdon MA, Dallas J, Whitney CG, Reid R, Santosham M: **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* 2006, **43**(1):8-15.

38. Watson K, Carville K, Bowman J, Jacoby P, Riley TV, Leach AJ, Lehmann D: **Upper respiratory tract bacterial carriage in Aboriginal and non-Aboriginal children in a semi-arid area of Western Australia**. *Pediatr Infect Dis J* 2006, **25**(9):782-790.

39. Huang SS, Finkelstein JA, Rifas-Shiman SL, Kleinman K, Platt R: **Community-level predictors of pneumococcal carriage and resistance in young children**. *Am J Epidemiol* 2004, **159**(7):645-654.

40. Smith CB, Golden C, Klauber MR, Kanner R, Renzetti A: **Interactions between viruses and bacteria in patients with chronic bronchitis**. *J Infect Dis* 1976, **134**(6):552-561.

41. Wadowsky RM, Mietzner SM, Skoner DP, Doyle WJ, Fireman P: **Effect of experimental influenza A virus infection on isolation of *Streptococcus pneumoniae* and other aerobic bacteria from the oropharynges of allergic and nonallergic adult subjects**. *Infect Immun* 1995, **63**(4):1153-1157.

42. Tong HH, Fisher LM, Kosunick GM, DeMaria TF: **Effect of adenovirus type 1 and influenza A virus on *Streptococcus pneumoniae* nasopharyngeal colonization and otitis media in the chinchilla**. *Ann Otol Rhinol Laryngol* 2000, **109**(11):1021-1027.
43. Peltola VT, Mccullers JA: **Respiratory viruses predisposing to bacterial infections: role of neuraminidase**. *The Pediatric infectious disease journal* 2004, **23**(1):S87-S97.
44. Avadhanula V, Rodriguez CA, Devincenzo JP, Wang Y, Webby RJ, Ulett GC, Adderson EE: **Respiratory viruses augment the adhesion of bacterial pathogens to respiratory epithelium in a viral species- and cell type-dependent manner**. *J Virol* 2006, **80**(4):1629-1636.
45. Greenberg D, Givon-Lavi N, Broides A, Blancovich I, Peled N, Dagan R: **The contribution of smoking and exposure to tobacco smoke to *Streptococcus pneumoniae* and *Haemophilus influenzae* carriage in children and their mothers**. *Clin Infect Dis* 2006, **42**(7):897-903.
46. Cardozo DM, Nascimento-Carvalho CM, Andrade AL, Silvany-Neto AM, Daltro CH, Brandao MA, Brandao AP, Brandileone MC: **Prevalence and risk factors for nasopharyngeal carriage of *Streptococcus pneumoniae* among adolescents**. *J Med Microbiol* 2008, **57**(Pt 2):185-189.
47. Smith T, Lehmann D, Montgomery J, Gratten M, Riley ID, Alpers MP: **Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children**. *Epidemiol Infect* 1993, **111**(1):27-39.
48. Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, Moxon ER, Crook DW, Peto TE: **Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children**. *J Infect Dis* 2006, **194**(5):682-688.
49. Hansman D, Morris S: **Pneumococcal carriage amongst children in Adelaide, South Australia**. *Epidemiol Infect* 1988, **101**(2):411-417.
50. O'Brien KL, Millar EV, Zell ER, Bronsdon M, Weatherholtz R, Reid R, Becenti J, Kvamme S, Whitney CG, Santosham M: **Effect of pneumococcal conjugate vaccine on nasopharyngeal colonization among immunized and unimmunized children in a community-randomized trial**. *J Infect Dis* 2007, **196**(8):1211-1220.

51. Brugger SD, Frey P, Aebi S, Hinds J, Muhlemann K: **Multiple colonization with *S. pneumoniae* before and after introduction of the seven-valent conjugated pneumococcal polysaccharide vaccine**. *PLoS One* 2010, **5**(7):e11638.
52. Huebner RE, Dagan R, Porath N, Wasas AD, Klugman KP: **Lack of utility of serotyping multiple colonies for detection of simultaneous nasopharyngeal carriage of different pneumococcal serotypes**. *Pediatr Infect Dis J* 2000, **19**(10):1017-1020.
53. Regev-Yochay G, Dagan R, Raz M, Carmeli Y, Shainberg B, Derazne E, Rahav G, Rubinstein E: **Association between carriage of *Streptococcus pneumoniae* and *Staphylococcus aureus* in Children**. *JAMA* 2004, **292**(6):716-720.
54. Bosch AA, van Houten MA, Bruin JP, Wijmenga-Monsuur AJ, Trzcinski K, Bogaert D, Rots NY, Sanders EA: **Nasopharyngeal carriage of *Streptococcus pneumoniae* and other bacteria in the 7th year after implementation of the pneumococcal conjugate vaccine in the Netherlands**. *Vaccine* 2016, **34**(4):531-539.
55. Grivea IN, Priftis KN, Giotas A, Kotzia D, Tsantouli AG, Douros K, Michoula AN, Syrogiannopoulos GA: **Dynamics of pneumococcal carriage among day-care center attendees during the transition from the 7-valent to the higher-valent pneumococcal conjugate vaccines in Greece**. *Vaccine* 2014, **32**(48):6513-6520.
56. Cohen R, Biscardi S, Levy C: **The multifaceted impact of pneumococcal conjugate vaccine implementation in children in France between 2001 to 2014**. *Hum Vaccin Immunother* 2016, **12**(2):277-284.
57. Hanieh S, Hamaluba M, Kelly DF, Metz JA, Wyres KL, Fisher R, Pradhan R, Shakya D, Shrestha L, Shrestha A *et al*: ***Streptococcus pneumoniae* carriage prevalence in Nepal: evaluation of a method for delayed transport of samples from remote regions and implications for vaccine implementation**. *PLoS One* 2014, **9**(6):e98739.
58. Roca A, Hill PC, Townend J, Egere U, Antonio M, Bojang A, Akisanya A, Litchfield T, Nsekpong DE, Oluwalana C *et al*: **Effects of community-wide vaccination with PCV-7 on pneumococcal nasopharyngeal carriage in the Gambia: a cluster-randomized trial**. *PLoS Med* 2011, **8**(10):e1001107.
59. Bogaert D, De Groot R, Hermans PW: ***Streptococcus pneumoniae* colonisation: the key to pneumococcal disease**. *Lancet Infect Dis* 2004, **4**(3):144-154.

60. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T *et al*: **Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates**. *Lancet* 2009, **374**(9693):893-902.

61. Walker CL, Rudan I, Liu L, Nair H, Theodoratou E, Bhutta ZA, O'Brien KL, Campbell H, Black RE: **Global burden of childhood pneumonia and diarrhoea**. *Lancet* 2013, **381**(9875):1405-1416.

62. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY *et al*: **Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010**. *The Lancet* 2012, **380**(9859):2095-2128.

63. Leibovitz E, Piglansky L, Raiz S, Greenberg D, Hamed KA, Ledeine JM, Press J, Leiberman A, Echols RM, Pierce PF *et al*: **Bacteriologic and clinical efficacy of oral gatifloxacin for the treatment of recurrent/nonresponsive acute otitis media: an open label, noncomparative, double tympanocentesis study**. *Pediatr Infect Dis J* 2003, **22**(11):943-949.

64. Takala AK, Jero J, Kela E, Ronnberg PR, Koskenniemi E, Eskola J: **Risk factors for primary invasive pneumococcal disease among children in Finland**. *JAMA* 1995, **273**(11):859-864.

65. Pastor P, Medley F, Murphy TV: **Invasive pneumococcal disease in Dallas County, Texas: results from population-based surveillance in 1995**. *Clin Infect Dis* 1998, **26**(3):590-595.

66. Lipsky BA, Boyko EJ, Inui TS, Koepsell TD: **Risk factors for acquiring pneumococcal infections**. *Arch Intern Med* 1986, **146**(11):2179-2185.

67. Robinson KA, Baughman W, Rothrock G, Barrett NL, Pass M, Lexau C, Damaske B, Stefonek K, Barnes B, Patterson J *et al*: **Epidemiology of invasive *Streptococcus pneumoniae* infections in the United States, 1995-1998: Opportunities for prevention in the conjugate vaccine era**. *JAMA* 2001, **285**(13):1729-1735.

68. Frankel RE, Virata M, Hardalo C, Altice FL, Friedland G: **Invasive pneumococcal disease: clinical features, serotypes, and antimicrobial resistance patterns in cases involving patients with and without human immunodeficiency virus infection**. *Clin Infect Dis* 1996, **23**(3):577-584.

69. Nuorti JP, Butler JC, Gelling L, Kool JL, Reingold AL, Vugia DJ: **Epidemiologic relation between HIV and invasive pneumococcal disease in San Francisco County, California**. *Ann Intern Med* 2000, **132**(3):182-190.

70. Kyaw MH, Rose CE, Jr., Fry AM, Singleton JA, Moore Z, Zell ER, Whitney CG: **The influence of chronic illnesses on the incidence of invasive pneumococcal disease in adults**. *J Infect Dis* 2005, **192**(3):377-386.

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

72. Nuorti JP, Butler JC, Farley MM, Harrison LH, McGeer A, Kolczak MS, Breiman RF: **Cigarette smoking and invasive pneumococcal disease. Active Bacterial Core Surveillance Team**. *N Engl J Med* 2000, **342**(10):681-689.

73. Johnson HL, Deloria-Knoll M, Levine OS, Stoszek SK, Freimanis Hance L, Reithinger R, Muenz LR, O'Brien KL: **Systematic evaluation of serotypes causing invasive pneumococcal disease among children under five: the pneumococcal global serotype project**. *PLoS Med* 2010, **7**(10).

74. Grabenstein JD, Musey LK: **Differences in serious clinical outcomes of infection caused by specific pneumococcal serotypes among adults**. *Vaccine* 2014, **32**(21):2399-2405.

75. Brueggemann AB, Peto TE, Crook DW, Butler JC, Kristinsson KG, Spratt BG: **Temporal and geographic stability of the serogroup-specific invasive disease potential of *Streptococcus pneumoniae* in children**. *J Infect Dis* 2004, **190**(7):1203-1211.

76. Hoge CW, Reichler MR, Dominguez EA, Bremer JC, Mastro TD, Hendricks KA, Musher DM, Elliott JA, Facklam RR, Breiman RF: **An epidemic of pneumococcal disease in an overcrowded, inadequately ventilated jail**. *N Engl J Med* 1994, **331**(10):643-648.

77. Nuorti JP, Butler JC, Crutcher JM, Guevara R, Welch D, Holder P, Elliott JA: **An outbreak of multidrug-resistant pneumococcal pneumonia and bacteremia among unvaccinated nursing home residents**. *N Engl J Med* 1998, **338**(26):1861-1868.

78. Berk SL, Gage KA, Holtsclaw-Berk SA, Smith JK: **Type 8 pneumococcal pneumonia: an outbreak on an oncology ward**. *South Med J* 1985, **78**(2):159-161.

79. Leimkugel J, Adams Forgor A, Gagneux S, Pfluger V, Flierl C, Awine E, Naegeli M, Dangy JP, Smith T, Hodgson A *et al*: **An outbreak of serotype 1 *Streptococcus pneumoniae* meningitis in northern Ghana with features that are characteristic of *Neisseria meningitidis* meningitis epidemics**. *J Infect Dis* 2005, **192**(2):192-199.

80. Jennifer Cornick CC, Simon R. Harris, Feyruz Yalsin, Madikay Senghore, Anmol M. Kiran, Shanil Govindpershad, Sani Ousmane, Mignon du Plessis, Gerd Pluschke, Chinelo Ebruke, Lesley McGee, Beutel Sigauque, Jean-Marc Collard, Martin Antoni, Anne von Gottberg, Neil French, Keith P. Klugman, Robert S. Heyderman, Stephen D. Bentley, Dean B. Everett, PAGe Consortium: **Region-specific diversification of the highly virulent serotype 1 *Streptococcus pneumoniae***. *Microbial Genomics* 2015, **1**(1).

81. Martin M, Turco JH, Zegans ME, Facklam RR, Sodha S, Elliott JA, Pryor JH, Beall B, Erdman DD, Baumgartner YY *et al*: **An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae***. *N Engl J Med* 2003, **348**(12):1112-1121.

82. Malke H: **Gross, Trevor, Jane Faull, Steve Ketteridge, and Derek Springham, Introductory Microbiology, XIV, 414 pages, illustrated, Chapman & Hall, London 1995, soft cover,£ 19.99**. In.: Urban & Fischer; 1995.

83. Andam CP, Hanage WP: **Mechanisms of genome evolution of *Streptococcus***. *Infect Genet Evol* 2015, **33**:334-342.

84. Griffith F: **The significance of pneumococcal types**. *Journal of Hygiene* 1928, **27**(02):113-159.

85. Avery OT, MacLeod CM, McCarty M: **Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III**. *The Journal of experimental medicine* 1944, **79**(2):137-158.

86. Tomasz A: **Model for the mechanism controlling the expression of competent state in Pneumococcus cultures**. *Journal of bacteriology* 1966, **91**(3):1050-1061.

87. Havarstein LS, Coomaraswamy G, Morrison DA: **An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae***. *Proc Natl Acad Sci U S A* 1995, **92**(24):11140-11144.

88. Claverys JP, Havarstein LS: **Extracellular-peptide control of competence for genetic transformation in *Streptococcus pneumoniae***. *Front Biosci* 2002, **7**:d1798-1814.

89. Peterson SN, Sung CK, Cline R, Desai BV, Snesrud EC, Luo P, Walling J, Li H, Mintz M, Tsegaye G *et al*: **Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays**. *Mol Microbiol* 2004, **51**(4):1051-1070.

90. Campbell EA, Choi SY, Masure HR: **A competence regulon in *Streptococcus pneumoniae* revealed by genomic analysis**. *Mol Microbiol* 1998, **27**(5):929-939.

91. Pestova EV, Morrison DA: **Isolation and characterization of three *Streptococcus pneumoniae* transformation-specific loci by use of a lacZ reporter insertion vector**. *J Bacteriol* 1998, **180**(10):2701-2710.

92. Chen I, Dubnau D: **DNA uptake during bacterial transformation**. *Nat Rev Microbiol* 2004, **2**(3):241-249.

93. Mortier-Barriere I, Velten M, Dupaigne P, Mirouze N, Pietrement O, McGovern S, Fichant G, Martin B, Noirot P, Le Cam E *et al*: **A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA**. *Cell* 2007, **130**(5):824-836.

94. Chen Z, Yang H, Pavletich NP: **Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures**. *Nature* 2008, **453**(7194):489-484.

95. Berge M, Mortier-Barriere I, Martin B, Claverys JP: **Transformation of *Streptococcus pneumoniae* relies on DprA- and RecA-dependent protection of incoming DNA single strands**. *Mol Microbiol* 2003, **50**(2):527-536.

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

97. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS *et al*: **Rapid pneumococcal evolution in response to clinical interventions**. *Science* 2011, **331**(6016):430-434.

98. Evans BA, Rozen DE: **Significant variation in transformation frequency in *Streptococcus pneumoniae***. *ISME J* 2013, **7**(4):791-799.

99. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M: **Population genomics of post-vaccine changes in pneumococcal epidemiology**. *Nat Genet* 2013, **45**(6):656-663.

100. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ *et al*: **Dense genomic sampling identifies highways of pneumococcal recombination**. *Nat Genet* 2014, **46**(3):305-309.

101. McDonnell M, Lain R, Tomasz A: **"Diplophage": a bacteriophage of Diplococcus pneumoniae**. *Virology* 1975, **63**(2):577-582.

102. Tiraby JG, Tiraby E, Fox MS: **Pneumococcal bacteriophages**. *Virology* 1975, **68**(2):566-569.

103. Ramirez M, Severina E, Tomasz A: **A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae***. *J Bacteriol* 1999, **181**(12):3618-3625.

104. Loeffler JM, Fischetti VA: **Lysogeny of *Streptococcus pneumoniae* with MM1 phage: improved adherence and other phenotypic changes**. *Infect Immun* 2006, **74**(8):4486-4495.

105. Yother J: **Capsules**. In: *The Pneumococcus.* Edited by Tuomanen EI. Washington D.C.: ASM Press; 2004.

106. Watson DA, Musher DM, Verhoef J: **Pneumococcal virulence factors and host immune responses to them**. *Eur J Clin Microbiol Infect Dis* 1995, **14**(6):479-490.

107. Calix JJ, Porambo RJ, Brady AM, Larson TR, Yother J, Abeygunwardana C, Nahm MH: **Biochemical, genetic, and serological characterization of two capsule subtypes among *Streptococcus pneumoniae* Serotype 20 strains: discovery of a new pneumococcal serotype**. *J Biol Chem* 2012, **287**(33):27885-27894.

108. Calix JJ, Nahm MH: **A new pneumococcal serotype, 11E, has a variably inactivated *wcjE* gene**. *J Infect Dis* 2010, **202**(1):29-38.

109. Jin P, Kong F, Xiao M, Oftadeh S, Zhou F, Liu C, Russell F, Gilbert GL: **First report of putative *Streptococcus pneumoniae* serotype 6D among nasopharyngeal isolates from Fijian children**. *J Infect Dis* 2009, **200**(9):1375-1380.

110. Oliver MB, van der Linden MP, Kuntzel SA, Saad JS, Nahm MH: **Discovery of *Streptococcus pneumoniae* serotype 6 variants with glycosyltransferases synthesizing two differing repeating units**. *J Biol Chem* 2013, **288**(36):25976-25985.

111. van Dam JE, Fleer A, Snippe H: **Immunogenicity and immunochemistry of*Streptococcus pneumoniae* capsular polysaccharides**. *Antonie van Leeuwenhoek* 1990, **58**(1):1-47.

112. Kauffmann F, Lund E, Eddy BE: **Proposal for a change in the nomenclature of *Diplococcus pneumoniae* and a comparison of the Danish and American type designations**. *International Journal of Systematic and Evolutionary Microbiology* 1960, **10**(1):31-40.

113. Henrichsen J: **Six newly recognized types of *Streptococcus pneumoniae***. *J Clin Microbiol* 1995, **33**(10):2759-2762.

114. Sorensen UB, Henrichsen J, Chen HC, Szu SC: **Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of *Streptococcus pneumoniae* revealed by immunochemical methods**. *Microb Pathog* 1990, **8**(5):325-334.

115. Cartee RT, Forsee WT, Jensen JW, Yother J: **Expression of the *Streptococcus pneumoniae* type 3 synthase in *Escherichia coli*. Assembly of type 3 polysaccharide on a lipid primer**. *J Biol Chem* 2001, **276**(52):48831-48839.

116. Dillard JP, Vandersea MW, Yother J: **Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae***. *J Exp Med* 1995, **181**(3):973-983.

117. Arrecubieta C, Lopez R, Garcia E: **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* 1994, **176**(20):6375-6383.

118. Llull D, Munoz R, Lopez R, Garcia E: **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* 1999, **190**(2):241-251.

119. Waite RD, Penfold DW, Struthers JK, Dowson CG: **Spontaneous sequence duplications within capsule genes *cap8E* and *tts* control phase variation in**

***Streptococcus pneumoniae* serotypes 8 and 37**. *Microbiology* 2003, **149**(Pt 2):497-504.

120. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen HB, Nahm MH: **Pneumococcal capsules and their types: past, present, and future**. *Clinical microbiology reviews* 2015, **28**(3):871-899.
121. Kim JO, Weiser JN: **Association of intrastrain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of *Streptococcus pneumoniae***. *J Infect Dis* 1998, **177**(2):368-377.
122. Cundell DR, Weiser JN, Shen J, Young A, Tuomanen EI: **Relationship between colonial morphology and adherence of *Streptococcus pneumoniae***. *Infect Immun* 1995, **63**(3):757-761.
123. Weiser JN, Kapoor M: **Effect of intrastrain variation in the amount of capsular polysaccharide on genetic transformation of *Streptococcus pneumoniae*: implications for virulence studies of encapsulated strains**. *Infect Immun* 1999, **67**(7):3690-3692.
124. Kim JO, Romero-Steiner S, Sorensen UB, Blom J, Carvalho M, Barnard S, Carlone G, Weiser JN: **Relationship between cell surface carbohydrates and intrastrain variation on opsonophagocytosis of *Streptococcus pneumoniae***. *Infect Immun* 1999, **67**(5):2327-2333.
125. Manso AS, Chai MH, Atack JM, Furi L, De Ste Croix M, Haigh R, Trappetti C, Ogunniyi AD, Shewell LK, Boitano M *et al*: **A random six-phase switch regulates pneumococcal virulence via global epigenetic changes**. *Nat Commun* 2014, **5**:5055.
126. Kolkman MA, Wakarchuk W, Nuijten PJ, van der Zeijst BA: **Capsular polysaccharide synthesis in *Streptococcus pneumoniae* serotype 14: molecular analysis of the complete cps locus and identification of genes encoding glycosyltransferases required for the biosynthesis of the tetrasaccharide subunit**. *Mol Microbiol* 1997, **26**(1):197-208.
127. Morona JK, Morona R, Paton JC: **Characterization of the locus encoding the *Streptococcus pneumoniae* type 19F capsular polysaccharide biosynthetic pathway**. *Mol Microbiol* 1997, **23**(4):751-763.
128. Mavroidi A, Godoy D, Aanensen DM, Robinson DA, Hollingshead SK, Spratt BG: **Evolutionary genetics of the capsular locus of serogroup 6 pneumococci**. *J Bacteriol* 2004, **186**(24):8181-8192.

129. Burton RL, Geno KA, Saad JS, Nahm MH: **Pneumococcus with the "6E" cps locus produces serotype 6B capsular polysaccharide**. *J Clin Microbiol* 2016.
130. van Selm S, van Cann LM, Kolkman MA, van der Zeijst BA, van Putten JP: **Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides**. *Infection and immunity* 2003, **71**(11):6192-6198.
131. Maynard GD: **An Enquiry Into the Etiology, Manifestations and Prevention of Pneumonia Amongst Natives on the Rand, Recruited from" tropical Areas."**: South African Institute for Medical Research; 1913.
132. Wright A, Morgan WP, Colebrook L, Dodgson R: **Observations ON PROPHYLACTIC INOCULATION AGAINST PNEUMOCOCCUS INFECTIONS, AND ON THE RESULTS WHICH HAVE BEEN ACHIEVED BY IT**. *The Lancet* 1914, **183**(4714):1-10.
133. Dochez AR, Avery OT: **The elaboration of specific soluble substance by pneumococcus during growth**. *The Journal of experimental medicine* 1917, **26**(4):477-493.
134. Dubos R, Avery OT: **Decomposition of the capsular polysaccharide of pneumococcus type III by a bacterial enzyme**. *The Journal of experimental medicine* 1931, **54**(1):51-71.
135. Ekwurzel G, Simmons J, Dublin LI, Felton LD: **Studies on immunizing substances in pneumococci: VIII. Report on field tests to determine the prophylactic value of a pneumococcus antigen**. *Public Health Reports (1896-1970)* 1938:1877-1893.
136. MacLeod CM, Hodges RG, Heidelberger M, Bernhard WG: **Prevention of pneumococcal pneumonia by immunization with specific capsular polysaccharides**. *The Journal of experimental medicine* 1945, **82**(6):445-465.
137. Gibbons JH, Lashof JC: **A Review of Selected Federal Vaccine and Immunization Policies, Based on Case**. 1979.
138. Austrian R, Douglas R, Schiffman G, Coetzee A, Koornhof H, Hayden-Smith S, Reid R: **Prevention of pneumococcal pneumonia by vaccination**. *Transactions of the Association of American Physicians* 1975, **89**:184-194.
139. SHAPIRO ED, CLEMENS JD: **A controlled evaluation of the protective efficacy of pneumococcal vaccine for patients at high risk of serious pneumococcal infections**. *Annals of internal medicine* 1984, **101**(3):325-330.

140. BOLAN G, BROOME CV, FACKLAM RR, PLIKAYTIS BD, FRASER DW, SCHLECH WF: **Pneumococcal vaccine efficacy in selected populations in the United States**. *Annals of internal medicine* 1986, **104**(1):1-6.

141. Sell SH, Wright PF, Vaughn WK, Thompson J, Schiffman G: **Clinical studies of pneumococcal vaccines in infants. I. Reactogenicity and immunogenicity of two polyvalent polysaccharide vaccines**. *Review of Infectious Diseases* 1981, **3**(Supplement 1):S97-S107.

142. Douglas R, Paton J, Duncan S, Hansman D: **Antibody response to pneumococcal vaccination in children younger than five years of age**. *Journal of infectious diseases* 1983, **148**(1):131-137.

143. Nuorti PJ, Butler JC, Breiman RF: **Prevention of pneumococcal disease: recommendations of the Advisory Committee on Immunization Practices (ACIP)**. *Morbidity and Mortality Weekly Report: Recommendations and Reports* 1997:i-24.

144. Avery OT, Goebel WF: **Chemo-immunological studies on conjugated carbohydrate-proteins: II. Immunological specificity of synthetic sugar-protein antigens**. *The Journal of experimental medicine* 1929, **50**(4):533.

145. Goebel WF, Avery OT: **CHEMO-IMMUNOLOGICAL STUDIES ON CONJUGATED CARBOHYDRATE-PROTEINS: I. THE SYNTHESIS OFp-AMINOPHENOL β-GLUCOSIDE, p-AMINOPHENOL β-GALACTOSIDE, AND THEIR COUPLING WITH SERUM GLOBULIN**. *The Journal of experimental medicine* 1929, **50**(4):521.

146. Rennels MB, Edwards KM, Keyserling HL, Reisinger KS, Hogerman DA, Madore DV, Chang I, Paradiso PR, Malinoski FJ, Kimura A: **Safety and immunogenicity of heptavalent pneumococcal vaccine conjugated to CRM197 in United States infants**. *Pediatrics* 1998, **101**(4):604-611.

147. Black S, Shinefield H, Fireman B, Lewis E, Ray P, Hansen JR, Elvin L, Ensor KM, Hackell J, Siber G *et al*: **Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Northern California Kaiser Permanente Vaccine Study Center Group**. *Pediatr Infect Dis J* 2000, **19**(3):187-195.

148. Croxtall JD, Keating GM: **Pneumococcal polysaccharide protein D-conjugate vaccine (Synflorix; PHiD-CV)**. *Paediatr Drugs* 2009, **11**(5):349-357.

149. Centers for Disease Control and Prevention (CDC): **Progress in introduction of pneumococcal, conjugate vaccine worldwide, 2000–2008**. *MMWR Morb Mortal Wkly Rep* 2008, **57**:1148-1151.
150. IVAC: **State of PCV Use and Impact Evaluations**. In.: Johns Hopkins Bloomberg School of Public Health; 2016.
151. Pilishvili T, Lexau C, Farley MM, Hadler J, Harrison LH, Bennett NM, Reingold A, Thomas A, Schaffner W, Craig AS *et al*: **Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine**. *J Infect Dis* 2010, **201**(1):32-41.
152. McFetridge R, Meulen AS, Folkerth SD, Hoekstra JA, Dallas M, Hoover PA, Marchese RD, Zacholski DM, Watson WJ, Stek JE *et al*: **Safety, tolerability, and immunogenicity of 15-valent pneumococcal conjugate vaccine in healthy adults**. *Vaccine* 2015, **33**(24):2793-2799.
153. Sobanjo-ter Meulen A, Vesikari T, Malacaman EA, Shapiro SA, Dallas MJ, Hoover PA, McFetridge R, Stek JE, Marchese RD, Hartzel J *et al*: **Safety, tolerability and immunogenicity of 15-valent pneumococcal conjugate vaccine in toddlers previously vaccinated with 7-valent pneumococcal conjugate vaccine**. *Pediatr Infect Dis J* 2015, **34**(2):186-194.
154. Lin PL, Michaels MG, Janosky J, Ortenzo M, Wald ER, Mason EO, Jr.: **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* 2003, **111**(4 Pt 1):896-899.
155. Whitney CG, Farley MM, Hadler J, Harrison LH, Bennett NM, Lynfield R, Reingold A, Cieslak PR, Pilishvili T, Jackson D *et al*: **Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine**. *N Engl J Med* 2003, **348**(18):1737-1746.
156. Kaplan SL, Mason EO, Jr., Wald ER, Schutze GE, Bradley JS, Tan TQ, Hoffman JA, Givner LB, Yogev R, Barson WJ: **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* 2004, **113**(3 Pt 1):443-449.
157. Block SL, Hedrick J, Harrison CJ, Tyler R, Smith A, Findlay R, Keegan E: **Community-wide vaccination with the heptavalent pneumococcal conjugate**

significantly alters the microbiology of acute otitis media. *Pediatr Infect Dis J* 2004, **23**(9):829-833.

158. Casey JR, Pichichero ME: **Changes in frequency and pathogens causing acute otitis media in 1995-2003**. *Pediatr Infect Dis J* 2004, **23**(9):824-828.

159. Weinberger DM, Malley R, Lipsitch M: **Serotype replacement in disease after pneumococcal vaccination**. *Lancet* 2011, **378**(9807):1962-1973.

160. Moore MR, Hyde TB, Hennessy TW, Parks DJ, Reasonover AL, Harker-Jones M, Gove J, Bruden DL, Rudolph K, Parkinson A *et al*: **Impact of a conjugate vaccine on community-wide carriage of nonsusceptible *Streptococcus pneumoniae* in Alaska**. *J Infect Dis* 2004, **190**(11):2031-2038.

161. Huang SS, Platt R, Rifas-Shiman SL, Pelton SI, Goldmann D, Finkelstein JA: **Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004**. *Pediatrics* 2005, **116**(3):e408-413.

162. Park SY, Moore MR, Bruden DL, Hyde TB, Reasonover AL, Harker-Jones M, Rudolph KM, Hurlburt DA, Parks DJ, Parkinson AJ *et al*: **Impact of conjugate vaccine on transmission of antimicrobial-resistant *Streptococcus pneumoniae* among Alaskan children**. *Pediatr Infect Dis J* 2008, **27**(4):335-340.

163. Pai R, Moore MR, Pilishvili T, Gertz RE, Whitney CG, Beall B: **Postvaccine genetic structure of *Streptococcus pneumoniae* serotype 19A from children in the United States**. *J Infect Dis* 2005, **192**(11):1988-1995.

164. Lipsitch M, O'Neill K, Cordy D, Bugalter B, Trzcinski K, Thompson CM, Goldstein R, Pelton S, Huot H, Bouchet V *et al*: **Strain characteristics of *Streptococcus pneumoniae* carriage and invasive disease isolates during a cluster-randomized clinical trial of the 7-valent pneumococcal conjugate vaccine**. *J Infect Dis* 2007, **196**(8):1221-1227.

165. Jacobs MR, Good CE, Bajaksouzian S, Windau AR: **Emergence of *Streptococcus pneumoniae* serotypes 19A, 6C, and 22F and serogroup 15 in Cleveland, Ohio, in relation to introduction of the protein-conjugated pneumococcal vaccine**. *Clin Infect Dis* 2008, **47**(11):1388-1395.

166. Kaplan SL, Barson WJ, Lin PL, Stovall SH, Bradley JS, Tan TQ, Hoffman JA, Givner LB, Mason EO, Jr.: **Serotype 19A Is the most common serotype causing invasive pneumococcal infections in children**. *Pediatrics* 2010, **125**(3):429-436.

167. Pelton SI, Huot H, Finkelstein JA, Bishop CJ, Hsu KK, Kellenberg J, Huang SS, Goldstein R, Hanage WP: **Emergence of 19A as virulent and multidrug**

resistant Pneumococcus in Massachusetts following universal immunization of infants with pneumococcal conjugate vaccine. *Pediatr Infect Dis J* 2007, **26**(6):468-472.

168. Brueggemann AB, Pai R, Crook DW, Beall B: **Vaccine escape recombinants emerge after pneumococcal vaccination in the United States**. *PLoS Pathog* 2007, **3**(11):e168.

169. Beall B, McEllistrem MC, Gertz RE, Jr., Wedel S, Boxrud DJ, Gonzalez AL, Medina MJ, Pai R, Thompson TA, Harrison LH *et al*: **Pre- and postvaccination clonal compositions of invasive pneumococcal serotypes for isolates collected in the United States in 1999, 2001, and 2002**. *J Clin Microbiol* 2006, **44**(3):999-1017.

170. Golubchik T, Brueggemann AB, Street T, Gertz RE, Jr., Spencer CC, Ho T, Giannoulatou E, Link-Gelles R, Harding RM, Beall B *et al*: **Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event**. *Nat Genet* 2012, **44**(3):352-355.

171. Moore MR, Link-Gelles R, Schaffner W, Lynfield R, Lexau C, Bennett NM, Petit S, Zansky SM, Harrison LH, Reingold A *et al*: **Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance**. *Lancet Infect Dis* 2015, **15**(3):301-309.

172. Hakenbeck R, Tarpay M, Tomasz A: **Multiple changes of penicillin-binding proteins in penicillin-resistant clinical isolates of *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 1980, **17**(3):364-371.

173. Spratt BG: **Distinct penicillin binding proteins involved in the division, elongation, and shape of *Escherichia coli* K12**. *Proc Natl Acad Sci U S A* 1975, **72**(8):2999-3003.

174. Spratt BG, Pardee AB: **Penicillin-binding proteins and cell shape in *E. coli***. *Nature* 1975, **254**(5500):516-517.

175. HAKENBECK R, BRIESE T, ELLERBROK H: **Antibodies against the benzylpenicilloyl moiety as a probe for penicillin-binding proteins**. *European Journal of Biochemistry* 1986, **157**(1):101-106.

176. Grebe T, Hakenbeck R: **Penicillin-binding proteins 2b and 2x of *Streptococcus pneumoniae* are primary resistance determinants for different classes of beta-lactam antibiotics**. *Antimicrob Agents Chemother* 1996, **40**(4):829-834.

177. Munoz R, Dowson CG, Daniels M, Coffey TJ, Martin C, Hakenbeck R, Spratt BG: **Genetics of resistance to third-generation cephalosporins in clinical isolates of *Streptococcus pneumoniae***. *Mol Microbiol* 1992, **6**(17):2461-2465.

178. Krauss J, Hakenbeck R: **A mutation in the D,D-carboxypeptidase penicillin-binding protein 3 of *Streptococcus pneumoniae* contributes to cefotaxime resistance of the laboratory mutant C604**. *Antimicrob Agents Chemother* 1997, **41**(5):936-942.

179. Schuster C, Dobrinski B, Hakenbeck R: **Unusual septum formation in *Streptococcus pneumoniae* mutants with an alteration in the D,D-carboxypeptidase penicillin-binding protein 3**. *J Bacteriol* 1990, **172**(11):6499-6505.

180. Severin A, Schuster C, Hakenbeck R, Tomasz A: **Altered murein composition in a DD-carboxypeptidase mutant of *Streptococcus pneumoniae***. *J Bacteriol* 1992, **174**(15):5152-5155.

181. Dowson CG, Coffey TJ, Kell C, Whiley RA: **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* 1993, **9**(3):635-643.

182. Dowson CG, Hutchison A, Brannigan JA, George RC, Hansman D, Linares J, Tomasz A, Smith JM, Spratt BG: **Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae***. *Proc Natl Acad Sci U S A* 1989, **86**(22):8842-8846.

183. Hakenbeck R, Briese T, Chalkley L, Ellerbrok H, Kalliokoski R, Latorre C, Leinonen M, Martin C: **Antigenic variation of penicillin-binding proteins from penicillin-resistant clinical strains of *Streptococcus pneumoniae***. *J Infect Dis* 1991, **164**(2):313-319.

184. Hakenbeck R, Briese T, Chalkley L, Ellerbrok H, Kalliokoski R, Latorre C, Leinonen M, Martin C: **Variability of penicillin-binding proteins from penicillin-sensitive *Streptococcus pneumoniae***. *J Infect Dis* 1991, **164**(2):307-312.

185. Weber B, Ehlert K, Diehl A, Reichmann P, Labischinski H, Hakenbeck R: **The *fib* locus in *Streptococcus pneumoniae* is required for peptidoglycan crosslinking and PBP-mediated beta-lactam resistance**. *FEMS Microbiol Lett* 2000, **188**(1):81-85.

186. Filipe SR, Tomasz A: **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* 2000, **97**(9):4891-4896.

187. Mascher T, Heintz M, Zahner D, Merai M, Hakenbeck R: **The CiaRH system of *Streptococcus pneumoniae* prevents lysis during stress induced by treatment with cell wall inhibitors and by mutations in pbp2x involved in beta-lactam resistance**. *J Bacteriol* 2006, **188**(5):1959-1968.

188. Tran TD, Kwon HY, Kim EH, Kim KW, Briles DE, Pyo S, Rhee DK: **Decrease in penicillin susceptibility due to heat shock protein ClpL in *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 2011, **55**(6):2714-2728.

189. Hansman D, Bullen MM: **A Resistant Pneumococcus**. *Lancet* 1967, **2**(7509):264-&.

190. Hansman D, Glasgow H, Sturt J, Devitt L, Douglas R: **Increased resistance to penicillin of pneumococci isolated from man**. *N Engl J Med* 1971, **284**(4):175-177.

191. Jacobs MR, Koornhof HJ, Robins-Browne RM, Stevenson CM, Vermaak ZA, Freiman I, Miller GB, Witcomb MA, Isaacson M, Ward JI *et al*: **Emergence of multiply resistant pneumococci**. *N Engl J Med* 1978, **299**(14):735-740.

192. Klugman KP: **Pneumococcal resistance to antibiotics**. *Clin Microbiol Rev* 1990, **3**(2):171-196.

193. Widdowson CA, Klugman KP: **The molecular mechanisms of tetracycline resistance in the pneumococcus**. *Microb Drug Resist* 1998, **4**(1):79-84.

194. Widdowson CA, Klugman KP, Hanslo D: **Identification of the tetracycline resistance gene, *tet(O)*, in *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 1996, **40**(12):2891-2893.

195. Evans W, Hansman D: **Tetracycline-resistant pneumococcus**. *The Lancet* 1963, **281**(7278):451.

196. Lee NY, Song JH, Kim S, Peck KR, Ahn KM, Lee SI, Yang YH, Li J, Chongthaleong A, Tiengrim S *et al*: **Carriage of antibiotic-resistant pneumococci among Asian children: A multinational surveillance by the Asian network for surveillance of resistant pathogens (ANSORP)**. *Clinical Infectious Diseases* 2001, **32**(10):1463-1469.

197. Monro RE, Marcker KA: **Ribosome-catalysed reaction of puromycin with a formylmethionine-containing oligonucleotide**. *J Mol Biol* 1967, **25**(2):347-350.

198. Monro RE, Vazquez D: **Ribosome-catalysed peptidyl transfer: effects of some inhibitors of protein synthesis**. *J Mol Biol* 1967, **28**(1):161-165.

199. Dang-Van A, Tiraby G, Acar JF, Shaw WV, Bouanchaud DH: **Chloramphenicol resistance in *Streptococcus pneumoniae*: enzymatic acetylation and possible plasmid linkage**. *Antimicrob Agents Chemother* 1978, **13**(4):577-583.

200. Shaw WV, Leslie AG: **Chloramphenicol acetyltransferase**. *Annu Rev Biophys Biophys Chem* 1991, **20**:363-386.

201. Vijayakumar MN, Priebe SD, Guild WR: **Structure of a conjugative element in *Streptococcus pneumoniae***. *J Bacteriol* 1986, **166**(3):978-984.

202. Jacobs MR, Felmingham D, Appelbaum PC, Gruneberg RN: **The Alexander Project 1998-2000: susceptibility of pathogens isolated from community-acquired respiratory tract infection to commonly used antimicrobial agents**. *J Antimicrob Chemother* 2003, **52**(2):229-246.

203. Weisblum B: **Erythromycin resistance by ribosome modification**. *Antimicrob Agents Chemother* 1995, **39**(3):577-585.

204. Roberts MC, Sutcliffe J, Courvalin P, Jensen LB, Rood J, Seppala H: **Nomenclature for macrolide and macrolide-lincosamide-streptogramin B resistance determinants**. *Antimicrob Agents Chemother* 1999, **43**(12):2823-2830.

205. Del Grosso M, Iannelli F, Messina C, Santagati M, Petrosillo N, Stefani S, Pozzi G, Pantosti A: **Macrolide efflux genes *mef(A)* and *mef(E)* are carried by different genetic elements in *Streptococcus pneumoniae***. *J Clin Microbiol* 2002, **40**(3):774-778.

206. Gay K, Baughman W, Miller Y, Jackson D, Whitney CG, Schuchat A, Farley MM, Tenover F, Stephens DS: **The emergence of *Streptococcus pneumoniae* resistant to macrolide antimicrobial agents: a 6-year population-based assessment**. *J Infect Dis* 2000, **182**(5):1417-1424.

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

208. Daly MM, Doktor S, Flamm R, Shortridge D: **Characterization and prevalence of MefA, MefE, and the associated *msr(D)* gene in *Streptococcus pneumoniae* clinical isolates**. *J Clin Microbiol* 2004, **42**(8):3570-3574.

209. Santagati M, Iannelli F, Cascone C, Campanile F, Oggioni MR, Stefani S, Pozzi G: **The novel conjugative transposon tn1207.3 carries the macrolide efflux gene *mef(A)* in *Streptococcus pyogenes***. *Microb Drug Resist* 2003, **9**(3):243-247.

210. Santagati M, Iannelli F, Oggioni MR, Stefani S, Pozzi G: **Characterization of a genetic element carrying the macrolide efflux gene *mef(A)* in *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 2000, **44**(9):2585-2587.

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

212. Kim SH, Song JH, Chung DR, Thamlikitkul V, Yang Y, Wang H, Lu M, So TM, Hsueh PR, Yasin RM *et al*: **Changing trends in antimicrobial resistance and serotypes of *Streptococcus pneumoniae* isolates in Asian countries: an Asian Network for Surveillance of Resistant Pathogens (ANSORP) study**. *Antimicrob Agents Chemother* 2012, **56**(3):1418-1426.

213. Courvalin P, Ounissi H, Arthur M: **Multiplicity of macrolide-lincosamide-streptogramin antibiotic resistance determinants**. *J Antimicrob Chemother* 1985, **16 Suppl A**:91-100.

214. Adrian PV, Klugman KP: **Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 1997, **41**(11):2406-2413.

215. Maskell JP, Sefton AM, Hall LM: **Multiple mutations modulate the function of dihydrofolate reductase in trimethoprim-resistant *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 2001, **45**(4):1104-1108.

216. Maskell JP, Sefton AM, Hall LM: **Mechanism of sulfonamide resistance in clinical isolates of *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 1997, **41**(10):2121-2126.

217. Enright M, Zawadski P, Pickerill P, Dowson CG: **Molecular evolution of rifampicin resistance in *Streptococcus pneumoniae***. *Microb Drug Resist* 1998, **4**(1):65-70.

218. Padayachee T, Klugman KP: **Molecular basis of rifampin resistance in *Streptococcus pneumoniae***. *Antimicrob Agents Chemother* 1999, **43**(10):2361-2365.

219. Breiman RF, Butler JC, Tenover FC, Elliott JA, Facklam RR: **Emergence of drug-resistant pneumococcal infections in the United States**. *JAMA* 1994, **271**(23):1831-1835.

220. Linares J, Pallares R, Alonso T, Perez JL, Ayats J, Gudiol F, Viladrich PF, Martin R: **Trends in antimicrobial resistance of clinical isolates of *Streptococcus pneumoniae* in Bellvitge Hospital, Barcelona, Spain (1979-1990)**. *Clin Infect Dis* 1992, **15**(1):99-105.

221. Doern GV, Brueggemann A, Holley HP, Jr., Rauch AM: **Antimicrobial resistance of *Streptococcus pneumoniae* recovered from outpatients in the United States during the winter months of 1994 to 1995: results of a 30-center national surveillance study**. *Antimicrob Agents Chemother* 1996, **40**(5):1208-1213.

222. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 1995, **269**(5223):496-512.

223. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ *et al*: **Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae***. *Science* 2001, **293**(5529):498-506.

224. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C *et al*: **Genomics. Genome project standards in a new era of sequencing**. *Science* 2009, **326**(5950):236-237.

225. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**(12):5463-5467.

226. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

227. Schuster SC: **Next-generation sequencing transforms today's biology**. *Nat Methods* 2008, **5**(1):16-18.

228. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.

229. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**(5910):133-138.

230. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA: **Pacific biosciences sequencing technology for genotyping and variation discovery in human data**. *BMC Genomics* 2012, **13**:375.

231. Granger D, Boily-Larouche G, Turgeon P, Weiss K, Roger M: **Molecular characteristics of *pbp1a* and *pbp2b* in clinical *Streptococcus pneumoniae* isolates in Quebec, Canada**. *J Antimicrob Chemother* 2006, **57**(1):61-70.

232. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Jr., Walker H, Hawkins PA, Tran T, Whitney CG, McGee L *et al*: **Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting beta-Lactam Resistance Levels in *Streptococcus pneumoniae***. *MBio* 2016, **7**(3).

233. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci U S A* 2005, **102**(39):13950-13955.

234. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J *et al*: **Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome**. *J Bacteriol* 2007, **189**(22):8186-8195.

235. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR *et al*: **Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species**. *Genome Biol* 2010, **11**(10):R107.

236. van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, Klugman KP, von Gottberg A, Bentley SD, Parkhill J *et al*: **Defining the estimated core genome of bacterial populations using a Bayesian decision model**. *PLoS Comput Biol* 2014, **10**(8):e1003788.

237. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST *et al*: **Complement factor H polymorphism in age-related macular degeneration**. *Science* 2005, **308**(5720):385-389.

238. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D: **Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter***. *Proc Natl Acad Sci U S A* 2013, **110**(29):11923-11927.

239. Chen PE, Shapiro BJ: **The advent of genome-wide association studies for bacteria**. *Curr Opin Microbiol* 2015, **25**:17-24.

240. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C *et al*: **Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes**. *PLoS Genet* 2014, **10**(8):e1004547.

241. Kristinsson KG: **Epidemiology of penicillin resistant pneumococci in Iceland**. *Microb Drug Resist* 1995, **1**(2):121-125.

242. Tomasson G, Gudnason T, Kristinsson KG: **Dynamics of pneumococcal carriage among healthy Icelandic children attending day-care centres**. *Scand J Infect Dis* 2005, **37**(6-7):422-428.

243. Hjaltested EK, Bernatoniene J, Erlendsdottir H, Kaltenis P, Bernatoniene G, Gudnason T, Haraldsson A, Kristinsson KG: **Resistance in respiratory tract pathogens and antimicrobial use in Icelandic and Lithuanian children**. *Scand J Infect Dis* 2003, **35**(1):21-26.

244. Kristinsson KG: **Effect of antimicrobial use and other risk factors on antimicrobial resistance in pneumococci**. *Microb Drug Resist* 1997, **3**(2):117-123.

245. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, Hakenbeck R, Hryniewicz W, Lefevre JC, Tomasz A *et al*: **Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network**. *J Clin Microbiol* 2001, **39**(7):2565-2571.

246. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

247. Gladstone RA, Jefferies JM, Tocheva AS, Beard KR, Garley D, Chong WW, Bentley SD, Faust SN, Clarke SC: **Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction**. *Vaccine* 2015, **33**(17):2015-2021.

248. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R *et al*: **The European Nucleotide Archive**. *Nucleic Acids Res* 2011, **39**(Database issue):D28-31.

249. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**(5):821-829.

250. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ *et al*: **Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain**. *Microbiology* 2012, **158**(Pt 4):1005-1015.

251. Weinstein MP, Klugman KP, Jones RN: **Rationale for revised penicillin susceptibility breakpoints versus *Streptococcus pneumoniae*: coping with antimicrobial susceptibility in an era of resistance**. *Clin Infect Dis* 2009, **48**(11):1596-1600.

252. Leclercq R, Cantón R, Brown DF, Giske CG, Heisig P, MacGowan AP, Mouton JW, Nordmann P, Rodloff AC, Rossolini GM: **EUCAST expert rules in antimicrobial susceptibility testing**. *Clinical Microbiology and Infection* 2013, **19**(2):141-160.

253. van Tonder AJ, Bray JE, Roalfe L, White R, Zancolli M, Quirk SJ, Haraldsson G, Jolley KA, Maiden MC, Bentley SD *et al*: **Genomics Reveals the Worldwide Distribution of Multidrug-Resistant Serotype 6E Pneumococci**. *J Clin Microbiol* 2015, **53**(7):2271-2285.

254. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE**. *Bioinformatics* 2011, **27**(4):578-579.

255. Nadalin F, Vezzi F, Policriti A: **GapFiller: a de novo assembly approach to fill the gap within paired reads**. *BMC Bioinformatics* 2012, **13 Suppl 14**:S8.

256. Ponstingl H, Ning Z: **SMALT-a new mapper for DNA sequencing reads**. *F1000 Posters* 2010, **1**:313.

257. Park IH, Park S, Hollingshead SK, Nahm MH: **Genetic basis for the new pneumococcal serotype, 6C**. *Infect Immun* 2007, **75**(9):4482-4489.

258. Bratcher PE, Kim KH, Kang JH, Hong JY, Nahm MH: **Identification of natural pneumococcal isolates expressing serotype 6D by genetic, biochemical and serological characterization**. *Microbiology* 2010, **156**(Pt 2):555-560.

259. Ko KS, Baek JY, Song JH: **Capsular gene sequences and genotypes of "serotype 6E" *Streptococcus pneumoniae* isolates**. *J Clin Microbiol* 2013, **51**(10):3395-3399.

260. Oliver MB, Jones C, Larson TR, Calix JJ, Zartler ER, Yother J, Nahm MH: ***Streptococcus pneumoniae* serotype 11D has a bispecific glycosyltransferase and expresses two different capsular polysaccharide repeating units**. *J Biol Chem* 2013, **288**(30):21945-21954.

261. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
262. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.
263. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies**. *Mol Biol Evol* 2006, **23**(2):254-267.
264. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J: **Roary: rapid large-scale prokaryote pan genome analysis**. *Bioinformatics* 2015, **31**(22):3691-3693.
265. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.
266. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome research* 2003, **13**(9):2178-2189.
267. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**(3):e9490.
268. Didelot X, Wilson DJ: **ClonalFrameML: efficient inference of recombination in whole bacterial genomes**. *PLoS Comput Biol* 2015, **11**(2):e1004041.
269. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data**. *Genetics* 2007, **175**(3):1251-1266.
270. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W475-478.
271. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2013, **41**(Database issue):D36-42.
272. Leinonen R, Sugawara H, Shumway M: **The sequence read archive**. *Nucleic Acids Res* 2011, **39**(Database issue):D19-21.
273. Wyres KL, van Tonder A, Lambertsen LM, Hakenbeck R, Parkhill J, Bentley SD, Brueggemann AB: **Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated prior to 1974**. *BMC Genomics* 2013, **14**:500.
274. Cody AJ, McCarthy NM, Wimalarathna HL, Colles FM, Clark L, Bowler IC, Maiden MC, Dingle KE: **A longitudinal 6-year study of the molecular epidemiology of clinical *Campylobacter* isolates in Oxfordshire, United kingdom**. *J Clin Microbiol* 2012, **50**(10):3193-3201.

275. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

276. Casella G, Berger RL: **Statistical inference**, vol. 70: Duxbury Press Belmont, CA; 1990.

277. Jaynes ET: **Probability theory: the logic of science**: Cambridge University Press; 2003.

278. Jukes THaC, C.R.: **Evolution of Protein Molecules**. In. New York: Academic Press; 1969: 21-132.

279. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome Res* 2009, **19**(9):1639-1645.

280. Silverman BW: **Density estimation for statistics and data analysis**, vol. 26: CRC press; 1986.

281. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T *et al*: **eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges**. *Nucleic Acids Res* 2012, **40**(Database issue):D284-289.

282. Bryant D, Moulton V: **Neighbor-net: an agglomerative method for the construction of phylogenetic networks**. *Mol Biol Evol* 2004, **21**(2):255-265.

283. Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC: **Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing**. *J Clin Microbiol* 2013, **51**(8):2526-2534.

284. Heinsbroek E, Ladhani S, Gray S, Guiver M, Kaczmarski E, Borrow R, Ramsay M: **Added value of PCR-testing for confirmation of invasive meningococcal disease in England**. *J Infect* 2013, **67**(5):385-390.

285. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Linares J, Jacobs MR, Kristinsson KG, Beall BW, Klugman KP *et al*: **The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success**. *Genome Biol* 2012, **13**(11):R103.

286. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2012, **40**(Database issue):D290-301.

287. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

288. Meyer Y, Buchanan BB, Vignols F, Reichheld JP: **Thioredoxins and glutaredoxins: unifying elements in redox biology**. *Annu Rev Genet* 2009, **43**:335-367.

289. Lu J, Vlamis-Gardikas A, Kandasamy K, Zhao R, Gustafsson TN, Engstrand L, Hoffner S, Engman L, Holmgren A: **Inhibition of bacterial thioredoxin reductase: an antibiotic mechanism targeting bacteria lacking glutathione**. *FASEB J* 2013, **27**(4):1394-1403.

290. Ren J, Kotaka M, Lockyer M, Lamb HK, Hawkins AR, Stammers DK: **GTP cyclohydrolase II structure and mechanism**. *J Biol Chem* 2005, **280**(44):36912-36919.

291. Poulsen P, Jensen KF: **Three genes preceding *pyrE* on the *Escherichia coli* chromosome are essential for survival and normal cell morphology in stationary culture and at high temperature**. *Res Microbiol* 1991, **142**(2-3):283-288.

292. Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, Goesmann A, Frosch M, Muller T, Vogel U *et al*: **Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome**. *PLoS One* 2011, **6**(4):e18441.

293. Kong Y, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, Alexander DC, Hao W: **Homologous Recombination Drives both Sequence Diversity and Gene Content Variation in *Neisseria meningitidis***. *Genome Biol Evol* 2013.

294. Castillo-Ramirez S, Harris SR, Holden MT, He M, Parkhill J, Bentley SD, Feil EJ: **The impact of recombination on dN/dS within recently emerged bacterial clones**. *PLoS Pathog* 2011, **7**(7):e1002129.

295. Boissy R, Ahmed A, Janto B, Earl J, Hall BG, Hogg JS, Pusch GD, Hiller LN, Powell E, Hayes J *et al*: **Comparative supragenomic analyses among the pathogens *Staphylococcus aureus, Streptococcus pneumoniae,* and *Haemophilus influenzae* using a modification of the finite supragenome model**. *BMC Genomics* 2011, **12**:187.

296. Hall BG, Ehrlich GD, Hu FZ: **Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing**. *Microbiology* 2010, **156**(Pt 4):1060-1068.

297. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA *et al*: **Evolution of MRSA during hospital transmission and intercontinental spread**. *Science* 2010, **327**(5964):469-474.

298. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H *et al*: **A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic**. *Genome Res* 2013, **23**(4):653-664.

299. Cirlan M, Saad M, Coman G, Bilal NE, Elbashier AM, Kreft D, Snijders S, van Leeuwen W, van Belkum A: **International spread of major clones of methicillin resistant *Staphylococcus aureus*: nosocomial endemicity of multi locus sequence type 239 in Saudi Arabia and Romania**. *Infect Genet Evol* 2005, **5**(4):335-339.

300. Edgeworth JD, Yadegarfar G, Pathak S, Batra R, Cockfield JD, Wyncoll D, Beale R, Lindsay JA: **An outbreak in an intensive care unit of a strain of methicillin-resistant *Staphylococcus aureus* sequence type 239 associated with an increased rate of vascular access device-related bacteremia**. *Clin Infect Dis* 2007, **44**(4):493-501.

301. Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J *et al*: **Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study**. *Lancet Infect Dis* 2013, **13**(2):130-136.

302. Hiller NL, Eutsey RA, Powell E, Earl JP, Janto B, Martin DP, Dawid S, Ahmed A, Longwell MJ, Dahlgren ME *et al*: **Differences in genotype and virulence among four multidrug-resistant *Streptococcus pneumoniae* isolates belonging to the PMEN1 clone**. *PLoS One* 2011, **6**(12):e28850.

303. Lefebure T, Bitar PD, Suzuki H, Stanhope MJ: **Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept**. *Genome Biol Evol* 2010, **2**:646-655.

304. Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, Hallin PF, Newell DG, Toszeghy M, Ridley A, Manning G *et al*: **Genomic characterization of *Campylobacter jejuni* strain M1**. *PLoS One* 2010, **5**(8):e12253.

305. Biggs PJ, Fearnhead P, Hotter G, Mohan V, Collins-Emerson J, Kwan E, Besser TE, Cookson A, Carter PE, French NP: **Whole-genome comparison of two**

Campylobacter jejuni **isolates of the same sequence type reveals multiple loci of different ancestral lineage**. *PLoS One* 2011, **6**(11):e27121.

306. Dunning Hotopp JC, Grifantini R, Kumar N, Tzeng YL, Fouts D, Frigimelica E, Draghi M, Giuliani MM, Rappuoli R, Stephens DS *et al*: **Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes**. *Microbiology* 2006, **152**(Pt 12):3733-3749.
307. Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O *et al*: **Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis***. *Proc Natl Acad Sci U S A* 2008, **105**(9):3473-3478.
308. Schoen C, Tettelin H, Parkhill J, Frosch M: **Genome flexibility in *Neisseria meningitidis***. *Vaccine* 2009, **27 Suppl 2**:B103-111.
309. Rusniok C, Vallenet D, Floquet S, Ewles H, Mouze-Soulama C, Brown D, Lajus A, Buchrieser C, Medigue C, Glaser P *et al*: **NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen *Neisseria meningitidis***. *Genome Biol* 2009, **10**(10):R110.
310. Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV *et al*: ***Neisseria meningitidis*** **is structured in clades associated with restriction modification systems that modulate homologous recombination**. *Proc Natl Acad Sci U S A* 2011, **108**(11):4494-4499.
311. Krauland MG, Dunning Hotopp JC, Riley DR, Daugherty SC, Marsh JW, Messonnier NE, Mayer LW, Tettelin H, Harrison LH: **Whole genome sequencing to investigate the emergence of clonal complex 23 *Neisseria meningitidis* serogroup Y disease in the United States**. *PLoS One* 2012, **7**(4):e35699.
312. Castillo-Ramirez S, Corander J, Marttinen P, Aldeljawi M, Hanage WP, Westh H, Boye K, Gulay Z, Bentley SD, Parkhill J *et al*: **Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus***. *Genome Biol* 2012, **13**(12):R126.
313. McClain MS, Shaffer CL, Israel DA, Peek RM, Jr., Cover TL: **Genome sequence analysis of *Helicobacter pylori* strains associated with gastric ulceration and gastric cancer**. *BMC Genomics* 2009, **10**:3.

314. Fischer W, Windhager L, Rohrer S, Zeiller M, Karnholz A, Hoffmann R, Zimmer R, Haas R: **Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer**. *Nucleic Acids Res* 2010, **38**(18):6089-6101.

315. Kawai M, Furuta Y, Yahara K, Tsuru T, Oshima K, Handa N, Takahashi N, Yoshida M, Azuma T, Hattori M *et al*: **Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes**. *BMC Microbiol* 2011, **11**:104.

316. Lara-Ramirez EE, Segura-Cabrera A, Guo X, Yu G, Garcia-Perez CA, Rodriguez-Perez MA: **New implications on genomic adaptation derived from the *Helicobacter pylori* genome comparison**. *PLoS One* 2011, **6**(2):e17300.

317. Gunaletchumy SP, Teh X, Khosravi Y, Ramli NS, Chua EG, Kavitha T, Mason JN, Lee HT, Alias H, Zaidan NZ *et al*: **Draft genome sequences of *Helicobacter pylori* isolates from Malaysia, cultured from patients with functional dyspepsia and gastric cancer**. *J Bacteriol* 2012, **194**(20):5695-5696.

318. Feikin DR, Klugman KP: **Historical changes in pneumococcal serogroup distribution: implications for the era of pneumococcal conjugate vaccines**. *Clin Infect Dis* 2002, **35**(5):547-555.

319. WHO: **Pneumococcal vaccines WHO position paper–2012–recommendations**. *Vaccine* 2012, **30**(32):4717-4718.

320. Park IH, Pritchard DG, Cartee R, Brandao A, Brandileone MC, Nahm MH: **Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae***. *J Clin Microbiol* 2007, **45**(4):1225-1233.

321. Hausdorff WP, Hoet B, Adegbola RA: **Predicting the impact of new pneumococcal conjugate vaccines: serotype composition is not enough**. *Expert Rev Vaccines* 2015, **14**(3):413-428.

322. Millar EV, Pimenta FC, Roundtree A, Jackson D, Carvalho Mda G, Perilla MJ, Reid R, Santosham M, Whitney CG, Beall BW *et al*: **Pre- and post-conjugate vaccine epidemiology of pneumococcal serotype 6C invasive disease and carriage within Navajo and White Mountain Apache communities**. *Clin Infect Dis* 2010, **51**(11):1258-1265.

323. Grant LR, O'Brien SE, Burbidge P, Haston M, Zancolli M, Cowell L, Johnson M, Weatherholtz RC, Reid R, Santosham M *et al*: **Comparative immunogenicity of**

7 and 13-valent pneumococcal conjugate vaccines and the development of functional antibodies to cross-reactive serotypes. *PLoS One* 2013, **8**(9):e74906.

324. Kieninger DM, Kueper K, Steul K, Juergens C, Ahlers N, Baker S, Jansen KU, Devlin C, Gruber WC, Emini EA *et al*: **Safety, tolerability, and immunologic noninferiority of a 13-valent pneumococcal conjugate vaccine compared to a 7-valent pneumococcal conjugate vaccine given with routine pediatric vaccinations in Germany**. *Vaccine* 2010, **28**(25):4192-4203.
325. Vesikari T, Karvonen A, Korhonen T, Karppa T, Sadeharju K, Fanic A, Dieussaert I, Schuerman L: **Immunogenicity of 10-valent pneumococcal nontypeable *Haemophilus influenzae* Protein D Conjugate Vaccine when administered as catch-up vaccination to children 7 months to 5 years of age**. *Pediatr Infect Dis J* 2011, **30**(8):e130-141.
326. Choi EH, Lee HJ, Cho EY, Oh CE, Eun BW, Lee J, Kim MJ: **Prevalence and genetic structures of *Streptococcus pneumoniae* serotype 6D, South Korea**. *Emerg Infect Dis* 2010, **16**(11):1751-1753.
327. Lee H, Cha JH, Nahm MH, Burton RL, Kim KH: **The 7-valent pneumococcal conjugate vaccine elicits cross-functional opsonophagocytic killing responses to *Streptococcus pneumoniae* serotype 6D in children**. *BMC Infect Dis* 2013, **13**:474.
328. Park IH, Geno KA, Yu J, Oliver MB, Kim KH, Nahm MH: **Genetic, biochemical, and serological characterization of a new pneumococcal serotype, 6H, and generation of a pneumococcal strain producing three different capsular repeat units**. *Clin Vaccine Immunol* 2015, **22**(3):313-318.
329. Elberse K, Witteveen S, van der Heide H, van de Pol I, Schot C, van der Ende A, Berbers G, Schouls L: **Sequence diversity within the capsular genes of *Streptococcus pneumoniae* serogroup 6 and 19**. *PLoS One* 2011, **6**(9):e25018.
330. Baek JY, Park IH, So TM, Lalitha MK, Shimono N, Yasin RM, Carlos CC, Perera J, Thamlikitkul V, Hsueh PR *et al*: **Prevalence and characteristics of *Streptococcus pneumoniae* "putative serotype 6E" isolates from Asian countries**. *Diagn Microbiol Infect Dis* 2014, **80**(4):334-337.
331. Kawaguchiya M, Urushibara N, Kobayashi N: **High prevalence of genotype 6E (putative serotype 6E) among noninvasive/colonization isolates of *Streptococcus pneumoniae* in northern Japan**. *Microb Drug Resist* 2015, **21**(2):209-214.

332. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA: **PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods**. *BMC Bioinformatics* 2012, **13**:87.

333. Coffey TJ, Berron S, Daniels M, Garcia-Leoni ME, Cercenado E, Bouza E, Fenoll A, Spratt BG: **Multiply antibiotic-resistant *Streptococcus pneumoniae* recovered from Spanish hospitals (1988-1994): novel major clones of serotypes 14, 19F and 15F**. *Microbiology* 1996, **142 (Pt 10)**:2747-2757.

334. Smith AM, Klugman KP: **Three predominant clones identified within penicillin-resistant South African isolates of *Streptococcus pneumoniae***. *Microb Drug Resist* 1997, **3**(4):385-389.

335. Sibold C, Wang J, Henrichsen J, Hakenbeck R: **Genetic relationships of penicillin-susceptible and -resistant *Streptococcus pneumoniae* strains isolated on different continents**. *Infect Immun* 1992, **60**(10):4119-4126.

336. Gherardi G, Whitney CG, Facklam RR, Beall B: **Major related sets of antibiotic-resistant Pneumococci in the United States as determined by pulsed-field gel electrophoresis and pbp1a-pbp2b-pbp2x-dhf restriction profiles**. *J Infect Dis* 2000, **181**(1):216-229.

337. Bratcher PE, Park IH, Oliver MB, Hortal M, Camilli R, Hollingshead SK, Camou T, Nahm MH: **Evolution of the capsular gene locus of *Streptococcus pneumoniae* serogroup 6**. *Microbiology* 2011, **157**(Pt 1):189-198.

338. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

339. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods**. *Mol Biol Evol* 2011, **28**(10):2731-2739.

340. Findlow H, Borrow R, Andrews N, Waight P, Sheasby E, Matheson M, England A, Goldblatt D, Ashton L, Findlow J *et al*: **Immunogenicity of a single dose of meningococcal group C conjugate vaccine given at 3 months of age to healthy infants in the United kingdom**. *Pediatr Infect Dis J* 2012, **31**(6):616-622.

341. Andrews NJ, Waight PA, Burbidge P, Pearce E, Roalfe L, Zancolli M, Slack M, Ladhani SN, Miller E, Goldblatt D: **Serotype-specific effectiveness and correlates of protection for the 13-valent pneumococcal conjugate vaccine: a postlicensure indirect cohort study**. *Lancet Infect Dis* 2014, **14**(9):839-846.

342. Burton RL, Nahm MH: **Development of a fourfold multiplexed opsonophagocytosis assay for pneumococcal antibodies against additional serotypes and discovery of serological subtypes in *Streptococcus pneumoniae* serotype 20**. *Clin Vaccine Immunol* 2012, **19**(6):835-841.

343. Turner P, Turner C, Jankhot A, Helen N, Lee SJ, Day NP, White NJ, Nosten F, Goldblatt D: **A longitudinal study of *Streptococcus pneumoniae* carriage in a cohort of infants and their mothers on the Thailand-Myanmar border**. *PLoS One* 2012, **7**(5):e38271.

344. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, Sa-Leao R, Song JH, Ko KS, Beall B *et al*: **Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone**. *BMC Biol* 2014, **12**(1):49.

345. Coffey TJ, Daniels M, Enright MC, Spratt BG: **Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of *Streptococcus pneumoniae* arose by large recombinational replacements of the *cpsA-pbp1a* region**. *Microbiology* 1999, **145 (Pt 8)**:2023-2031.

346. Coffey TJ, Enright MC, Daniels M, Morona JK, Morona R, Hryniewicz W, Paton JC, Spratt BG: **Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae***. *Mol Microbiol* 1998, **27**(1):73-83.

347. Coffey TJ, Enright MC, Daniels M, Wilkinson P, Berron S, Fenoll A, Spratt BG: **Serotype 19A variants of the Spanish serotype 23F multiresistant clone of *Streptococcus pneumoniae***. *Microb Drug Resist* 1998, **4**(1):51-55.

348. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Linares J, Jacobs MR, Kristinsson KG, Beall BW, Klugman KP *et al*: **Pneumococcal Capsular Switching: An Historical Perspective**. *J Infect Dis* 2012.

349. Beall BW, Gertz RE, Hulkower RL, Whitney CG, Moore MR, Brueggemann AB: **Shifting genetic structure of invasive serotype 19A pneumococci in the United States**. *J Infect Dis* 2011, **203**(10):1360-1368.

350. Fitzwater SP, Chandran A, Santosham M, Johnson HL: **The worldwide impact of the seven-valent pneumococcal conjugate vaccine**. *Pediatr Infect Dis J* 2012, **31**(5):501-508.

351. Satzke C, Dunne EM, Porter BD, Klugman KP, Mulholland EK: **The PneuCarriage Project: A Multi-Centre Comparative Study to Identify the**

Best Serotyping Methods for Examining Pneumococcal Carriage in Vaccine Evaluation Studies. *PLoS Med* 2015, **12**(11):e1001903; discussion e1001903.

352. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, Feltwell T, Fookes MC, Harris SR, Dordel J, Salter SJ *et al*: **Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection**. *PLoS Genet* 2013, **9**(10):e1003868.
353. Cornick JE, Chaguza C, Harris SR, Yalcin F, Senghore M, Kiran AM, Govindpershad S, Ousmane S, Du Plessis M, Pluschke G: **Region-specific diversification of the highly virulent serotype 1 *Streptococcus pneumoniae***. *Microbial Genomics* 2015, **1**(2).
354. Yother J: **Capsules of *Streptococcus pneumoniae* and other bacteria: paradigms for polysaccharide biosynthesis and regulation**. *Annu Rev Microbiol* 2011, **65**:563-581.
355. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.
356. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis version 6.0**. *Mol Biol Evol* 2013, **30**(12):2725-2729.
357. Pimenta FC, Gertz RE, Jr., Roundtree A, Yu J, Nahm MH, McDonald RR, Carvalho Mda G, Beall BW: **Rarely occurring 19A-like cps locus from a serotype 19F pneumococcal isolate indicates continued need of serology-based quality control for PCR-based serotype determinations**. *J Clin Microbiol* 2009, **47**(7):2353-2354.
358. Menezes AP, Reis JN, Ternes YM, Andrade AL, Pimenta FC, Carvalho Mda G, Beall B: **Update of pneumococcal PCR serotyping assay for detection of a commonly occurring type 19F wzy variant in Brazil**. *J Clin Microbiol* 2013, **51**(7):2470-2471.
359. Dunne EM, Tikkanen L, Balloch A, Gould K, Yoannes M, Phuanukoonnon S, Licciardi PV, Russell FM, Mulholland EK, Satzke C *et al*: **Characterization of 19A-like 19F pneumococcal isolates from Papua New Guinea and Fiji**. *New Microbes New Infect* 2015, **7**:86-88.
360. Ding F, Tang P, Hsu MH, Cui P, Hu S, Yu J, Chiu CH: **Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant *Streptococcus pneumoniae* serotype 14**. *BMC Genomics* 2009, **10**:158.

361. Coffey TJ, Dowson CG, Daniels M, Zhou J, Martin C, Spratt BG, Musser JM: **Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae***. *Mol Microbiol* 1991, **5**(9):2255-2260.

362. Munoz R, Coffey TJ, Daniels M, Dowson CG, Laible G, Casal J, Hakenbeck R, Jacobs M, Musser JM, Spratt BG *et al*: **Intercontinental spread of a multiresistant clone of serotype 23F *Streptococcus pneumoniae***. *J Infect Dis* 1991, **164**(2):302-306.

363. Munoz R, Musser JM, Crain M, Briles DE, Marton A, Parkinson AJ, Sorensen U, Tomasz A: **Geographic distribution of penicillin-resistant clones of *Streptococcus pneumoniae*: characterization by penicillin-binding protein profile, surface protein A typing, and multilocus enzyme analysis**. *Clin Infect Dis* 1992, **15**(1):112-118.

364. Lefevre JC, Bertrand MA, Faucon G: **Molecular analysis by pulsed-field gel electrophoresis of penicillin-resistant *Streptococcus pneumoniae* from Toulouse, France**. *Eur J Clin Microbiol Infect Dis* 1995, **14**(6):491-497.

365. McDougal LK, Rasheed JK, Biddle JW, Tenover FC: **Identification of multiple clones of extended-spectrum cephalosporin-resistant *Streptococcus pneumoniae* isolates in the United States**. *Antimicrob Agents Chemother* 1995, **39**(10):2282-2288.

366. Sloas MM, Barrett FF, Chesney PJ, English BK, Hill BC, Tenover FC, Leggiadro RJ: **Cephalosporin treatment failure in penicillin- and cephalosporin-resistant *Streptococcus pneumoniae* meningitis**. *Pediatr Infect Dis J* 1992, **11**(8):662-666.

367. Marton A, Gulyas M, Munoz R, Tomasz A: **Extremely high incidence of antibiotic resistance in clinical isolates of *Streptococcus pneumoniae* in Hungary**. *J Infect Dis* 1991, **163**(3):542-548.

368. Shi ZY, Enright MC, Wilkinson P, Griffiths D, Spratt BG: **Identification of three major clones of multiply antibiotic-resistant *Streptococcus pneumoniae* in Taiwanese hospitals by multilocus sequence typing**. *J Clin Microbiol* 1998, **36**(12):3514-3519.

369. Syrogiannopoulos GA, Grivea IN, Beratis NG, Spiliopoulou AE, Fasola EL, Bajaksouzian S, Appelbaum PC, Jacobs MR: **Resistance patterns of**

Streptococcus pneumoniae **from carriers attending day-care centers in southwestern Greece**. *Clin Infect Dis* 1997, **25**(2):188-194.

370. Syrogiannopoulos GA, Ronchetti F, Dagan R, Grivea I, Ronchetti MP, Porat N, Davies TA, Ronchetti R, Appelbaum PC, Jacobs MR: **Mediterranean clone of penicillin-susceptible, multidrug-resistant serotype 6B *Streptococcus pneumoniae* in Greece, Italy and Israel**. *Int J Antimicrob Agents* 2000, **16**(3):219-224.
371. Porat N, Trefler R, Dagan R: **Persistence of two invasive *Streptococcus pneumoniae* clones of serotypes 1 and 5 in comparison to that of multiple clones of serotypes 6B and 23F among children in southern Israel**. *J Clin Microbiol* 2001, **39**(5):1827-1832.
372. Brueggemann AB, Spratt BG: **Geographic distribution and clonal diversity of *Streptococcus pneumoniae* serotype 1 isolates**. *J Clin Microbiol* 2003, **41**(11):4966-4970.
373. Henriques Normark B, Kalin M, Ortqvist A, Akerlund T, Liljequist BO, Hedlund J, Svenson SB, Zhou J, Spratt BG, Normark S *et al*: **Dynamics of penicillin-susceptible clones in invasive pneumococcal disease**. *J Infect Dis* 2001, **184**(7):861-869.
374. Enright MC, Spratt BG: **A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease**. *Microbiology* 1998, **144 (Pt 11)**:3049-3060.
375. Brandileone MC, Di Fabio JL, Vieira VS, Zanella RC, Casagrande ST, Pignatari AC, Tomasz A: **Geographic distribution of penicillin resistance of *Streptococcus pneumoniae* in Brazil: genetic relatedness**. *Microb Drug Resist* 1998, **4**(3):209-217.
376. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, **30**(14):3059-3066.
377. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M *et al*: **eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences**. *Nucleic Acids Res* 2016, **44**(D1):D286-293.

378. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J: **Hierarchical and spatially explicit clustering of DNA sequences with BAPS software**. *Mol Biol Evol* 2013, **30**(5):1224-1228.

379. van Opijnen T, Bodi KL, Camilli A: **Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms**. *Nat Methods* 2009, **6**(10):767-772.

380. Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA: **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database**. *Bioinformatics* 2008, **24**(23):2672-2676.

381. Sullivan MJ, Petty NK, Beatson SA: **Easyfig: a genome comparison visualizer**. *Bioinformatics* 2011, **27**(7):1009-1010.

382. Hakenbeck R, Grebe T, Zahner D, Stock JB: B**eta-lactam resistance in *Streptococcus pneumoniae*: penicillin-binding proteins and non-penicillin-binding proteins**. *Mol Microbiol* 1999, **33**(4):673-678.

383. Lambertsen L, Brendstrup M, Friis H, Christensen JJ: **Molecular characterization of invasive penicillin non-susceptible *Streptococcus pneumoniae* from Denmark, 2001 to 2005**. *Scand J Infect Dis* 2010, **42**(5):333-340.

384. Aruni AW, Mishra A, Dou Y, Chioma O, Hamilton BN, Fletcher HM: ***Filifactor alocis*--a new emerging periodontal pathogen**. *Microbes Infect* 2015, **17**(7):517-530.

385. Guerillot R, Da Cunha V, Sauvage E, Bouchier C, Glaser P: **Modular evolution of TnGBSs, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading**. *J Bacteriol* 2013, **195**(9):1979-1990.

386. Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, Bentley SD, Garcia E, Mitchell TJ: **Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages**. *J Bacteriol* 2009, **191**(15):4854-4862.

387. Syrogiannopoulos GA, Grivea IN, Davies TA, Katopodis GD, Appelbaum PC, Beratis NG: **Antimicrobial use and colonization with erythromycin-resistant *Streptococcus pneumoniae* in Greece during the first 2 years of life**. *Clin Infect Dis* 2000, **31**(4):887-893.

388. Syrjanen RK, Kilpi TM, Kaijalainen TH, Herva EE, Takala AK: **Nasopharyngeal carriage of *Streptococcus pneumoniae* in Finnish children younger than 2 years old**. *J Infect Dis* 2001, **184**(4):451-459.

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

390. Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY *et al*: **Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes**. *Nat Commun* 2016, **7**:12797.

391. Read TD, Massey RC: **Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology**. *Genome Med* 2014, **6**(11):109.

392. Balding DJ: **A tutorial on statistical methods for population association studies**. *Nat Rev Genet* 2006, **7**(10):781-791.

393. Price AL, Zaitlen NA, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies**. *Nat Rev Genet* 2010, **11**(7):459-463.

394. Tu AH, Fulgham RL, McCrory MA, Briles DE, Szalai AJ: **Pneumococcal surface protein A inhibits complement activation by *Streptococcus pneumoniae***. *Infect Immun* 1999, **67**(9):4720-4724.

395. Shaper M, Hollingshead SK, Benjamin WH, Jr., Briles DE: **PspA protects *Streptococcus pneumoniae* from killing by apolactoferrin, and antibody to PspA enhances killing of pneumococci by apolactoferrin [corrected]**. *Infect Immun* 2004, **72**(9):5031-5040.

396. Wani JH, Gilbert JV, Plaut AG, Weiser JN: **Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of *Streptococcus pneumoniae***. *Infection and immunity* 1996, **64**(10):3967-3974.

397. Rizk G, Lavenier D, Chikhi R: **DSK: k-mer counting with very low memory usage**. *Bioinformatics* 2013:btt020.

398. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

399. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.
400. Turner SD: **qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots**. *bioRxiv* 2014.
401. Bah T: **Inkscape: Guide to a Vector Drawing Program (Digital Short Cut)**: Pearson Education; 2009.
402. Rajakovich LJ, Tomlinson J, Dos Santos PC: **Functional Analysis of Bacillus subtilis Genes Involved in the Biosynthesis of 4-Thiouridine in tRNA**. *J Bacteriol* 2012, **194**(18):4933-4940.
403. Webb E, Claas K, Downs DM: **Characterization of *thiI*, a new gene involved in thiazole biosynthesis in *Salmonella typhimurium***. *J Bacteriol* 1997, **179**(13):4399-4402.
404. Mueller EG, Palenchar PM, Buck CJ: **The role of the cysteine residues of ThiI in the generation of 4-thiouridine in tRNA**. *J Biol Chem* 2001, **276**(36):33588-33595.
405. Attali C, Frolet C, Durmort C, Offant J, Vernet T, Di Guilmi AM: ***Streptococcus pneumoniae* choline-binding protein E interaction with plasminogen/plasmin stimulates migration across the extracellular matrix**. *Infect Immun* 2008, **76**(2):466-476.
406. Mann B, Orihuela C, Antikainen J, Gao G, Sublett J, Korhonen TK, Tuomanen E: **Multifunctional role of choline binding protein G in pneumococcal pathogenesis**. *Infect Immun* 2006, **74**(2):821-829.
407. Cieslewicz MJ, Kasper DL, Wang Y, Wessels MR: **Functional analysis in type Ia group B *Streptococcus* of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci**. *J Biol Chem* 2001, **276**(1):139-146.
408. Mollerach M, Lopez R, Garcia E: **Characterization of the *galU* gene of *Streptococcus pneumoniae* encoding a uridine diphosphoglucose pyrophosphorylase: a gene essential for capsular polysaccharide biosynthesis**. *J Exp Med* 1998, **188**(11):2047-2056.
409. Song XM, Connor W, Hokamp K, Babiuk LA, Potter AA: ***Streptococcus pneumoniae* early response genes to human lung epithelial cells**. *BMC Res Notes* 2008, **1**:64.

410. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL *et al*: **Identifying lineage effects when controlling for population structure improves power in bacterial association studies**. *Nature Microbiology* 2016, **1**:16041.

411. Sandgren A, Albiger B, Orihuela CJ, Tuomanen E, Normark S, Henriques-Normark B: **Virulence in mice of pneumococcal clonal types with known invasive disease potential in humans**. *J Infect Dis* 2005, **192**(5):791-800.

412. Briles DE, Hollingshead SK, King J, Swift A, Braun PA, Park MK, Ferguson LM, Nahm MH, Nabors GS: **Immunization of humans with recombinant pneumococcal surface protein A (rPspA) elicits antibodies that passively protect mice from fatal infection with *Streptococcus pneumoniae* bearing heterologous PspA**. *J Infect Dis* 2000, **182**(6):1694-1701.

413. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, Barash SC, Rosen CA, Masure HR, Tuomanen E *et al*: **Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection**. *Infect Immun* 2001, **69**(3):1593-1598.

414. Darrieux M, Goulart C, Briles D, Leite LC: **Current status and perspectives on protein-based pneumococcal vaccines**. *Crit Rev Microbiol* 2015, **41**(2):190-200.

415. Giefing C, Meinke AL, Hanner M, Henics T, Bui MD, Gelbmann D, Lundberg U, Senn BM, Schunn M, Habel A *et al*: **Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies**. *J Exp Med* 2008, **205**(1):117-131.

416. Ling E, Feldman G, Portnoi M, Dagan R, Overweg K, Mulholland F, Chalifa-Caspi V, Wells J, Mizrachi-Nebenzahl Y: **Glycolytic enzymes associated with the cell surface of *Streptococcus pneumoniae* are antigenic in humans and elicit protective immune responses in the mouse**. *Clin Exp Immunol* 2004, **138**(2):290-298.

417. Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC: **A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes**. *BMC Genomics* 2014, **15**:1138.

418. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP: **Diversification of bacterial genome content through distinct mechanisms over different timescales**. *Nat Commun* 2014, **5**:5471.