The Population Structure of Bacterial Pathogens

DPhil Thesis

Eleanor Rose Watkins

Department of Zoology

Merton College, University of Oxford

Trinity Term 2015

Abstract

Populations of many bacterial pathogens are structured into discrete strains despite frequent genetic exchange. Although insights from several studies suggest that different selection pressures operate in different areas of the genome, few theoretical and conceptual frameworks of bacterial population structure examine the effects of distinct competitive processes acting on different loci in the genome, and fewer still have considered the metabolic genes despite their increasingly-recognised importance in virulence. Buckee *et al*. (2008) investigated the combined effects of ecological competition operating at metabolic genes and immunological competition at antigenic genes on strain diversity using a stochastic model. A key prediction was that prevalent strains should show stable, non-overlapping associations between alleles of antigenic and metabolic genes. In this work, the roles of ecological and immunological competition in structuring bacterial pathogen populations is explored further using a multidisciplinary approach of mathematical modelling and whole genome analysis, focusing in particular on the well-characterised pathogens *Neisseria meningitidis* and *Streptococcus pneumoniae* as detailed case studies. We find support for the hypothesis that ecological and immunological competition play roles in structuring pathogen populations – even if present at relatively low levels – including support at the whole genome level. After extending the mathematical framework to encompass virulence-associated genes, we explore the effects of strain-targeted vaccination. We find that metabolic and virulence-associated genes from vaccine-strains are observed following vaccination in association with non-vaccine strains. The final chapter demonstrates how multilocus mathematical models coupled with whole genome analysis can be used together to gain additional insight into the evolution and population structure of bacterial pathogens.

Acknowledgements

I am indebted to Sunetra and Martin for their unwaning enthusiasm and support, and a surprising amount of faith in me ever since my time as an undergraduate during my summer project. I am grateful for their inspiring confidence in my abilities and helping me believe in myself – as well as seeing the lighter side of things – throughout the past six years. In addition, of course, to their valuable guidance, intellectual (and non-intellectual) discussions, and time spent with me. I'm not sure I would have taken up a DPhil – and let alone submit a thesis – if it wasn't for them.

I am very grateful to Merton College, Oxford University, for offering me the Jackson Scholarship which has funded my fees and stipend throughout my DPhil. I am grateful also for their Prize Scholarship and the Travel Grant awarded to me. Financial support aside, I would like to thank Merton College for a wonderful graduate experience.

The genetic data analysed in this thesis has come from several studies, each of which has required large amounts of time and money and involved many people. I am indebted to those who were involved with each of these studies, and for making their data publically available.

I would like to thank the current and past members of Sunetra and Martin's groups for providing advice, support and a lot of fun. In particular, I am grateful to Ben Ashby the MATLAB man for uncovering the many hidden joys of MATLAB, and Bridget Penman for her kind guidance throughout my time in Sunetra's group. I must also thank Julia Bennett for teaching me basic lab skills and Odile Harrison for helping me fathom genomics. Thanks also to KFOS and the gang for stimulating discussions on niche theory and chimaera-faced mugs.

Special thanks goes to Caroline Buckee for advice and guidance throughout my DPhil, and always bringing new ideas. I am grateful to Caroline for hosting my visit to the Harvard School of Public Health in June 2013, where I met some extraordinary people and had a highly intellectually-rich and rewarding time.

Thanks to my friends in Oxford for providing wholesome fun and support – both in Zoology and Merton. Thank you especially to Carinna Hockham and Jess Raj.

Thank you to Simon and Deborah Phillips for their support, advice, positivity and insight since we met 7 years ago.

I am eternally grateful to Mum, Dad and Pip for their endless patience and love. I can't thank them enough for everything they have ever done for me. As a constant source of support, comfort and re-assurance, they have always been my pillars of strength and I owe all of my achievements to them.

To Nick, my wonderful and kind husband, who has been the secret weapon behind everything I do. I must thank him for his love, encouragement and wisdom; his ability to see the positives; and always seeing the best in me.

List of Abbreviations

AT	Antigenic Type
ELISA	Enzyme-linked immunosorbent assay
ET	Electrophoretic Type
fHbp	Factor H binding protein
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MI	Mutual information
MLST	Multi Locus Sequence Typing
MLEE	Multi Locus Enzyme Electrophoresis
MPI	Modal percentage identity
MT	Metabolic Type
NadA	Neisserial Adhesin A
NHBA	Neisseria heparin binding antigen
NVT	Non-vaccine type
OMP	Outer membrane protein
OMV	Outer membrane vesicle
PAMP	Pathogen Associated Molecular Pattern
PcpA	Pneumococcal choline-binding protein
PCR	Polymerase Chain Reaction
PCV7	Heptavalent pneumococcal conjugate vaccine
PFGE	Pulsed-field gel electrophoresis
PspA	Pneumococcal surface protein A
PsaA	Pneumococcal surface antigen
RFLP	Restriction fragment length polymorphism
rMLST	Ribosomal Multi Locus Sequence Typing
SIR	Susceptible-Infected-Recovered

SLV	Single locus variant
SNP	Single nucleotide polymorphism
ST	Sequence type
VIMS	Vaccine Induced Metabolic Shift
VISR	Vaccine Induced Strain Replacement
RFLP	Variable number of tandem repeats
VT	Vaccine Type
wgMLST	Whole Genome Multi Locus Sequence Typing

Table of Contents

Chapter 1.

Introduction

1.1 Background

Many bacterial pathogens exhibit large amounts of diversity, which allows for the escape of recognition by the host immune response, and also complicates disease control interventions. The reasons for understanding the mechanisms responsible for the generation and maintenance of diversity in pathogen populations are twofold. Firstly, for the practical benefits of disease control, including predicting the effects of interventions such as vaccination and the circumstances under which antibiotic resistance may arise. Secondly, from a biological perspective, understanding the processes which maintain diversity in populations has remained an important question in ecology as well as infectious disease epidemiology, and several parallels can be drawn between the two.

Although bacterial pathogen populations manifest extensive diversity, many are simultaneously highly structured: members of such populations can be collected into 'strains' on the basis of a number of different properties, including antigenic or metabolic variants, as well as levels of virulence and drug resistance. The definition of a strain varies widely depending upon the methods used for discrimination and the question being asked.

In this work, I investigate the selective forces that maintain genetic diversity and structure in bacterial pathogen populations using deterministic mathematical models and whole genome analysis. I focus particularly on the well-characterised population structures of *Neisseria meningitidis* (the meningococcus), a major cause of meningitis and septicaemia worldwide,

and *Streptococcus pneumoniae* (the pneumococcus), which represents the main cause of pneumonia in children and adults and is also an important cause of meningitis and septicaemia. I present theoretical frameworks in which pathogen strains are divided into multiple genomic components, supported by analyses of genotypic and whole genomic data. Subsequently, I use such frameworks to simulate strain-targeted vaccination. In my final chapter, I provide an example of how such models can act as an analytical platform on which to base and interpret whole genome studies, in a genomic analysis of pandemic meningococcal lineages. Here, I introduce the work with a brief overview of the typing methods used to characterise genetic diversity in bacterial pathogen populations, followed by a discussion of the prevailing hypotheses and theoretical frameworks of bacterial population structure. I subsequently introduce the biology and epidemiology of *N. meningitidis* and *S. pneumoniae,* and conclude with a short section summarising various approaches to genomic analysis and the insights gained into bacterial population structure.

1.2 Genetic Diversity in Bacterial Populations

1.2.1 Characterisation of Antigenic Diversity

Several surface-exposed proteins and carbohydrates are highly variable in bacterial populations, as a result of their interactions with the host immune system and the high levels of diversifying selection thereby imposed. Many bacterial pathogens exhibit extensive antigenic diversity within the host to facilitate immune evasion. The antigenic pilin and Opa proteins of *Neisseria gonorrhoeae*, for example, have been demonstrated to vary during infection (Jerse *et al*. 1994; Seifert *et al*. 1994). Others vary less during the course of infection but display high levels of between-host diversity at the population level, such as *S. pneumoniae*

and *N. meningitidis*. It is primarily this population-level of antigenic diversity which this work seeks to explore further and explain.

Historically, methods of characterising antigenic diversity were based upon studying the reactions and cross-reactions of antigen extracts with their respective antisera. The Quellung reaction, first described in the early twentieth century by Fred Neufeld, involves the binding of antibodies to the antigen, causing it to become opaque and appear to enlarge (Neufeld & Handel 1909). The Quellung reaction was first described for characterising the polysaccharide capsule which surrounds *S. pneumoniae*, and contributed to the identification of over 90 capsular types, or serotypes. Sero-agglutination assays based on similar principles, such as latex agglutination and co-agglutination, are used today for a number of pathogens, including *S. pneumoniae* and *N. meningitidis* (Thirumoorthi & Dajani 1979; Lalitha *et al*. 1996). Enzyme-linked immunosorbent assays (ELISA) are frequently used also, which use antibodies linked to enzymes, and allow the quantification of antigen present (Engvall & Perlmann 1971).

However, for a number of antigenic targets, the panel of immunological reagents available is not comprehensive, with the consequence that not all isolates can be typed. Further, the target antigen is only recognised if it is expressed by the organism. Since the advent of molecular techniques, sequencing reactions of the genetic regions coding for antigens have been frequently used to characterise antigenic diversity. Many sequences coding for antigenic structures show variation "hotspots" with a high density of polymorphisms, which may correspond to structures which interact with the host immune system. It is possible, therefore, to gain an understanding of the diversity of these antigens through just sequencing those sequences which are highly variable. Variable regions of two outer membrane protein (OMP) antigens in *N. meningitidis*, for example, have been amplified as part of molecular typing schemes (Thompson *et al*. 2003; Russell *et al*. 2004). The porin PorA, which allows for the passage of ions across the cell membrane, has a number of membrane-exposed loops in its

protein structure, and the nucleotide sequences of three of these are highly genetically variable (VR1, VR2 and VR3) (Van der Ley *et al*. 1991). Regions VR1 and VR2 have been characterised as part of routine typing for a number of years, primarily using antibodies ("serosubtyping") but later replaced by sequencing reactions. FetA (otherwise known as FrpB) is an iron-regulated OMP with one hypervariable region (Saleem *et al*. 2013). This region of FetA in addition to VR1 and VR2 of PorA have been used in meningococcal typing schemes (Jolley *et al*. 2007).

Alternatively, many antigens in bacterial populations are part of large operons with a number of genetic components involved in their regulation and production, such as the polysaccharide capsule operon in *S. pneumoniae*. A number of multiplex PCR protocols have been developed to characterise the pneumococcal polysaccharide capsule, based on a combination of sequencing the more conserved *cpsA* and *cpsB* genes, and other serotype-specific or serogroup-specific genes, such as *wzy* and *wsz* (Brito *et al*. 2003; Kong & Gilbert 2003).

The decreasing costs of whole genome sequencing in recent years have enabled a growing number of analyses of antigenic diversity that exploit whole genomes. Nucleotide sequences coding for antigens – and allelic variants thereof – can be identified directly from genome sequences, therefore allowing the characterisation of a large number of antigens from one organism. Frequently, antigens which are controlled by phase variation lie in areas of the genome rich in repetitive sequences, which may not sequence and/or assemble well using current short-read sequencing technology. Further, analysing DNA/protein sequence variation alone does not provide information on the expression of such antigens, which may vary between different host sites or stages of infection.

1.2.2 Characterisation of Non-Antigenic Markers

In addition to variation among antigenic genes, a number of pathogens also exhibit extensive diversity in genes which are not responsible for surface-exposed structures. The patterns and structuring exhibited by these genes is thought to be different to those of the antigens, as many of the non-antigenic genes analysed in bacterial typing methods are housekeeping genes, which are considered to be subject to stabilising selection for conservation of function rather than diversifying selection. This use of "neutral" loci can provide information regarding the relatedness between different strains in populations. However, as discussed in subsequent sections, it is conceivable that variation in collections of metabolic housekeeping genes in bacterial pathogens may still contribute to important phenotypic differences among strains.

Conventionally, protein electrophoresis was used to compare phenotypic differences among variants of several enzymes, by comparing migration rates on a starch gel. Early studies of *Escherichia coli* isolates, for example, concluded that 94% of loci were polymorphic, with estimates of 100-1000 electrophoretic types (ETs) (Whittam *et al*. 1983). Although the technique of multilocus enzyme electrophoresis (MLEE) played an important role in establishing our current understanding of the extent of diversity in bacterial populations, it had several shortfalls also (reviewed by Maiden 2006). Its reliance upon phenotypic differences between strains meant that it would only detect those mutations which affected a protein's electrophoretic properties: approximately one twentieth of all possible mutations. Further, enzymes may have different amino acid sequences without having sufficiently different electrophoretic mobilities to give distinct bands on a gel. It is also difficult to compare results between different laboratories.

Multilocus sequence typing (MLST), developed in the late 1990s, also uses variation in housekeeping genes to discriminate among bacterial isolates, but makes use of nucleotide

sequence data (Maiden *et al*. 1998). MLST entails the sequencing of allele fragments at seven housekeeping loci which are dispersed around the chromosome (Tables 1.1 and 1.2). For each locus, unique allele fragments are assigned distinct identification numbers, and the alleles at each of the seven loci define the allelic profile, or sequence type (ST). An important reason for developing MLST was to account for the high rates of recombination in pathogen populations; allele-based methods such as MLST are able to account for the observation than polymorphisms between strains are more often generated by recombination than point mutation, as alleles are considered to be equally distinct from each other regardless of the number of polymorphisms they show.

STs are subsequently grouped into clonal complexes using a variety of techniques, including UPGMA-based methods using dendrograms (unweighted pair-group method with arithmetic averages), or the more recent clustering algorithm eBURST (Feil *et al*. 2004). For *N. meningitidis*, the STs within a clonal complex share at least four identical MLST loci, whereas clonal complexes for *S. pneumoniae* are defined more conservatively (by eBURST) as a cluster in which all STs are linked as SLVs (single locus variants) to at least one other ST. As well as being highly discriminatory, portable and universal, MLST data has been made freely available over the internet, on the PubMLST database (www.pubmlst.org; Jolley & Maiden 2010).

MLST Locus	Encoded enzyme/transporter	Function
abcZ	ABC transporter	Transport substrates across cell membranes
adk	Adenylate kinase	Catalyses the transfer of a phosphoryl group from ATP to AMP
aroE	Shikimate dehydrogenase	Required for the biosynthesis of chorismate
fumC	Fumarate hydratase	Part of the TCA cycle
gdh	Glucose-6-phosphate 1-dehydrogenase	Catalyses the initial step of the pentose phosphate pathway
pdhC	Pyruvate dehydrogenase subunit E1	Converts pyruvate into acetyl-CoA
pgm	Phosphoglucomutase	Catalyses the conversion of glucose 6-phosphate to glucose 1-phosphate

***Table 1.1. MLST loci used for* N. meningitidis *and the functions of the encoded genes*.**

MLST Locus	Encoded enzyme/transporter	Function
aroE	Shikimate dehydrogenase	Required for the biosynthesis of chorismate
ddl	D-alanine-D-alanine ligase	Assembles one of the sub-units used for peptidoglycan crosslinking; required for bacterial cell wall synthesis
gdh	Glucose-6-phosphate 1-dehydrogenase	Catalyses the initial step of the pentose phosphate pathway
gki	Glucose kinase	Facilitates phosphorylation of glucose to glucose-6-phosphate
recP	Transketolase	Involved in the pentose phosohate pathway
spi	Signal peptidase I	Cleaves the amino-terminal signal peptide from proteins that are translocated across membranes
xpt	Xanthine phosphoribosyltransferase	*De novo* synthesis of purine nucleotides

***Table 1.2. MLST loci used for* S. pneumoniae *and the functions of the encoded genes*.**

Other typing techniques which rely on genotypic differences between strains include the analysis of restriction fragment length polymorphism (RFLPs), single nucleotide polymorphisms (SNPs) and Multiple Locus Variable-number Tandem Repeat (VNTR) Analysis (MLVA). In RFLP analysis, restriction enzymes are used to cleave DNA at specific sequences, the locations of which vary in different strains. The resulting restriction fragments are then separated by gel electrophoresis. SNPs are variations observed at a single nucleotide site. SNP-based methods are particularly useful for some genetically monomorphic pathogens, whose levels of diversity are too low to be detected by MLST. MLVA analysis takes advantage of the variation in repeat unit sizes and repeat sequences among bacterial strains. Selected VNTRs are usually amplified by PCR, and the size of each amplified locus measured using either agarose gel electrophoresis or, more recently, capillary sequencers (van Belkum 2007). However, there are concerns that MLVA is not suitable for broad phylogenetic analyses due to the high mutation rate of loci amplified and thus large number of homoplasies (Achtman 2008).

Advances in sequencing technologies and bioinformatic analyses have led to exponential growth in the study of whole genomes in recent years (Koonin & Wolf 2008). Whole genome sequencing enables the analysis of variation across all shared loci between strains, providing extremely high resolution (discussed further in Section 1.6).

1.2.3 Bacterial Population Structures

Insights from typing techniques such as MLEE and MLST revealed that a myriad of population structures exist for bacteria: from populations with very low genetic diversity to those exhibiting such high levels of variation that almost no two strains are the same. The *Mycobacterium tuberculosis* complex for example, a closely related group of bacteria which

give rise to tuberculosis, display 99.9% nucleotide sequence identity (Sreevatsan *et al*. 1997). In contrast, for other bacteria, the presence of identical housekeeping sequences across strains is very rare. In a study of *Helicobacter pylori*, only three sequences were identical out of seven housekeeping loci and two virulence loci, across 20 strains (Achtman *et al*. 1999).

Results from MLEE surveys provided strong support for the clonal model, in which genetic exchange is effectively absent. Each member of the population is genetically identical to its parent cell, and genetic variation is generated only via mutation. As mutations accumulate, different genotypic clusters emerge, as changes are limited to the descendant of the cell in which the mutation occurs. The results of the MLEE experiments showed strong linkage disequilibrium in *E. coli* and *Salmonella* species, which implied clonality as recombination rates were not sufficient to break up clonal associations (Ochman & Selander 1984; Whittam *et al*. 1983).

The amount of genetic variation in the datasets generated by MLEE, however, seemed lower than expected for populations with large genetically effective population sizes such as *E. coli.* Several electrophoretically identical and closely related clones were identified in a number of unrelated hosts, suggesting that only a limited number of *E. coli* lineages exist (Selander & Levin 1980). Thus the clonal model was modified to account for reductions in genetic variation, which occur as a result of bottlenecks, leading to genetic drift, or periodic selection, in which fitter variants repeatedly eliminate the genetic diversity that has accumulated over time (Levin 1981). This was inferred as an important mechanism for the shaping of diversity in meningococcal populations during epidemic spread (Morelli *et al*. 1997).

The apparent widespread applicability of the clonal model was questioned in the late 1980s with the increasing accessibility to nucleotide sequencing. As more studies employed nucleotide sequencing to examine genetic variability in pathogen populations, evidence for the

occurrence of genetic exchange accumulated, in the form of mosaic genes, for example. Accumulating evidence for "localized sex" (Maynard Smith *et al*. 1991) led to the development of the non-clonal model of bacterial population structure. Recombination enables mutations to escape the lineage in which they arose, thereby preventing the emergence of stable clones. Given that recombination randomises genetic variation, non-clonal or panmictic populations do not depart from linkage equilibrium. *N. gonorrhoeae* and *H. pylori,* for example, both display high levels of polymorphism and random associations between genetic markers (Smith *et al*. 1993*;* Suerbaum *et al*. 1998).

Holmes *et al*. (1999) investigated the extent to which recombination disrupts the population structuring processes imposed by clonal reproduction in *N. meningitidis* using gene fragments of 12 house-keeping loci. They found that different loci yielded incongruent Maximum Likelihood trees with distinct branching patterns, and that identical alleles were present across genetically distinct isolates. They also found low levels of linkage disequilibrium when strains belonging to hyperinvasive lineages were removed from the analysis. Feil *et al*. also investigated the relative contributions of recombination and point mutations to the genetic divergence among meningococcal strains using seven-locus MLST data, and calculated a per-site recombination to mutation parameter (r/m) of 80 (Feil *et al*. 1999). Jolley *et al.* (2005) used a coalescent-based approach to estimate the rates of mutation and recombination in carried and invasive meningococcal populations, and found ratios of recombination to mutation rates ranging from 0.08 to 2.66. The results of all three studies suggest that recombination is frequent and an important genetic mechanism operating in meningococcal populations. Recombination is also believed to be an important process in structuring pneumococcal populations. Feil *et al*. (2000) used MLST data to estimate than that recombination has generated new alleles at a frequency approximately 10-fold higher than mutation. Indeed, rates may exceed those in meningococcal populations (Feil *et al*. 2001). The results of more recent studies using whole

genome data suggest that rates of recombination are not uniform across pneumococci and differ significantly between lineages (Croucher *et al*. 2013; Chewapreecha *et al*. 2014).

Few bacterial populations tend to exhibit strict clonal or non-clonal pathogen structures however, and most appear to occupy an intermediate position between the two extremes, with extensive genetic variation which is highly structured. *N. meningitidis* and *S. pneumoniae* manifest such "intermediate" population structures: despite frequent recombination, and a large number of STs identified, the high levels of diversity in these pathogens are highly structured into clonal complexes. Over 10000 STs have been identified in *N. meningitidis*, for example, which are grouped with similar strains into over 50 clonal complexes (10402 STs and 51 clonal complexes were found following a pubmlst.org database search on 22/07/15).

A number of clonal complexes/STs among several bacterial pathogens are associated with a higher risk of invasive disease or severe infection relative to others, some of which have been recorded to persist for many years and spread globally (Caugant *et al*. 1986; Musser *et al*. 1989; Enright & Spratt 1998; Brueggemann *et al*. 2003; Wertheim *et al*. 2005). In *N. meningitidis* for example, the ST-32 complex has been responsible for serogroup B epidemics worldwide over a 40 year period, since the late 1960s (Harrison *et al*. 2015). These observations seem paradoxical given that invasive disease in many of these pathogens does not contribute to the transmission process.

Further, despite high rates of recombination, many bacterial pathogens show associations between their major antigen(s) and MLST-type or MLEE-type. Non-overlapping associations are have been found between ST and serotype in *S. pneumoniae* in a number of countries worldwide (Table 1.3; Figure 1.1), and are also well-documented between clonal complex and several OMPs across a number of studies of *N. meningitidis* (Table 1.4). Natural fluctuations between pneumococcal ST and serotype and meningococcal ST and OMP type are also

apparent in some countries, however, as particular clones expand and decline over time – even in the absence of conjugate vaccination (Jefferies *et al*. 2010; Feikin & Klugman 2002; Buckee *et al*. 2008). Associations between antigenic type and MLST-type/MLEE-type are indeed found across a range of recombinogenic bacterial pathogens: Table 1.4 shows a list of bacterial pathogens with such associations; each pathogen has been previously been shown to demonstrate frequent recombination, which should homogenise such associations.

Intriguingly, these associations between antigenic type and MLST-type or MLEE-type appear to be stable over long time periods in a number of pathogens. Among *N. meningitidis* for example, identical associations between MLEE-type/clonal complex and several OMP antigens have been recorded over several decades. Caugant *et al.* (1987) used MLEE to investigate the associations between OMP antigens, MLEE type and serogroup for a variety of isolates, and found that stable combinations of OMP-type and ET were recorded over a period of 15 years. Urwin *et al*. (2004) analysed the diversity of three OMP antigens (PorA, PorB and FetA) among a representative global sample of 108 invasive meningococcal isolates, and akin to the results of Caugant *et al.* (1987), they found associations between clonal complex and OMP type which persisted over many years. Maximum Likelihood trees for both the MLST and OMP alleles were also congruent despite high rates of recombination. Buckee *et al*. also noted non-overlapping associations between PorA VR1 and VR2 and clonal complex (Buckee *et al*. 2008), and later FetA (Buckee *et al*. 2010), in a carriage sample from the Czech Republic spanning 27 years. Similarly, Bambini *et al*. (2013) found non-overlapping associations between clonal complex and the OMPs fHbp, NHBA and NadA over periods of approximately 20 years in the Netherlands. Identical associations between serotype and ST in *S. pneumoniae* have also been observed over many years (Table 1.5).

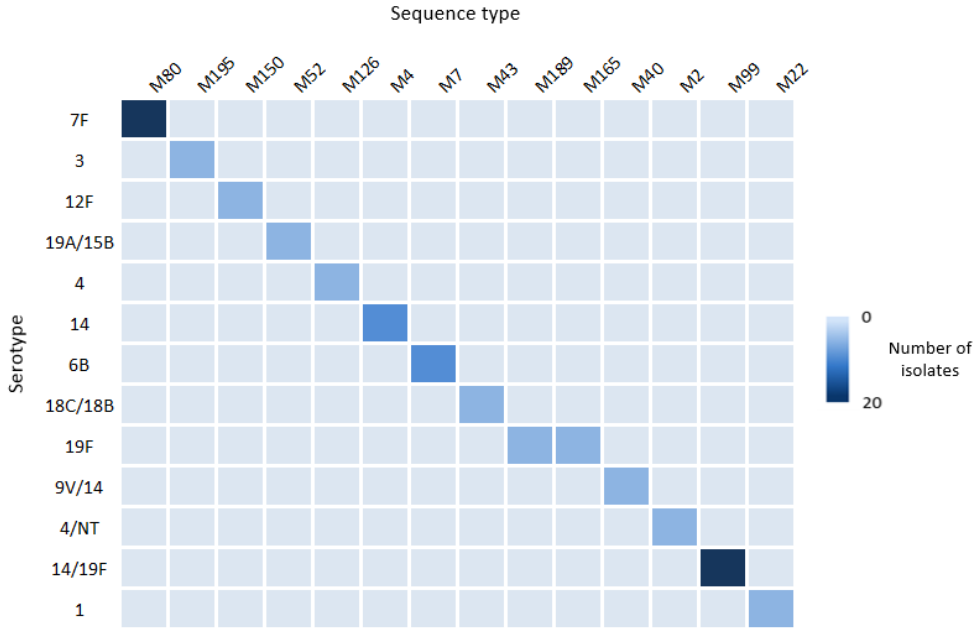


Figure 1.1. Associations between the most frequent pneumococcal serotypes and sequence types in a global sample *(using data from Enright & Spratt, 1998). The sequence types are named after the reference isolate for each sequence type, as stated in the original article.*

Sequence type/ Electrophoretic type	Country/countries included	No. isolates	Reference
ST	Scotland	217	(Clarke *et al*. 2006)
ST	Australia, Canada, Denmark, Finland, the Netherlands, Sweden, Great Britain, Uruguay	274	(Enright & Spratt 1998)
ST	UK	30	(Enright *et al*. 2000)
ST	UK	501	(Brueggemann *et al*. 2003)
ST	UK	310	(Tocheva *et al*. 2013)
ST	USA	1168	(Beall *et al*. 2006)
ST	Finland	224	(Hanage *et al*. 2005)
ST	Finland	437	(Hanage *et al*. 2004)
ST	Columbia	629	(Parra *et al*. 2014)
ST	UK	1030	(Pichon *et al*. 2013)
ST	Japan	66	(Tanaka *et al*. 2012)
ST	Taiwan	68	(Hsieh *et al*. 2009)
ET	Spain, Hungary, US	342	(Munoz *et al*. 1992)
ST	US	616	(Croucher *et al*. 2013)

Table 1.3. Studies in which non-overlapping associations between serotype and ET/ST have been observed among **Streptococcus pneumoniae.**

Pathogen	Common disease(s)	Antigenic Type	Metabolic Type	Details	No. isolates	Study showing frequent recombination
Eschericha coli	Gastroenteritis	O antigen (lipopolysaccharide)	ET	Good correlation between O group and ET; the majority of isolates of an ET expressed a common O antigen (Maslow *et al*. 1995)	187	(Wirth *et al*. 2006)
Haemophilus influenzae	Bacteraemia, meningitis, cellulitis, epiglottitis, septic arthritis, pneumonia	Serotype	ET	No ETs were shared among isolates of different serotypes (Musser *et al*. 1988)	2209	(Mell *et al*. 2011)
		Serotype	Clonal complex	On a minimum evolution tree of concatenated MLST nucleotide data, monophyletic groups were formed according to several serotypes (Meats *et al*. 2003)	131	
Neisseria gonorrhoeae	Gonorrhoea	11 Opa genes (opacity related protein)	ST	Distantly-related STs manifested unique opa allelic profiles. Minimal opa variation was found among isolates of the same ST (Bilek *et al*. 2009)	14	(Orourke & Stevens 1993; Bilek *et al*. 2009)
Neisseria lactamica	Asymptomatic colonisation of nasopharynx in young children	FetA	ST	Particular VRs were associated with STs (Bennett *et al*. 2009)	275	(Alber *et al*. 2001)
Neisseria meningitidis	Meningitis, septicaemia	PorA	Clonal complex	Non-overlapping PorA combinations were associated with clonal complexes (Buckee *et al*. 2008)	977	(Jolley *et al*. 2005)
		PorA, FetA	Clonal complex	Non-overlapping PorA:FetA combinations stably associated with clonal complexes (Buckee *et al*. 2010).	1054	
		fHbp (factor H binding protein)	Clonal complex	Subfamily/variant fHbp alleles clustered with particular clonal complexes (Brehony *et al*. 2009)	107	

		fHbp, NadA and NHBA	Clonal complex	Variants of fHbp, NadA and NHBA clustered with clonal complex in non-overlapping associations (Bambini *et al*. 2013)	165	
		Opa (opacity related protein)		For each of the clonal complexes examined, particular opa alleles were consistently observed at each locus (Callaghan *et al*. 2006).	77	
Pasteurella trehalosi	Systemic infections in sheep	Serotype, outer membrane protein, LPS	ET	Isolates of the same ET were generally associated with the same combination of serotype, LPS type and outer-membrane protein (OMP) type (Davies *et al*. 1997)	60	(Davies *et al*. 2002)
Salmonella enterica	Gastroenteritis	Serovar (O and H antigens)	eBGs (eBurst Groups)	Strong correlation between eBGs and serovars (except those in lineage 3) (Achtman *et al*. 2012)	4257	(Didelot *et al*. 2011)
Staphylococcus aureus	Opportunistic infections of skin	Spa types (Staphylococcal Protein A), ClfA & ClfB (clumping factor A & B)	Clonal complex	97% (98 out of 101) of spa types were associated with a single clonal complex (Murphy *et al*. 2011) .	224	(Everitt *et al*. 2014)
		Spa types	ST	ClfA and ClfB sequence variants were closely associated with clonal complex (Tavares *et al*. 2014).	182	
Streptococcus pyogenes	Opportunistic infections (including necrotizing fasciitis, streptococcal toxic shock syndrome, tonsillitis, impetigo)	Emm types (M protein)	ST	95% (208 out of 220) of STs were associated with a single emm type (McGregor *et al*. 2004)	495	(Bessen 2009)

	Emm types (M protein)	ST	97% (97/100) of STs were associated with a single emm type (Enright *et al*. 2001)	212

Table 1.4. Associations among antigenic and metabolic genes for 9 bacterial pathogens, for which frequent recombination has been documented*. Metabolic types (MTs) refer to: (i) allelic variants of metabolic genes, or (ii) sequence types (STs), clonal complexes or electrophoretic types (ETs) from MLST schemes based solely on genes coding for enzymes involved in metabolic activity. Bacteria which had been typed using MLST schemes based on non-coding DNA, such as ribosomal intergenic spacers, were not included. The antigenic type (AT) is represented by: (i) variants of single antigens which generate distinct immunogenic responses, or (ii) multi-locus variants exhibiting unique combinations of alleles at a number of antigenic loci.*

Sequence Type	Serotype	Timespan (years)	Frequency (no. isolates)
15	14	37	84
53	8	25	48
63	14	23	137
81	19F	22	80
81	23F	30	338
90	6B	28	126
113	18C	32	32
156	9V	24	136
172	19A	34	67
247	4	31	25
447	37	11	31

Table 1.5. Number of years over which combinations of serotype and ST have been recorded for* S. pneumoniae*, according to the PubMLST website*. Owing to the nature of the database, these figures represent the minimum length of time over which such isolates have existed. Data were taken from the PubMLST website (http://pubmlst.net/) on 22/07/15.*

A number of questions regarding the maintenance of intermediate population structures such as *N. meningitidis* and *S. pneumoniae* can therefore be formulated. First, the persistence of discrete lineages over time and space, in the face of frequent recombination. Second, the stable associations among many bacterial pathogens of antigenic variants with combinations of metabolic housekeeping alleles, despite the genetic distance between such loci and the high rates of recombination observed. Third, for meningococci, the global spread and long-term

persistence of lineages which are associated with invasive disease, despite the presumed transmission cost. In the following sections, the approaches used to understand and rationalise these patterns of genetic diversity, and the prevailing hypotheses and models offered for the observations among both antigenic and non-antigenic genes, shall be discussed.

1.3 Understanding the Processes Underlying Observed Patterns of Diversity

1.3.1 Epidemiological Models of Disease Dynamics

Mathematical models have become important tools for analysing the spread and control of infectious diseases, and in turn for supporting the development of management strategies. Mathematical epidemiology dates to Daniel Bernoulli's model in 1760, which was used to evaluate the effectiveness of variolation with the smallpox virus. However, it was not until the late nineteenth century that the germ theory gained acceptance, which proposed that disease was caused by specific micro-organisms. The foundations of mathematical epidemiology were laid thereafter, in the early twentieth century, when several physicians and biological scientists introduced a range of mathematical models on the time courses of epidemics. Advances in computational power allowed for the widespread use of stochastic models, including agent-based models as well as spatial models, thus providing epidemiologists with a diverse range of models with which to work. In addition to using theoretical frameworks to model the prevalence of individual strains, it is possible to introduce multiple loci into the strains modelled in a system; it is these multilocus models which I explore in my DPhil in deterministic frameworks.

Kermack and McKendrick's epidemic model (Kermack & McKendrick 1991) was developed to explain the rise and fall in the number of infected individuals in epidemics such as cholera

and the plague. It is a compartmental model, in which the host population is divided into distinct categories according to their status of the disease in question:

Susceptible (S): individuals with no immunity to the infectious agent in question, and are thus able to be infected upon exposure.

Infectious (I): individuals currently infected, who are able to transmit the infectious agent to susceptible individuals.

Recovered (R): individuals who are immune to the infection.

The population, N, is fixed, with the total population size given by: N = S + I + R. The model comprises the following differential equations:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Where γ represents the recovery rate and β represents the transmission coefficient, a measure of transmission that accounts for contact frequency and probability of infection given a close contact with an infected individual. Every individual has an equal probability as every other individual of contracting the disease; therefore in a fully susceptible population, an infected individual is able to transmit the infection to βN others per unit time. This established the threshold theory, which posits that the introduction of a few infectious individuals into a fully susceptible population can only give rise to an epidemic if the number of susceptible individuals exceeds a critical value. The basic reproductive number, R_0, is defined as the expected number of secondary cases produced by a single infectious individual in a completely

susceptible population. In order for an epidemic to occur, R_0 must exceed one, and the outbreak will become extinct if it decreases below one. In order for mass vaccination to eradicate a pathogen, the critical proportion of a population to be vaccinated is $1 - 1/R_0$. This refers to herd immunity – if a sufficient number of susceptible individuals in the population have either vaccine-induced or naturally-acquired immunity, the disease will not spread, thus protecting the susceptible individuals from infection also.

1.3.2 Population Genetics Models and Phylogenetic Approaches

Population genetics and phylogenetic approaches are also used to understand the processes underlying the diversity and evolution of pathogens, and are complementary to mathematical models of disease dynamics. Population genetics involves the study of genetic variation within populations, and examination of changes in the frequencies of genes and alleles in populations over space and time. The aim is to understand the evolutionary and epidemiological processes which give rise to the genetic diversity observed, using genetic data and realistic evolutionary models. Such tools can be used to further understand the origin and history of pathogens, the selective processes operating in populations, and the role of recombination in generating diversity (reviewed by Wilson *et al*. 2005). Coalescent approaches use mathematical models to understand the shape of genealogies through population-level processes. Phylogenetic methods, in contrast, make no assumptions about the processes which govern the shape of phylogenies. (A phylogeny is a diagram which displays the evolutionary relationships of organisms or genes). Phylogenetic approaches can be used to date the origin of organisms/genes and epidemics (if a known nucleotide or amino acid substitution rate is available), to detect recombination, and to infer information on the size and structure of historical populations.

1.3.3 Hypotheses Offered for Antigenic Diversity and Structuring

A number of explanations have been proposed to account for the generation and maintenance of antigenic diversity and structuring in bacterial pathogen populations, with support from genotypic data, population genetics approaches and/or mathematical models. The explanations summarised in this section are focused mainly around mathematical models of disease dynamics, as they form the primary focus of this thesis. Many explanations for antigenic diversity observed within pathogen populations invoke selection pressures imposed by the host immune response, which can operate at within- or between- host levels. Immunity can either be cross-acting (affecting all strains) or strain-specific (affecting only those strains of a given antigenic determinant).

It is important to note that there is also increasing evidence for a number of explanations of antigenic diversity which do not concern the actions of the immune system. Antigens have a range of diverse functions including nutrient acquisition, regulation of osmotic pressure and host cell binding. There are therefore other ecological factors which may favour the selection of antigenic variability, as any selection pressures in the pathogen's environment which favour phenotypic variability may indirectly affect antigenic variability (reviewed by Lipsitch & O'Hagan 2007).

1.3.3.1 Population-level Diversity

Simple models of strain competition, in which cross-immunity between all strains is perfect, exhibit competitive exclusion culminating in the dominance of the strain of highest transmission fitness (R_0) (Anderson & May 1991; Bremermann & Thieme 1989; Levin & Pimentel 1981). If cross-immunity is not perfect, then co-existence of strains can occur, with

the degree of cross-immunity dictating the range of differences in R_0 that the system can support (Gupta *et al*. 1994a).

Strong cross-immunity can cause pathogen populations to segregate into discrete strains characterised by non-overlapping antigenic determinants (Gupta *et al*. 1996; Gupta *et al*. 1998) (Figure 1.2). If infection with a given strain generates a variant-specific immune response, there is a degree of cross-protection conferred between strains that share antigenic variants. Therefore, if the immune response is sufficiently strong, strains that share antigenic variants are at a transmission disadvantage, leading to the segregation of the population into strains defined by unique combinations of antigenic alleles. At weak levels of immune selection, discrete strain structure does not emerge, and all strains co-exist at the same equilibrium level. At intermediate levels of immune selection, pathogens exist as a set of strains which fluctuate in either a chaotic or cyclical manner.

Support for the immune selection hypothesis of Gupta and colleagues was first observed among two variable regions (VR1 and VR2) of the outer membrane antigen PorA from a sample of disease isolates of *N. meningitidis* in the UK (Gupta *et al*. 1996) (Figure 1.3), but have been subsequently noted in a number of carried and invasive datasets worldwide (e.g. Urwin *et al*. 2004; Buckee *et al*., 2008). However, linkage disequilibrium between VR1 and VR2 of PorA may occur for a number of alternative reasons. Firstly, they are located very close together on the chromosome; thus strong genetic linkage may account for the associations. Secondly, as they are surface-exposed variable regions on the same antigenic structure, there may be epistatic interactions governing the associations observed. Subsequent analyses showing non-overlapping associations between FetA and PorA VR1 and VR2 provided additional support for the hypothesis, as the two antigens are distantly located on the chromosome (Buckee *et al*. 2010; Buckee *et al*. 2011). These antigenic associations also appear to be stable: associations between PorA VR1 and VR2 were observed for 27 years in a longitudinal carriage dataset from

the Czech Republic (Buckee *et al*. 2008), for example, with subsequent analyses extending to the FetA antigen (Buckee *et al*. 2010). Non-overlapping associations between Opa variants (antigenic adhesins required for invasion) have also been observed at the population level (Callaghan *et al*. 2008). Similar non-overlapping patterns are observed in *Streptococcus pyogenes*, in which allelic variants of two outer membrane proteins display distinct associations (Johnson *et al*. 2006). Non-overlapping antigenic variants have also been observed in non-bacterial pathogens, including *Plasmodium falciparum* (Gupta *et al*. 1994b).

Antigenic clustering among discrete strains has been observed in a number of other mathematical models. Gomes *et al*. (2002) used a model formulation in which the pathogen structure was distributed as a continuous variable (the "strain space"). The results depended on the levels of the parameter p, which determined the degree to which strains are clustered around a reference strain. At higher levels of p, most strains will offer a lot of cross-protection to a given strain. Simulations showed that, when p is negative, the population falls into stationary clusters of different strains – a result consistent with the hypothesis of Gupta and colleagues. When p is close to zero, oscillations are more likely to occur. Gog & Grenfell used a status-based formulation to model the effects of cross-immunity, in which the host variables are tracked according to any impact they are going to have on future infections by all other strains (Gog & Grenfell 2002). The results depended on the length of the infectious period, and two main behaviours were observed. Discrete clusters of pathogen strains emerge over "strain space" if the infectious period is prolonged relative to the host lifespan. If the infectious time is short relative to the host lifespan, a single cluster dominates at any given time, and the population jumps between clusters of strains.

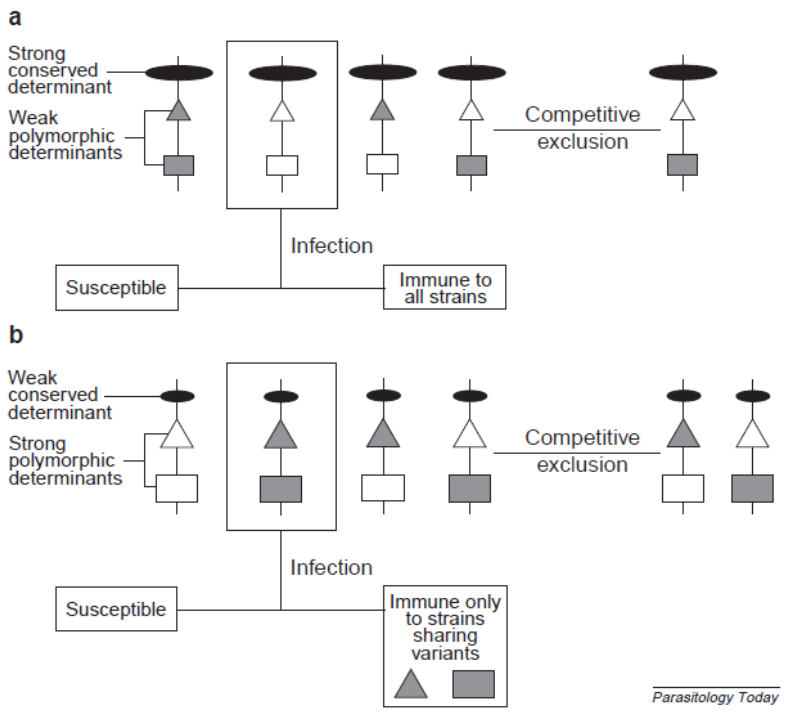


Figure 1.2. Effects of cross-protection on a range of antigenic determinants (from Gupta & Anderson, 2001). *(A) Immune selection acting on strains with strong conserved determinants leads to the emergence of a single strain through competitive exclusion.* ***(****B) Immune selection operating on strains with polymorphic antigens leads to a set of strains with non-overlapping antigenic determinants.*

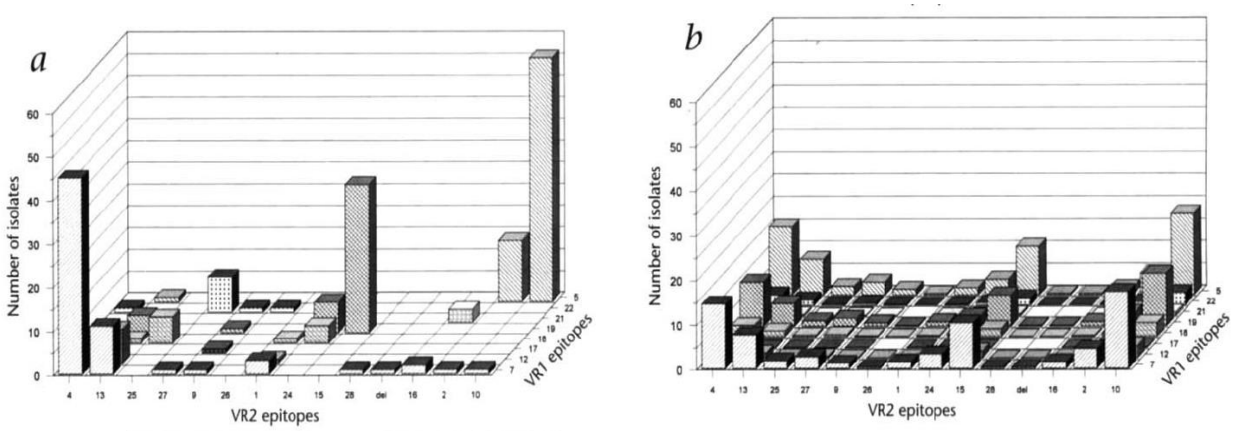


Figure 1.3. Non-overlapping patterns among meningococcal antigens (taken from Gupta et al., 1996.) *(A) Associations between the VR1 and VR2 epitopes of the porin PorA among 222* N. meningitidis *disease isolates. (B) Randomly distributed associations between VR1 and VR2 regions.*

There are a number of frameworks which invoke both strain-specific and cross-reactive immunity. Gupta & Galvani consider the case where a proportion of the host population responds principally to a variable antigen, and the remainder responds principally to a conserved antigen (Gupta & Galvani 1999). They found that discrete strain structure (with strains showing non-overlapping antigenic repertoires) was observed even with a high proportion of hosts that respond cross-reactively rather than in a variant-specific manner. Several models incorporating both strain-specific and cross-reactive immunity have been used to understand the genetic diversity of *S. pneumoniae*, and the effects of serotype-specific vaccination (Cobey & Lipsitch 2012; Flasche *et al*. 2013). These studies generally showed that increasing levels of cross-protective immunity between capsular serotypes led to a reduction in the diversity of serotypes, whereas increasing levels of serotype-specific immunity led to an increase in the number of circulating serotypes (see also Section 8.3). A balance between the two was important for re-creating the patterns of diversity observed.

1.3.3.2 Within-Host and Population-Level Diversity

Most theoretical frameworks of pathogen evolution have focused either on the evolution of within-host diversity (particularly in chronic pathogens), or population-level diversity (particularly in highly transmissible, short-lived pathogens). There are few frameworks which examine both processes (Gilchrist & Sasaki 2002).

Callaghan *et al*. (2008) used a deterministic model and genetic data to examine the effects of cross-immunity at both within-host and between-host levels. At the level of individual pathogens, immune selection is predicted to decrease within-repertoire antigenic diversity if identical alleles are able to occupy different loci. It is probable that strains which express a diverse set of antigenic variants will encounter hosts which have previously been infected with

a variant from their repertoire, and selection should therefore favour strains which express identical antigenic variants at multiple loci. This hypothesis is supported by data on the opacity-associated (Opa) outer membrane proteins of *N. meningitidis*. The Opa proteins are involved in adhesion and invasion and interact with the CEACAM family of proteins and cell saccharides. There are 3-4 *opa* gene loci in each meningococcus, the variants of which are highly diverse, with much variation occurring in two hypervariable surface-exposed regions (HV regions). Different *opa* loci within the same bacterium may express identical or diverse HV regions. Callaghan *et al*. showed found the observed number of identical HV variants within individual meningococci was significantly greater than that expected from a random distribution, among a sample of 216 carried meningococci from the Czech Republic. At a population level, combinations of HV1 and HV2 demonstrated a non-random and non-overlapping structure between strains. These observations were consistent with the results of the mathematical model: rather than acting to increase antigenic diversity both within and between strains as predicted by the actions of diversifying selection, strong cross-immunity led to an increase in between-strain diversity, but a decrease in within-strain diversity.

In Chapter 2, I extend the framework by Callaghan *et al*. (2008) to explore how within- and between-host immune selection shape the diversity and structuring of antigens in pathogens with different durations of infection and distinct life histories. I apply the model to the contrasting structure and diversity of Opa proteins observed across meningococci and gonococci.

1.3.4 Hypotheses Offered for Diversity of Non-Antigenic Markers and Maintenance of Intermediate Bacterial Population Structures

A number of theories have been offered to explain the persistence of lineages defined by non-antigenic markers in highly recombining pathogen populations. Firstly, high levels of linkage disequilibrium may result from genetic drift (Maynard Smith *et al*. 1993). According to neutral models of evolution, observed patterns of genetic variation are the result of chance events of mutation and recombination with no selective processes occurring. Fraser *et al*. (2005) constructed a mathematical model based purely on neutral mutational drift and recombination which was not sufficient to account for the population structure of *N. meningitidis, S. pneumoniae* and *S. aureus*. A neutral model which accounted for the impact of transmission in local populations – known as the neutral microepidemic model – was able to fit the relatedness distributions of data from these pathogens (Fraser *et al*. 2005).

However, Jolley *et al*. (2005) used a coalescent-based approach to demonstrate that the number of meningococcal STs generated as expected through neutral processes alone significantly underestimated the number of STs observed in a sample of meningococci from the Czech Republic in 1993. Further, Buckee *et al*. (2008) carried out simulations for the lifespan of allele combinations assuming the microepidemic clustering process adopted by Fraser *et al*. (2005). Results showed that the majority of allele combinations persisted for one year, and none persisted for more than 7 years; however, meningococcal STs can persist for decades (Urwin *et al*. 2004; Buckee *et al*. 2008). Thus, although neutral processes are likely to play an important role in the population structure of bacterial pathogens, it might be drawn from these studies that alone they may not be sufficient.

Others have proposed that selective forces play a role in the maintenance of clonal complexes. The epidemic model (Maynard Smith *et al*., 1993) posits that short-term clonal expansion can

occur within a recombining population if the founding genotypes have a selective advantage. Although this model was originally suggested to explain the electrophoretic data for *N. meningitidis*, it cannot account for the hyperinvasive nature of particular clonal complexes in meningococcal, pneumococcal and other bacterial populations, or their longevity over many years (Caugant *et al*. 1986; Musser *et al*. 1989; Enright & Spratt 1998; Brueggemann *et al.* 2003; Hanage *et al*. 2005; Wertheim *et al*. 2005).

The structuring of bacterial populations by purely mechanistic processes, as a result of differences in restriction modification systems, has been proposed by Budroni *et al*. (2011). They concluded that clades of *N. meningitidis* identified using whole genome sequencing were associated with different restriction modification systems, which prevented the exchange of DNA between clades, thus maintaining population structure despite substantial recombination (Budroni *et al*. 2011). However, the association between clades and restriction-modification systems could indeed be interpreted as a consequence of the diversification processes caused by other mechanisms, such as variation in transmissibility and immune selection, as opposed to the cause of the observed population structure.

1.3.4.1 Ecological and Immunological Competition

Although a number of theories have been put forward to explain the maintenance of strains in highly recombining populations, various aspects of the population structure of *N. meningitidis*, *S. pneumoniae* and other bacterial pathogens remain unexplained. These include the emergence of discrete lineages and their unique associations with antigenic repertoires despite high rates of recombination (Table 1.4). For *N. meningitidis*, this also includes the proclivity of certain lineages to cause invasive disease despite the severe transmission cost. A series of models developed by Buckee *et al.* (2008) are able to account for these observations, and can account

for the diversity and structuring observed among both housekeeping and antigenic alleles of meningococcal populations.

Traditionally, bacterial population structure has been investigated through studying the patterns of diversity at antigenic and/or "housekeeping" loci, as summarised above. These "housekeeping" loci, such as those encoded by the MLST loci, have important functions which are critical to the internal functioning of the bacterium, and it is often assumed that they are under strong stabilising selection. The majority of housekeeping loci encode metabolic functions. Indeed, 73% of the core meningococcal genome comprises metabolic genes (Bratcher *et al*. 2014), and all of the MLST loci comprise fragments of genes encoding enzymes involved in either transport or metabolic processes (Maiden *et al*. 1998) (Table 1.1). Buckee *et al*. proposed that clonal complexes correspond to constellations of co-adapted metabolic complexes. This encourages the emergence of lineage structure, as hybrid genotypes generated frequently by recombination may have a suboptimal combination of protein metabolic variants, which may affect the correct functioning of protein complexes, etc. In highly integrated bacterial metabolism, even small differences in amino acid sequence may affect the functioning of protein interaction networks (Noirot & Noirot-Gros 2004). A key assumption of their models was that STs vary slightly in their transmission fitness, as differences in metabolic efficiency encoded by different STs may lead to small differences in the ability to assimilate resources and transmit between hosts. According to Buckee *et al*., clonal complexes may, therefore, represent units of selection rather than entities produced by neutral processes.

Buckee *et al*. published two separate multilocus mathematical models. Their stochastic model featured strains which comprised two antigenic loci and one housekeeping locus (corresponding to the ST). It was assumed that strong ecological competition prevented co-infection by two identical STs. At high levels of cross-immunity (immunological competition), stable non-overlapping associations between antigenic types and STs were observed at small

transmission differences between STs, and fluctuating associations emerged at intermediate differences in transmission fitness between STs. The models results were consistent with the long-lived associations between clonal complex and OMP antigens of *N. meningitidis* observed in a number of studies (e.g. Caugant *et al.*, 1987; Urwin *et al*., 2004; Buckee *et al*. 2008), as previously mentioned, in addition to fluctuating associations between certain combinations of OMP antigens and clonal complex in a sample of isolates from the Czech Republic (Buckee *et al*. 2008; Buckee *et al*. 2010).

The second model by Buckee *et al*. was a deterministic framework, in which strains were defined by an ST and another locus conferring a property of excess virulence. This property had the effect of decreasing the R_0 of a strain through shortening its infectious period. Simulations showed that, at low differences in transmissibility between STs, all STs were able to co-exist and afford excess virulence. As these differences were increased, the strains of lower transmissibility were competitively excluded, until the population consisted of a few STs. Of these, only the more transmissible strains were able to afford excess virulence. The frequency distribution of STs belonging to carriage and invasive disease was simulated according to the model (at intermediate differences in fitness between STs), and was able to reproduce observed patterns of diversity among *N. meningitidis* isolates.

Although developed for the meningococcus, the results of the models by Buckee *et al*., (of associations between antigenic and metabolic alleles), are consistent with observations from other bacterial pathogens, and it could be hypothesised that strong ecological and immunological competition at metabolic and antigenic loci respectively may play a role in the maintenance of the population structure of other pathogens (discussed further in Chapter 8). The majority of the MLST loci in *S. pneumoniae* are involved in metabolic processes (Table 1.2); thus it could be postulated that the widespread associations observed between serotype and ST (Table 1.3), which can fluctuate over time (Jefferies *et al*. 2010; Feikin & Klugman

2002), may also be explained by the models by Buckee *et al*. Consistent with this hypothesis, Croucher *et al*. (2015) examined three alternative explanations for the patterns of within-serogroup switching observed within a sample of 616 genomes of *S. pneumoniae* (constraints of genetic transformation; proximity to beta-lactam resistance genes; and epistasis between the capsule and genes in the rest of the genome) and found that none may be likely to account for the distribution of capsule types observed. They did stress, however, that more detailed searches for loci that may interact epistatically with the *cps* locus are required (see also Section 8.2).

In the following sections, the epidemiology and structure of meningococcal and pneumococcal populations will be summarised, given that data from these pathogens primarily form the basis of the chapters of this thesis.

1.4 Epidemiology of *Neisseria meningitidis*

1.4.1 *Neisseria* species

The *Neisseria* genus comprises several Gram negative oxidase-positive diplococci, which colonise the mucosal surfaces of humans and a range of animal hosts. As of 14/08/2014, the PubMLST database (www.pubmlst.org) listed 23 *Neisseria* species. The majority of human *Neisseria* species colonise the oropharynx transiently as harmless commensals, such as *N. lactamica*, which is common in young children (Bennett *et al*. 2005). Two closely-related species, however, are important human pathogens: *N. meningitidis* is a primary cause of bacterial meningitis; and *N. gonorrhoeae* (the gonococcus) is responsible for the sexually transmitted infection gonorrhoea.

1.4.2 Invasive Disease by *N. gonorrhoeae*

N. gonorrhoeae contributes a substantial burden of morbidity, mortality and infertility worldwide. Gonorrhoea is treatable and curable, but no vaccine is available. *N. gonorrhoeae* primarily colonises the genito-urinary tract, in addition to the pharynx and rectum. Most infections are local, but a minority of cases (approximately 2%) lead to systemic infection - known as disseminated gonococcal infection - in which gonococci spread to the bloodstream and other parts of the body (Ross 1996). Severe cases of gonococcal infection in women can also lead to pelvic inflammatory disease, resulting in infertility, ectopic pregnancy, and chronic pelvic pain (Buchan *et al*. 1993).

N. lactamica, meningitidis and gonorrhoeae are closely related genetically, suggesting a relatively recent divergence. A study comparing seven housekeeping MLST alleles found a relatively low diversity among gonococcal isolates, suggesting that gonococci arose fairly recently from a meningococcal ancestor, by colonising the urogenital tract (Bennett *et al*. 2007). The results of a more recent study comparing the three species at the whole genome level supported this hypothesis (Bennett *et al*. 2010).

1.4.3 Meningococcal Carriage and Invasive Disease

N. meningitidis gives rise to over 500,000 cases of meningococcal disease worldwide each year, a figure exaggerated by large epidemic outbreaks (Tikhomirov *et al*. 1997). The mortality rate is 5-15% - a rate relatively unchanged since the 1930s - and an equal number of survivors develop long-term sequelae including hearing loss, mental retardation, and limb amputation (Swartz 2004). The most common presentation of meningococcal disease is meningitis, characterised by an inflammation of the protective membranes covering the brain and spinal

cord, the meninges. To cause meningitis, meningococci must traverse the epithelial cell layer of the nasopharyngeal epithelium, enter the bloodstream, and subsequently cross the blood-brain barrier and multiply in the cerebrospinal fluid. Around 20-30% of cases manifest septicaemia, and 12% may have features of both conditions (Thompson *et al*. 2006).

Despite its pathogenic potential, *N. meningitidis* evolved as a commensal of the human nasopharyngeal mucosa and is frequently carried asymptomatically. In industrialised countries, carriage rates are approximately 10% for the population as a whole (Claus *et al*. 2005), with the lowest prevalence in infants and young children and a peak in adolescents and young adults (Cartwright *et al*. 1987). Carriage prevalence in sub-Saharan Africa seem to be more variable, with carriage rates reported between 3% and 30% across a range of ages (reviewed by Trotter & Greenwood 2007). Carried populations of *N. meningitidis* exhibit high levels of genetic diversity in comparison to isolates from invasive disease (Yazdankhah *et al*. 2004; Jolley *et al*. 2000). As discussed previously, pathogenesis does not contribute to transmission among hosts, and it is not understood as to why these "accidental pathogens" occasionally cause disease (Maiden 2008). However, lineages with an increased propensity to cause disease – referred to as hyperinvasive – still persist for decades (Table 1.6) (Caugant *et al*. 1986; Urwin *et al*. 2004).

Clonal complex	MLEE grouping	Dominant serogroup	Main origin
ST-1 complex	Subgroup I/II	A	Russia, China, Africa
ST-5 complex	Subgroup III	A	Africa
ST-8 complex	Cluster A4	B, C	Europe
ST-11 complex	ET-37 complex	B, C, W	Worldwide
ST-18 complex	Cluster J1	B	Czech Republic, Poland
ST-32 complex	ET-5 complex	B	Worldwide
ST-269 complex	Lineage 4	B	Worldwide

Table 1.6. Details of meningococcal hyperinvasive clonal complexes*. (Adapted from Caugant & Maiden, 2009).*

The majority of meningococci express a polysaccharide capsule on their exterior, which are divided into 13 different types on the basis of immunogenicity reactions and biochemical composition, known as serogroups. Of 12 recognised serogroups, only six (A, B, C, W, X, Y) are responsible for the majority of invasive disease worldwide (Harrison *et al*. 2013). A minority of meningococci – referred to as capsule null – do not express a capsule. Although the vast majority of disease cases are attributed to capsulated meningococci, a small number caused by capsule null variants have been noted (Findlow *et al*. 2007).

The highest incidence of meningococcal disease occurs in sub-Saharan Africa, which can exceed 1000 cases per 100,000 during epidemics (Trotter & Greenwood 2007). The meningitis belt, extending from Senegal in the West to Ethiopia in the East, experiences periodic yet unpredictable epidemics every 6-14 years. Disease incidence increases in the dry season (from December to May), and decreases with the onset of the rains at the beginning of the wet season. It is not known why these outbreaks show such pronounced seasonality. The most popular hypothesis suggests that the extreme environmental conditions present during the dry season – including dry, dusty Harmattan winds, high temperature and low absolute humidity – enhance susceptibility to invasive disease through damage to the mucosal defences (Greenwood 1999).

The first epidemics of meningococcal disease in sub-Saharan Africa were recorded in the early 1900s (Greenwood 1999) with serogroup A representing the primary cause of disease during the 20th century. The largest serogroup A epidemic to date occurred in 1996-1997 with over 250,000 cases and a widespread distribution across the belt in 10 countries (Roberts 2008). In recent years, other serogroups have emerged as important causes of meningococcal disease in the belt. Although previously associated with small sporadic outbreaks, extensive epidemics of serogroup W in Burkina Faso in 2002 resulted in 12,000 cases, introduced via pilgrims from Mecca (Koumare *et al*. 2007). Serogroup X outbreaks have also increasingly been reported in recent years from a number of countries across the belt, including Niger, Ghana, Togo, Burkina

Faso and Kenya, but have not yet given rise to epidemics of the same scale as the serogroup A and W epidemics (Xie *et al*. 2013).

Disease incidence in Europe, the Americas and Australasia is much lower, ranging from under 1 case per 100,000 in Europe, to 0.5-1.5 cases per 100,000 in the United States. Serogroups B, C, W and Y predominate in industrialised countries (Halperin *et al*. 2012). Serogroup A outbreaks were recorded in the early part of the twentieth century in the United States and United Kingdom, but have not been observed since the Second World War (Olyhoek *et al*. 1987). Serogroup B is hyperendemic in many regions and is associated with sporadic disease, and serogroup C is associated with epidemics, frequently among adolescents (Caugant & Maiden 2009). Serogroup Y has emerged more recently as an important cause of disease over the past ten years. In the United States for example, serogroup Y accounted for only 2% of disease cases in 1989-1991 but now account for a third of disease isolates (Cohn *et al*. 2010). Although previously rare in Europe, recent increases in Y disease have also been reported from a number of countries, including Sweden, Finland, Norway and the UK (Broeker *et al*. 2012).

The epidemiology of meningococcal disease in Asia is less understood owing to lower levels of surveillance data available. In China, it is believed that epidemic periods, caused by serogroup A, occur periodically every 8-10 years, in addition to low levels of endemic disease caused by serogroups B and C. Incidence was reported to reach 500 cases per 100,000 and 0.2-1 cases per 100,000 during epidemic and endemic periods respectively (Zhang *et al*. 2008). A recent review of meningococcal disease in Asia suggested that the predominant serogroup causing epidemics in most countries was serogroup A (Vyse *et al*. 2011).

1.4.4 Virulence Factors

Virulence in this thesis is defined as the ability of a pathogen to cause disease. A virulence factors is defined as any agent (e.g. molecules, structures) of a pathogen which is necessary to cause disease. The polysaccharide capsule is often regarded as an important virulence factor in *N. meningitidis*. The capsule protects against phagocytosis and complement-mediated lysis, and only encapsulated meningococci are able to cause invasive disease – although a few exceptions have been reported (Findlow *et al*. 2007). However, the capsule is by no means sufficient to cause invasive disease, as the vast majority of encapsulated meningococci are avirulent.

No studies to date have found the same genetic determinants consistently associated with invasive disease (Schoen *et al*., 2014). A number of virulence factors have been shown to be necessary for colonisation and/or disease, including several adhesins (Johswich *et al*. 2013) and certain lipooligosaccharides (Kahler & Stephens 1998). However, such virulence factors are also present in commensal strains, and there is a distinct lack of a gene repertoire which definitively distinguish invasive meningococci from carried meningococci. The ability to cause disease is still thought to be polygenic and dependent on combinations of genes that are also present in less pathogenic meningococci.

1.4.4.1 Role of Metabolism in Virulence

Most bacterial pathogens encounter a variety of different environments within the host during infection, and must adapt accordingly to make use of alternative nutrient sources: a phenomenon which has been termed "nutritional virulence" (Abu Kwaik & Bumann 2013). Indeed, there is an increasing body of support that metabolism genes play important roles in

virulence in *N. meningitidis*. The first functional genomic study which demonstrated the importance of metabolism in pathogenesis was undertaken by Sun *et al*., in which they used signature-tagged mutagenesis to identify 73 genes essential for bacteraemia in an infant rat model (Sun *et al*. 2000). Of these, approximately half encoded enzymes involved in metabolism and nutrient transport. Subsequent *ex vivo* transcriptomic analyses showed that a number of metabolic genes were upregulated among meningococci grown in blood (Echenique-Rivera *et al*. 2011), including those involved in energy metabolism, transport and binding, amino acid biosynthesis, regulatory functions, cellular processes and cell envelope synthesis. Schoen and colleagues compared the sets of genes which were differentially expressed between meningococcal cells adhering to human epithelial cells and cells grown in blood from a number of studies, and found several differences in the metabolic profiles of colonising and invading cells (Schoen *et al*. 2014). Among the colonisation-associated genes, there was a slightly higher proportion of genes involved in carbohydrate transport and metabolism, whereas a larger proportion of genes involved in ion transport and metabolism, energy production and conversion, and amino acid transport and metabolism featured in the invasion-associated genes.

1.4.5 Vaccines

The capsule acts as a primary target for mucosal and humoral immunity, and several conjugate vaccines targeting serogroups A, C, W and Y have been developed. Conjugation of the bacterial polysaccharide to a protein carrier induces a T-cell-dependent immune response, which confers increased immunogenicity among infants and prolonged duration of protection (Kelly *et al*. 2006). Conjugate vaccines provide protection against carriage as well as invasive disease, therefore preventing transmission and promoting herd immunity. The MenC conjugate

vaccine, for example, introduced in the UK in 1999, resulted in a significant decrease in the carriage of serogroup C meningococci (Maiden *et al*. 2008). Low levels of carriage of Serogroup A have also been recorded following the introduction of the MenAfriVac conjugate vaccine in Chad and Burkina Faso (Daugla *et al*. 2014; Kristiansen *et al*. 2014).

Unfortunately, capsule-based vaccines against serogroup B may not be effective, owing to concerns of inducing autoimmune responses (Jodar *et al*. 2002). An alternative to capsule-based vaccines entails surface-exposed OMPs, which are incorporated in or attached to the outer membrane. There are a number of OMPs which have been used in vaccines, including the porin PorA and the iron-regulated receptor FetA used in typing schemes (Van der Ley *et al*. 1991;Thompson *et al*. 2003). OMPs can be incorporated into vaccines through outer membrane vesicles (OMV), as recombinant fusion proteins, or alone. A number of OMV vaccines have been designed and implemented to address serogroup B outbreaks in New Zealand, Cuba and Norway, each containing outer membrane antigens specific to the outbreak strain (Bjune *et al*. 1991; Oster *et al*. 2005; Sierra *et al*. 1991). The recently licensed BEXSERO vaccine, developed by Novartis, contains three antigens in the form of recombination proteins, and PorA as an OMV component (Gossger *et al*. 2012). The formulation comprises: Neisseria heparin binding antigen (NHBA) conjugated to GNA1030; factor H binding protein (fHbp) conjugated to GNA2091; and the single antigen Neisserial adhesin A (NadA).

However, the large amount of diversity manifested at these antigenic loci – generated through extensive recombination – makes the design of OMV vaccines difficult as the coverage provided by using only one set of antigenic alleles may be low. Further, the extent of cross-protection provided by these antigens varies (Bambini *et al*. 2013): antibodies against FHBP are highly variant-specific; cross-protection has been observed for NadA against some variants but not others; and relatively high cross-protection has been observed for NHBA. There are also concerns over the long-term immunogenicity of these vaccines, and whether they provide

herd immunity against serogroup B strains (Gossger *et al*. 2012; Snape *et al*. 2013; Read *et al*. 2014).

1.5 Epidemiology of *Streptococcus pneumoniae*

1.5.1 *Streptococcus* species

Streptococci are a group of Gram Positive spheroidal bacteria which characteristically group together in chains, resembling a string of beads ("streptococcus" means "twisted berry" in Greek). The genus contains a variety of species, some of which are carried asymptomatically in humans and animals, and others of which cause invasive disease.

A number of classification schemes have been used over the last few decades to distinguish between Streptococcal groupings (Facklam 2002). Haemolytic reactions on blood agar medium are frequently used to classify such groupings, and can be divided into alpha, beta and gamma haemolysis (Shottmuller, 1903). Alpha haemolysis is characterised by the development of a green halo around colonies, resulting from the incomplete haemolysis of the surrounding red blood cells. Streptococci belonging to this group are commonly referred to as "Viridans" Streptococci, which can be divided into five groups according to 16s rRNA sequencing, of which *S. pneumoniae* belongs to the Mitis group (Kawamura *et al*. 1995). Beta haemolysis is characterised by a zone of complete lysis of red cells in the immediate vicinity of colonies. The two primary streptococci exhibiting beta haemolysis are *Streptococcus pyogenes* (Group A Streptococci) and *Streptococcus agalactiae* (Group B Streptococci). *S. pyogenes* can give rise to mild infections, such as impetigo, but also a number of severe infections, including Scarlet Fever, rheumatic fever, "strep throat" and toxic shock syndrome. *S. agalactiae* can cause infections of the bladder and uterus in pregnant women and is the most common cause of

neonatal sepsis. The third type of haemolytic reaction is gamma haemolysis, in which strains exhibit no haemolysis on blood agar. Species from the genus *Enterococcus* form the majority of this group, which are predominantly commensal but can give rise to opportunistic disease.

1.5.2 Pneumococcal Carriage and Invasive Disease

S. pneumoniae represents the leading cause of pneumonia in all ages, and also gives rise to meningitis and septicaemia. Invasive pneumococcal disease (IPD) gives rise to an estimated 1.6 million deaths each year, with the highest disease rates occurring in infants, and an increasing incidence of IPD in over 65 year olds (Sleeman *et al*. 2001). Individuals with chronic cardiopulmonary disease and immunocompromised individuals of all ages are also at increased risk of infection. As well as invasive disease, *S. pneumoniae* is responsible for a number of non-invasive infections, including otitis media, sinusitis and bronchitis.

Akin to the meningococcus, most pneumococci inhabit the nasopharynx as harmless commensals and are transmitted asymptomatically between hosts in water droplets. Carriage is highest in the first few years of life and declines thereafter, with rates between 30 and 62% described for children under 2 years (Bogaert *et al*. 2004). MLST studies have shown than pneumococcal carriage populations are more genetically diverse than those isolated from invasive disease (Robinson *et al*. 2001). There is also a significant inverse relationship between invasive disease and carriage prevalence - suggesting that the most invasive serotypes are least common in carriage, and the most frequently carried serotypes are less likely to cause invasive disease (Brueggemann *et al*. 2004).

Most pneumococci express a polysaccharide capsule, which is an important virulence factor: acapsulate pneumococci are rarely able to cause invasive disease in humans, and are less able

to initiate pathogenesis in animal models (Mitchell & Mitchell 2010). Over 90 capsule types, known as serotypes, have been described on the basis of immunogenicity reactions and chemical differences in capsular polysaccharides (which can be further grouped into 46 serogroups owing to cross-reactivity between certain serotypes). However, only ten or so serotypes are responsible for the majority of invasive disease worldwide. A review of more than 70 studies concluded that approximately ten serogroups were responsible for most disease in each continent, with serogroups 1, 4, 5, 6, 7, 9, 14, 18, 19 and 23 ranking highly in each region (Hausdorff *et al*. 2000). The serotypes accounting for disease in young children compared to older children and adults were similar, with serotype 3 accounting for increased disease in the latter group. A more recent review on paediatric pneumococcal disease alone, which made use of more than 140 studies, concluded that seven serotypes (1, 5, 6A, 6B, 14, 19F, and 23F) were the most common globally, corresponding to the seven most common in both Africa and Asia, and accounted for 58%–66% of IPD in every region (Johnson *et al*. 2010). However, although there were similarities in patterns across regions - for example serotype 14 was the most common across all regions - the proportion of disease caused by the seven most common serotypes varied across and within regions. Historically, it was believed that serotypes 1 and 5 were particular to developing countries; however, serotype 1 was an important cause of disease in Europe as well as Asia and Africa, contravening this view. It should be made clear that the serotype distributions reported in these studies were based on data before the introduction of pneumococcal conjugate vaccines; the post-vaccine serotype epidemiology is markedly different in many countries (reviewed in Weinberger *et al*., 2011).

1.5.3 Virulence Factors and Determinants of Invasive Disease Potential

A number of pneumococcal components necessary for causing invasive disease have been identified (Table 1.7). However, defining a list of genes present in all disease strains and absent from carriage strains has proved unsuccessful. As for the meningococcus, the factors necessary for pathogenesis are usually present in carriage strains also. Many virulence factors do, however, show extensive variation between strains, such as the capsule.

The capsule can inhibit several aspects of host immunity, including neutrophil extracellular traps and both complement-dependent and complement-independent neutrophil phagocytosis (Hyams *et al*. 2010; Wartha *et al*. 2007). Capsular switch studies – carried out in animal models and also at the genomic level – have demonstrated a strong link between serotype and virulence (Briles *et al*. 1992; Hu *et al*. 2012; Kelly *et al*. 1994), as have retrospective meta-analysis studies (Weinberger *et al*. 2010). A possible mechanistic link between polysaccharide biochemical structure and differences in invasive disease potential and carriage behaviour was suggested by Weinberger and colleagues (2009). They hypothesised that the structure of a given polysaccharide determines the degree of encapsulation, according to how energetically expensive a given polysaccharide is to produce. They found using that more heavily encapsulated pneumococci were more resistant to neutrophil-mediated killing, and more prevalent in carriage. Another mechanism for differences in virulence was suggested by (Hyams *et al*. 2013), who found using 20 capsule switch variants of strain TIGR4 that differences in the alternative pathway inhibitor Factor H binding negatively correlated with differences in Factor B binding, C3b binding and sensitivity to complement. This suggests that differences in the binding of Factor H could cause variations in complement resistance and invasive potential between capsular serotypes.

Virulence Factor	Role in virulence	Reference
Capsule	Prevents phagocytosis	(Yother 2011)
Pneumolysin	Forms membrane pores in host cells that lead to haemolysis; inhibit complement deposition	(Lock *et al*. 1996)
Hyaluronidase	Degrades hyaluronic acid	(Berry *et al*. 1994)
Neuraminidase A	Cleaves N-acetylneuraminic acid; unmasks binding sites	(Manco *et al*. 2006)
Serine protease	Unknown	(Bethe *et al*. 2001)
Autolysin	Bacterial autolysis; degradation products released induced inflammation	(Berry *et al*. 1989)
Pneumococcal surface protein A (PspA)	Interferes with the complement system. Binds lactoferrin and apolactoferrin.	(Ren *et al*. 2003)
Pneumococcal surface protein C (PspC)	Binds to complement regulatory Factor H, and to the human secretory IgA	(Zhang *et al*. 2000)
Pneumococcal surface adhesin A (PsaA)	Adheres to E-cadherin	(Anderton *et al*. 2007)
Pneumococcal adherence and virulence factor A	Binds fibronectin; prevents phagocytosis	(Holmes *et al*. 2001)

***Table 1.7. Major virulence factors identified in* S. pneumoniae.** (*Adapted from Nieto et al. 2013).*

The importance of pneumococcal serotype relative to genotype in determining the invasive disease potential of a strain has been the subject of debate for many years. A number of studies have calculated the empirical odds ratio of the relative probability of a given serotype or ST causing invasive disease (as opposed to being carried). In studies in both Oxford and Finland, there were no statistically significant differences in the invasive potential of different STs within the same serotype (Brueggemann *et al*. 2003; Hanage *et al*. 2005), whereas a study in Sweden did find such differences in OR between STs belonging to serotype 14 (Sandgren *et al*. 2004). It should be noted, however, that the majority of the invasive isolates analysed by Sandgren and colleagues were sampled from adults (257 out of 273), whereas the 246 carriage

isolates were obtained solely from infants (aged between 1 and 6 years); their results should therefore be interpreted with caution. The finding by Sandgren *et al*. was supported by a study in Israel which used mouse models to show that inoculation with pneumococci from two different serotypes (14 and 9V) but similar genetic backgrounds (resulting from a capsular switch) led to statistically insignificant differences in mortality rates whereas inoculation of a serotype 14 strain with a highly divergent PFGE (pulsed-field gel electrophoresis) pattern did not result in mortality (Nebenzahl *et al*. 2004). Serotype-independent differences in resistance to complement were also found by Hyams and colleagues (Hyams *et al*. 2011).

1.5.3.1 Role of Metabolism in Virulence

As with *N. meningitidis*, there are a number of genes which are correlated with differences in disease outcome and involved with metabolism and transport. Increasing attention has been given in recent years to the potential importance of metabolism genes in pathogenesis, particularly carbohydrate catabolism. A number of genome-wide screens of pneumococcal pathogenesis have identified genes involved in carbohydrate catabolism as virulence factors important in invasive disease (Table 1.8).

The concentration of free sugars in the normal human airway is low (Philips *et al*. 2003), but the human airway – including the epithelial cell surface, the mucin layer, secreted molecules, immune cells, and many of the bacterial flora – is heavily glycosylated. Pneumococci are able to synthesise maltodextrins, sucrose and glycans from the breakdown of host glycoproteins through a range of glycosidases. Complex carbohydrate metabolism has the potential to affect pneumococcal colonisation and pathogenesis in a number of ways, including acquisition of nutrients for growth, revealing receptors for adherence to host cells, and interfering with host immunity proteins (reviewed by Buckwalter & King 2012). Several studies suggest that

receptors for pneumococcal adherence on airway epithelial cells are sialylated and are revealed by neuraminidase activity (neuraminidases catalyse the hydrolysis of terminal sialic acid residues). King and colleagues, for example, demonstrated decreased adherence to human epithelial cells in nanA (neuraminidase A) knockout mutants, which was then restored with the addition of neuraminidase (King *et al*. 2006). There is also evidence suggesting a strong link between catabolite control protein A and virulence from a number of studies. CcpA represses the expression of genes involved in complex carbohydrate utilisation in the presence of a readily metabolisable carbohydrate such as glucose. Iyer *et al*. (2005) showed that Ccp-A inactivated strains show marked decreased in virulence in mouse models.

Gene	**Putative function**	**Reference**
nanA	Hydrolysis of N-acetylneuramic acid	(Polissi *et al*. 1998)
msmK	Sugar transport	
aga	Alpha-galactosidase	(Lau *et al*. 2001)
fco	Alpha-fucosidase	
manL	Mannose transporter	
strH	Beta-N-acetylglucosaminidase	(Hava & Camilli 2002)
pulA	Pullulanase	
bgaA	Beta-galactosidase	
glgB	Glycogen biosynthesis	
lacA	Lactose catabolism	
amy	Alpha-amylase	
malX	Maltodextrin-binding protein	
msmK	Sugar transport	
manN	Mannose transporte	(Orihuela *et al*. 2004)
lacA/lacB	Lactose metabolism	
malA	Maltose metabolism	
fba	Fructose-biphosphate aldolase	

Table 1.8 Examples of pneumococcal carbohydrate catabolism genes shown to be important for invasive disease, identified by genome-wide analyses*. (Adapted from Shelburne et al., 2008).*

Aside from carbohydrate metabolism, Piet and colleagues demonstrated that arginine synthesis genes also play a role in virulence (Piet *et al*. 2014). The *arg* locus is a variable region not

present in all pneumococcal strains, and was associated particularly with serotype 7F. Mouse models showed that pneumococci without these genes were severely attenuated in growth or cleared from lung, blood and CSF. Interestingly, arginine synthesis genes have been shown to be important for pathogenesis in other bacterial pathogens, such as *Mycobacterium tuberculosis* (Gordhan *et al*. 2002).

Studies of gene expression have also found metabolic genes upregulated at different stages of pathogenesis. (Orihuela *et al*. 2004) compared the gene expression of pneumococci in infected blood, cerebrospinal fluid, and adhering to epithelial cells, in a microarray study. A number of metabolic genes (including those pertaining to energy metabolism, amino acid biosynthesis, fatty acid metabolism, cofactor metabolism and nucleotide metabolism) were significantly upregulated in the bloodstream and cerebrospinal fluid.

1.5.4 Vaccines

The first licensed pneumococcal vaccine, targeting 14 serotypes, was licensed in the United States in 1977, and later replaced by the 23-valent vaccine PPSV23 in 1983. Both were polysaccharide vaccines, comprising purified capsular polysaccharide antigen. Although effective at preventing disease in the short-term, polysaccharide vaccines are not able to prevent carriage, do not induce effective immune responses in infants, and provide short-term immunogenicity (Pollard *et al*. 2009). As discussed in Section 1.4, the development of conjugate vaccines overcame these shortfalls: attached to a carrier protein, the polysaccharide antigen is able to induce a T cell dependent immune response, thus providing long-lasting immunity in infants and adults, and protecting against carriage (Ota *et al*. 2011). The first pneumococcal conjugate vaccine to be licensed, and introduced in a number of countries worldwide, was the heptavalent pneumococcal conjugate vaccine (PCV7).

1.5.4.1 Introduction and Coverage of the Heptavalent Pneumococcal Conjugate Vaccine

PCV7 includes the capsular polysaccharide of 7 serotypes (4,6B, 9V, 14, 18C, 19F, 23F), each coupled to a nontoxic variant of diphtheria toxin, CRM197. PCV7 was first licensed in the United States in 2000 and Europe in 2001. By March 2010, the vaccine was licensed in more than 90 of the 193 member states of the WHO and was included in the routine or national immunisation programs in more than 40 of these countries (McIntosh & Reinert 2011). Vaccine uptake varied extensively, from 92.6% in the US to 15.9% in Taiwan (among children aged 19-35 months for ≥3 doses). The serotypes included in the vaccine were selected on the basis of their rank order of disease incidence in the US. Given that serotypical epidemiology varies extensively across geographic regions, vaccine coverage is not uniform among and within countries: although approximately 90% of disease-causing serotypes are covered in the US, only ~45% are covered in a number of developing countries (Hausdorff *et al*. 2000).

1.5.4.2 Changes in Disease Incidence and Observations of Serotype Replacement

In the majority of countries in which PCV7 has been implemented, the incidence of invasive disease caused by serotypes included in PCV7 has significantly declined. There is evidence suggesting, however, that disease incidence caused by serotypes not included in the vaccine has also increased since the introduction of the vaccine (non-vaccine types, NVTs) (reviewed by Weinberger *et al*. 2011). Concomitant increases have also been observed in carriage of NVTs in many countries, to a greater extent than in disease (Hanage *et al*. 2010). This phenomenon of serotype replacement has the potential to attenuate the benefits of vaccine introduction. Although the net incidence of IPD has decreased following vaccination in the

majority of countries, the increase in disease caused by NVTs in some countries has outweighed the decline in disease caused by VTs, resulting in a net increase in invasive disease. A study in Barcelona, for example, demonstrated a 58% increase in the overall incidence of IPD among children aged <2 years from 1997 to 2006 (Munoz-Almagro *et al*. 2008). Furthermore, a study of Aboriginal people in Western Australia demonstrated a significant net increase in IPD in aboriginal young adults driven by an increase in NVTs (Lehmann *et al*. 2010). Evidence for serotype replacement has arisen in a number of countries on a global scale, including Norway, France, Portugal, Germany, Spain, the United States and Canada. To overcome this, further conjugate vaccines targeting 10 and 13 serotypes were licensed in 2008 and 2009 respectively – but won't necessarily prevent subsequent vaccine replacement in the future.

There are a number of caveats with interpreting the increase in NVTs following mass vaccination, however. It is difficult to attribute changes in epidemiology directly to vaccination; indeed, fluctuations in the incidence of pneumococcal serotypes have been observed naturally in the absence of pneumococcal vaccines (Jefferies *et al*. 2010; Feikin & Klugman 2002). Further, levels of vaccine coverage, rate of uptake and regional differences must be considered when interpreting the effects of vaccination. Spain, for example had relatively low initial vaccine coverage and some regional differences in uptake (Weinberger *et al*., 2011). Increases in reported prevalence of NVTs may also be due partly to the phenomenon of unmasking (Lipsitch 2001). Typical isolation methods involve the assessment of only one colony, and the subsequent identification of one serotype (which is, on average, the most common); thus, other less prevalent strains may have been present in a host but missed. Removing the common vaccine-type serotypes from the population therefore increases the likelihood of these less common serotypes being detected.

Weinberger *et al* (2011) argue that, for the increase in NVT prevalence to *not* have been caused by conjugate vaccination, one of two unlikely scenarios must hold. Firstly, the increase is due entirely to unmasking. Such an explanation is improbable as it requires that all hosts carrying a vaccine-type before vaccination are colonised by both vaccine and non-vaccine serotypes. The second scenario is that the invasiveness ratio (of cases to carriers) has declined since vaccination. This is also improbable, as there is no plausible mechanism by which vaccination may decrease invasiveness, and it appears that invasiveness ratios of serotypes generally remain consistent over years and countries (Brueggemann *et al*. 2004). Weinberger *et al*. also argue that increases in NVTs are unlikely to be due to chance because the vaccine was introduced at different times to different countries.

1.5.4.3 Protein-based Vaccines

To avoid population-level replacement by non-vaccine serotypes, vaccines should in theory contain immunogenic components able to offer broad protection across all strains. Several vaccines based on conserved surface-expressed non-capsular proteins have been in development for several years. There are a number of protein vaccine candidates including pneumolysin, PspA (Pneumococcal Surface Protein A), PcpA (pneumococcal choline-binding protein) and PsaA (Pneumococcal surface antigen), and a variety of options for incorporating proteins into pneumococcal vaccines, each of which has a number of advantages and drawbacks (reviewed by Ginsburg *et al*. 2012). Protein-only vaccines are potentially less expensive to manufacture than conjugate vaccines; however, the regulatory pathways as well as correlates of protection are not known. A major concern with stand-alone protein vaccines is that they may not be as efficacious as conjugate vaccines, based on experience with preclinical animal models, and it is not known whether such vaccines will result in herd immunity. This is because the polysaccharide capsule may cover the pneumococcal proteins to

such an extent that antibodies are not able to opsonise effectively. Other potential approaches include adding proteins to conjugate vaccines, and also using proteins as carriers for conjugate vaccines.

1.6 Whole Genome Sequencing

The models of bacterial population structure introduced in the preceding sections were developed and conceptualised primarily using typing data from a small number of loci. Built on the principles of Sanger sequencing originally developed in the 1960s, a number of whole genome sequencing technologies have advanced in recent years (Table 1.9). The resulting accumulation of large amounts of genome sequence data requires the development of tools and methods with which to analyse and interpret such data. In the following sections, the SNP- and allele- based approaches to genomic analyses are briefly summarised, and subsequently the insights into the biology and population structure of bacterial pathogens from whole genome sequence data are considered.

SNP-based methods are commonly used to characterise genetic variants within bacterial populations, in which single nucleotide differences are identified among isolates. This involves aligning either assembled sequences to, or mapping short-reads against, a reference genome. Such methods have been used to explore the evolution and phylogenetics of single-clone pathogens at high resolution (e.g. Grad *et al*. 2012; Bryant *et al*. 2013), but may be more problematic for large numbers of highly diverse and recombinogenic genomes, and are limited by the requirement of a whole genome alignment or reference sequence.

Company	Roche GS FLX	Ion Torrent Systems Inc. (Life Technologies)	Illumina – Sollexa	Pacific Biosciences	Oxford Nanopore Technologies
Sequencing Platform	GS FLX Titanium, GS Junior (454)	Ion Torrent	HiSeq 2000	SMRT	MinION™, GridION™
Method of feature generation	Bead-based/emulsion PCR	Bead-based/emulsion PCR	Isothermal bridge amplification on flow cell surface	Single molecule real time sequencing by synthesis	Patterns in ionic current from analyte interactions with the nanopore
Chemistry	Pyrosequencing	Pyrosequencing	Reversible dye terminators	Phospho-linked fluorescent Nucleotides	Nanopore sensing
Raw accuracy	99.5%	98%	98.5%	Single read 87%; consensus 99.9%	96%
Sequencing run time	2 hours	2 hours	2-5 days	<1 hour	No fixed run time
Read lengths	400 bases	Up to 400 bases	36 bases	>1000 bases	>10,000 bases
Cost per Mb	~$60	$1	~$2	-	-

Table 1.9. High-throughput next-generation sequencing platforms*. Information from Shendure & Ji (2008), Pareek et al. (2011), Chin et al. (2013) and www.nanoporetech.com.*

Akin to MLST of seven housekeeping genes, genomic allele-based methods utilise alleles as the unit of comparison as opposed to nucleotide sequences. As explained in Section 1.2, allele-based methods are able to account for high rates of recombination in pathogen populations. They are also scalable to the number of loci of interest. Ribosomal multi locus sequence typing (rMLST), for example, indexes the variation in the protein-encoding ribosomal genes (Jolley *et al*. 2012). Based on similar principles to those of conventional seven-locus MLST

approaches, the ribosomal genes are under stabilising selection. In contrast to MLST however, which does not always provide sufficient resolution to resolve relationships among very closely-related bacteria, rMLST is able to provide characterisation at higher resolution. Given that the ribosomal genes are universally present in all bacteria, rMLST can also be used to characterise relationships between phylogenetically distant bacterial species.

Typing resolution can be increased further by analysing a larger number of loci in the genome. Whole Genome MLST (wgMLST) entails the characterisation of allelic diversity of every coding sequence in the genome using a gene-by-gene approach (Maiden *et al*. 2013). The Genome Comparator Tool allows users to compare the allelic diversity at every coding sequence present in a selected reference genome (Jolley & Maiden 2010).

1.6.1 Insights from Whole Genome Sequencing

Findings from genomic studies have led to an improved understanding of bacterial genome dynamism and structure. Bacterial genomes are considered to comprise two parts, the core genome and accessory genome, which together make up the pan-genome. The core genome is usually defined as the set of genes which are present in all isolates of a given grouping (e.g. species, genus), whereas genes in the accessory genome may not be present in all strains.

Accessory genes are thus considered dispensable as they can be deleted from the genome, and are often involved in interactions with the external environment (e.g. colonisation factors, virulence factors) or supplementary biochemical pathways. Accessory genes are not always 'native' to the genome and may have been imported from unrelated lineages by lateral gene transfer. Indeed, many accessory genes are associated with mobile genetic elements (such as phages or transposons) which allow them to propagate between genomes. They can also cluster

together to form genomic islands, which may be recognised by an atypical G+C content. Genomic islands which feature a large number of pathogenicity associated genes are often referred to as pathogenicity islands (Lee 1996).

The core genome is thought to encode essential genes involved in housekeeping functions and information processing genes which enable the organism to survive and replicate. The results of a recent study by Bratcher and colleagues for example, identified 1760 coding sequences which were present in all 108 meningococcal genomes in a representative dataset (Bratcher *et al*. 2014). For defining the core genome, however, the threshold was relaxed to 95% in order to accommodate for missing genes in the draft genomes in the dataset, thus identifying 1605 core genes. When the functions of the core genes were considered, the majority coded for metabolic functions (73%), followed by genetic information processing (20%), and environmental information processing (6%).

Although the core genomes of bacterial pathogens are primarily composed of metabolic genes, there are an increasing number of studies which show that several metabolic and nutrient uptake genes are also part of the variable accessory genome. For example, the construction of pangenome-scale metabolic models for six *E. coli* strains showed a large number of strain-specific differences in the total number and functional classification of metabolic coding sequences and metabolic reactions in each organism (Baumler *et al*. 2011). The total number of metabolic reactions in each strain ranged from 2362 to 2428, and the metabolic pangenome – constructed from the interrogation of 16 finished genomes – contained 79 ORFs and 32 reactions not present in the core metabolic genome. Whole genome functional studies have also shown variation in the metabolic genes of *N. meningitidis*. Joseph and colleagues analysed 29 isolates using comparative genome hybridisation using microarrays (mCGH), and found that 40% of the core genes were involved in intragenic recombination, primarily among

metabolic genes. Indeed, almost all metabolic pathways were affected by recombination, including those for energy conversion (Joseph *et al*. 2011).

1.7 Outline and Aims of DPhil Project

Observations that metabolic genes comprise the majority of coding sequences in bacterial genomes, and also contribute significantly to the diversity observed in populations, suggest that the metabolic genes may be important factors in the evolution of bacterial pathogens. The role of the metabolic genes has also received more attention in recent years owing to an accumulation of evidence for the importance of metabolic genes in virulence and pathogenesis. With the current rate of accumulation of whole genome sequence data, the increasing number of functional genomics studies, and growing interest in metabolic modelling (Baart *et al*. 2007; Baart *et al*. 2008), it is likely in the next 5-10 years that we will gain an extensive understanding of the roles of metabolism in bacterial pathogen virulence, transmission and competition, as well as greater insight into the diversity observed across the metabolome of different strains. Large increases in genomic and functional data may require analytical frameworks in order to understand the functions and diversity observed. Using multilocus mathematical models to divide strains into different traits and assigning them behaviours can be expanded for several different types of genes in the genome, including virulence-associated genes and genes conferring antibiotic resistance, which can in turn be tested at the whole genome level. Such models may therefore act as analytical platforms on which to base and interpret the accumulating data acquired from whole genome studies.

The models of Buckee *et al*. (2008) form a novel hypothesis for the explanation of meningococcal and pneumococcal diversity and structuring. However, the predictions of the

framework are currently supported by studies which include only a small number of loci. Given that the majority of MLST genes in *S. pneumoniae*, *N. meningitidis* and other bacterial pathogens encode enzymes involved in metabolic processes, one could hypothesise that the predictions of the models of Buckee *et al*. (2008) for STs may be extended to other metabolic loci in the genome, and therefore that non-overlapping associations may be observed between antigenic alleles and alleles at metabolic loci across the genome (corresponding to a metabolic type or profile).

In this project, I aim to further our understanding of the processes maintaining bacterial population structure and diversity using deterministic multilocus mathematical models combining multiple traits, complemented with analyses of genotypic and whole genome data. The model and genotypic analyses in Chapter 2 build on previous theoretical frameworks of antigenic diversity in multilocus systems, and includes only antigenic loci. The models and genotypic/genomic analyses in Chapters 3, 4, 5 and 6 incorporate metabolic loci, and subsequently, virulence-associated loci are added to the model framework in Chapter 7.

In Chapter 2, the conflicting effects of within- and between- host immune selection on antigenic diversity are investigated, and applied to genotypic analysis of meningococcal and gonococcal Opa repertoires. In Chapter 3, I explore the effects of relaxing immunological competition at antigenic loci and ecological competition at metabolic loci in deterministic analogues of the stochastic model by Buckee *et al*. (2008), and examine the extent to which non-overlapping associations between antigenic types and metabolic types are maintained at varying strengths of competition. In Chapter 4, I use an extensive, public dataset of *N. meningitidis* isolates to investigate if the prediction of non-overlapping associations between antigenic and metabolic alleles is upheld across 3460 isolates. In Chapter 5, I present an analysis of the metabolic genes of 616 *S. pneumoniae* whole genomes to investigate if the model predictions of associations between antigenic and metabolic type are upheld at a whole genome level. Subsequently, I

incorporate virulence-associated alleles into the model framework in Chapter 6, and investigate the effects of strain-targeted vaccination in a model comprising antigenic, metabolic and virulence-associated genes, in the context of pneumococcal conjugate vaccination. In Chapter 7, I present a case study to demonstrate how insights from the multilocus strain models can be applied to inform our interpretation of genomic data, in which a succession of pandemic meningococcal strains belonging to Serogroup A is analysed using the gene-by-gene approach.

Chapter 2.

The effects of within- and between-host immune selection on the *Neisseria* Opa repertoire

2.1 Public Dissemination of Work

These findings were published in the article:

Eleanor R. Watkins, Yonatan H. Grad, Sunetra Gupta & Caroline O. Buckee (2014) Contrasting within- and between-host immune selection shapes *Neisseria* Opa repertoires. *Nature Scientific Reports*. 4 :6554.

For the published article, ERW performed all model simulations; ERW and COB wrote the manuscript; and COB, YHG and SG designed the study and gave advice.

2.2 Abstract

Pathogen evolution is influenced strongly by the host immune response. Previous theoretical studies on the effects of herd immunity on the antigenic diversity and structure of directly transmitted, short-lived pathogens have primarily focused on the impact of competition for hosts. In contrast, for long-lived infections, theoretical work has focused on the mechanisms promoting antigenic variation within the host. In reality, successful transmission requires that pathogens balance both within- and between-host immune selection. The Opa adhesins in the *Neisseria* genus provide a unique system to study the evolution of the same antigens across two major pathogens which exploit contrasting life history strategies: while *N. meningitidis* is an airborne, respiratory pathogen which colonises the nasopharynx, *N. gonorrhoeae* causes sexually transmitted infections. In this chapter, we aim to investigate how the duration of infection of a pathogen affects within- and between- host antigenic diversity, using a simple

mathematical model. We then analyse published genomic data of Opa repertoires in *N. meningitidis* and *N. gonorrhoeae* to qualitatively evaluate the extent to which the data support the results of the model. As gonococci evolved to cause chronic infection in hosts, we would hypothesise that gonococci would manifest greater within-host antigenic diversity and lower levels of between-host antigenic diversity than meningococci.

2.3 Introduction

Pathogen life histories are shaped by natural selection to maximise reproductive fitness: pathogens must avoid clearance by the host immune system long enough to ensure transmission to a new host lacking pre-existing immunity. Host immune defences therefore exert strong selective pressures on both the within- and between-host components of pathogen life histories. These processes have generated structured patterns of antigenic diversity within pathogen genomes that facilitate the evasion of herd and within-host immunity (Wisniewski-Dye & Vial 2008), promoting both high levels of allelic diversity at the population level (Henrichsen 1995; Buckee *et al*. 2010) and antigenic variation during infection (Van der Ploeg *et al*. 1984). The relative impact of within- and between-host selection on pathogen diversity is expected to depend on the mode of transmission: compared to airborne pathogens transmitted through the respiratory route, sexually transmitted or vector-borne pathogens must contend with often infrequent and uncertain opportunities for transmission, and may cause long-lived infections as a result, facilitated by frequent antigenic variation (Ronald 1996; McMichael & Phillips 1997; Lipsitch & O'Hagan 2007; Lange & Ferguson 2009).

The effects of different types of immunity on the structuring and dynamics of pathogen populations have been a focus of theoretical studies for several decades (Castillochavez *et al*. 1989; Nowak *et al*. 1990; Gupta *et al*. 1994a; Ferguson *et al*. 2003), but most have focused

either on herd immunity as a driver of population level diversity in short-term infections, or on the interactions with an individual immune system that orchestrate antigenic variation in long-term infections. Since pathogen diversity will ultimately reflect both these processes, understanding population structure will require the unification of the study of within-host and between-host immune selection pressures.

The Opa proteins of the bacterial *Neisseria* genus provide a unique framework for studying the impact of within- and between-host evolutionary pressures on the structuring of antigenic diversity in two distinct and globally important pathogens: the meningococcus and the gonococcus. These pathogens colonise distinct sites within the human host, with highly contrasting invasive disease outcomes. *N. meningitidis* colonises the human nasopharynx and although is primarily carried asymptomatically, can give rise to meningitis and septicaemia; whereas the gonococcus primarily colonises the genito-urinary tract in addition to the pharynx and rectum, and causes gonorrhoea and other invasive infections.

There are relatively few studies which have measured the duration of carriage/infection of meningococci and gonococci, and of these, it is difficult to conclude definitely that they accurately reflect the natural duration of infection (Table 2.1). For *N. meningitidis*, the duration of colonisation has been reported to vary between 30 days and 9 months on average; although it should be noted that measuring the duration of colonisation is limited by the length of the longitudinal studies themselves, and it is likely that some meningococcal strains may continue to stably colonise hosts for periods which are longer than the duration of these studies. Obtaining a single set of concrete estimates is particularly challenging given that certain lineages may have longer durations of infection than others. Caugant *et al*. (2006), for example, showed that the ST-23 complex has a longer duration of colonisation relative to other clonal complexes, whereas the ST-11 complex was not recorded in any carriers in the study (suggestive perhaps of a very short duration of infection). For *N. gonorrhoeae*, the estimates

available on durations of infection for gonorrhoeal disease date from the post-antibiotic era, but it is likely that the duration of most gonococcal infections were much longer before antibiotics were introduced. Before the introduction of antibiotics, gonorrhoea was traditionally viewed as giving rise to chronic and lifelong infection (for example, the medieval word "gleet" refers to chronic inflammation as a result of gonorrhoeal infection).

Although the estimates of the durations of infection of the two pathogens stated in Table 2.1 show some overlap, (and that the "true" durations of colonisation of *N. meningitidis* are indeed likely to be longer than stated in the studies below), the natural life history of gonococcal infection is chronic, lasting for many years. Indeed, in medieval times, it was believed that almost all women carried gonorrhoea and transmitted it to their unwitting male partners until their death (Mitchinson, 2013). There are also several writings dating back to the 16th century in which gonorrhoea is described as lifelong (such as those quoted by Norris, 1913). Thus, it is likely that the estimates of natural colonisation are likely to be much longer than the published estimates for both pathogens, but that the duration of infection of gonococci is still longer than that of meningococci.

	Carriage/asymptomatic	Disease/symptomatic
N. meningitidis (from Mueller *et al.*, 2007)	30 days	3-7 days
N. meningitidis (from Caugant *et al.* 2007a)	90% of carriers harboured the same strain for 5-6 months	
N. meningitidis (from Blakebrough *et al*., 1982)	Half-life of carriage was 3 months	
N. meningitidis (from De Wals & Bouckaert 1985)	9 months	
N. gonorrhoeae (estimates from Wiesner & Thompson, 1980)	3-12 months (men) 3-6 months (women)	3-45 days (men) 2-30 days (women)
N. gonorrhoeae (WHO 2012)	5-6 months (men) 6 months (women)	2-3 weeks (men) 3-4 weeks (women)

***Table 2.1 Estimates of the duration of symptomatic and asymptomatic infection with* Neisseria meningitidis *and* Neisseria gonorrhoeae.**

Multiple *opa* loci are present in each genome, encoding adhesins that play a central role in attachment to, and invasion of, host epithelium cells. Meningococci contain 3-4 *opa* loci (Tettelin *et al*. 2000), whereas the Opa repertoires of gonococci are larger, comprising 11-12 loci (Bhat *et al*. 1991). These sequences are not only highly diverse on a population level, owing to high rates of recombination and gene conversion, but also exhibit flexible expression within individual infections (Hobbs *et al*. 1994). Thus, they provide a unique opportunity to study the effects of the selection by the immune system at both the population level and within individual isolates in two related pathogens with contrasting life histories. In addition, the Opa proteins have been shown to elicit bactericidal antibodies and are currently being considered as vaccine candidates against meningococcal infection (Callaghan *et al*. 2011). Understanding the processes underlying the maintenance of diversity and structuring of the Opa proteins therefore has implications for the prospective design and evolutionary consequences of such vaccines.

The sequence diversity among *Neisserial* Opa proteins is primarily limited to two hypervariable (HV) regions, corresponding to two surface exposed loops that have been shown to interact with the host immune system (Malorny *et al*. 1998; de Jonge *et al*. 2003). Each *opa* locus is constitutively transcribed and protein expression is controlled by phase variation (due to insertions and deletions of pentanucleotide repeats in the leader coding sequence of each *opa* allele) (Stern *et al*. 1986). The expression of different Opa proteins has been shown to vary during the course of gonococcal infection, suggesting that Opa antigenic variation could prolong infection through the avoidance of clearance by the humoral host immune response (Swanson *et al*. 1988; Jerse *et al*. 1994). Within-host variation at the Opa loci has yet to be observed for meningococci, but the presence of mechanisms required for phase variation suggests that it is possible (Stern *et al*. 1986).

Previously, Callaghan *et al*. (2008) have shown that strong cross-immunity between hosts can structure the diversity of meningococcal Opa. Here, we extend this mathematical model to explain the differences in Opa repertoire structures that we observe between gonococci and meningococci, and to examine the trade-offs faced by pathogens generally between maximising within-host fitness and competing on a population level for hosts. Despite their relatively recent shared evolutionary history, gonococci and meningococci have exploited divergent life strategies. Here we show that the contrasting structures of their Opa antigenic repertoires can be parsimoniously explained by a different balance of within- and between-host selection pressures conferred by the host immune response, mediated by different infectious periods.

2.4 Methods

2.4.1 Calculating Opa Repertoire Diversity

We examined the structure and diversity of published Opa repertoires of meningococcal and gonococcal isolates (Callaghan *et al*. 2008; Bilek *et al*. 2009) in the context of allelic genomic repertoire diversity. Opa allelic data for *N. gonorrhoeae* were taken from the isolates published by Bilek *et al.* (2009), and for *N. meningitidis* from the isolates presented by Callaghan *et al.* (2008). Only the unique Opa repertoires were included in order to avoid bias by the prevalence of particularly widespread and successful strains. First, unique Opa repertoires were identified, and subsequently those isolates expressing shared alleles at a minimum of two loci were selected. Out of a total of 216 *N. meningitidis* isolates, 148 Opa profiles were unique and of these, 64 showed identical Opa alleles at two or more loci. Out of a total of 150 *N. gonorrhoeae* isolates, there were 138 unique Opa profiles and of these, 95 showed identical Opa alleles at

two or more loci. For each Opa profile showing shared alleles, the percentage of expressed Opa alleles which were shared in the repertoire was calculated.

2.4.2. Mathematical Model

We developed a simple mathematical model to represent the idea that within-host antigenic diversity can prolong individual infections through antigenic variation, but that strains sharing antigenic alleles will be competing for hosts through herd immunity. The mathematical model was based on an extension of the framework by Callaghan *et al* (2008), in which two *opa* loci each comprised two HV regions, which were permitted to express one of two possible alleles (Figure 2.1a). In this framework, HV1 could express either "a" or "b", and HV2 could express either "x" or "y," thereby creating nine different strains in total: *ax*, *bx*, *ay*, *by*, *axy*, *bxy*, *abx*, *aby* and *abxy*. No dose dependence was assumed (*i.e.* strains *ax* and *axax* were taken as immunologically identical), and immunological cross-immunity (referring to immune responses against antigenically similar strains) was included (Figure 2.1b).

To incorporate antigenic variation into the model, we assumed that increased Opa repertoire diversity was associated with longer durations of infection: strain *abxy* has a longer duration of infection than *aby*, which in turn has a longer duration of infection than *ab*. Thus, antigen repertoire diversity facilitated chronicity in the host, but also increased the probability of sharing antigenic alleles with other pathogen strains in the population. Our goal was to determine whether the simple within- versus between-host trade-offs imposed by allele-specific immunity could robustly explain the patterns of Opa repertoire diversity observed in both of these pathogens. `

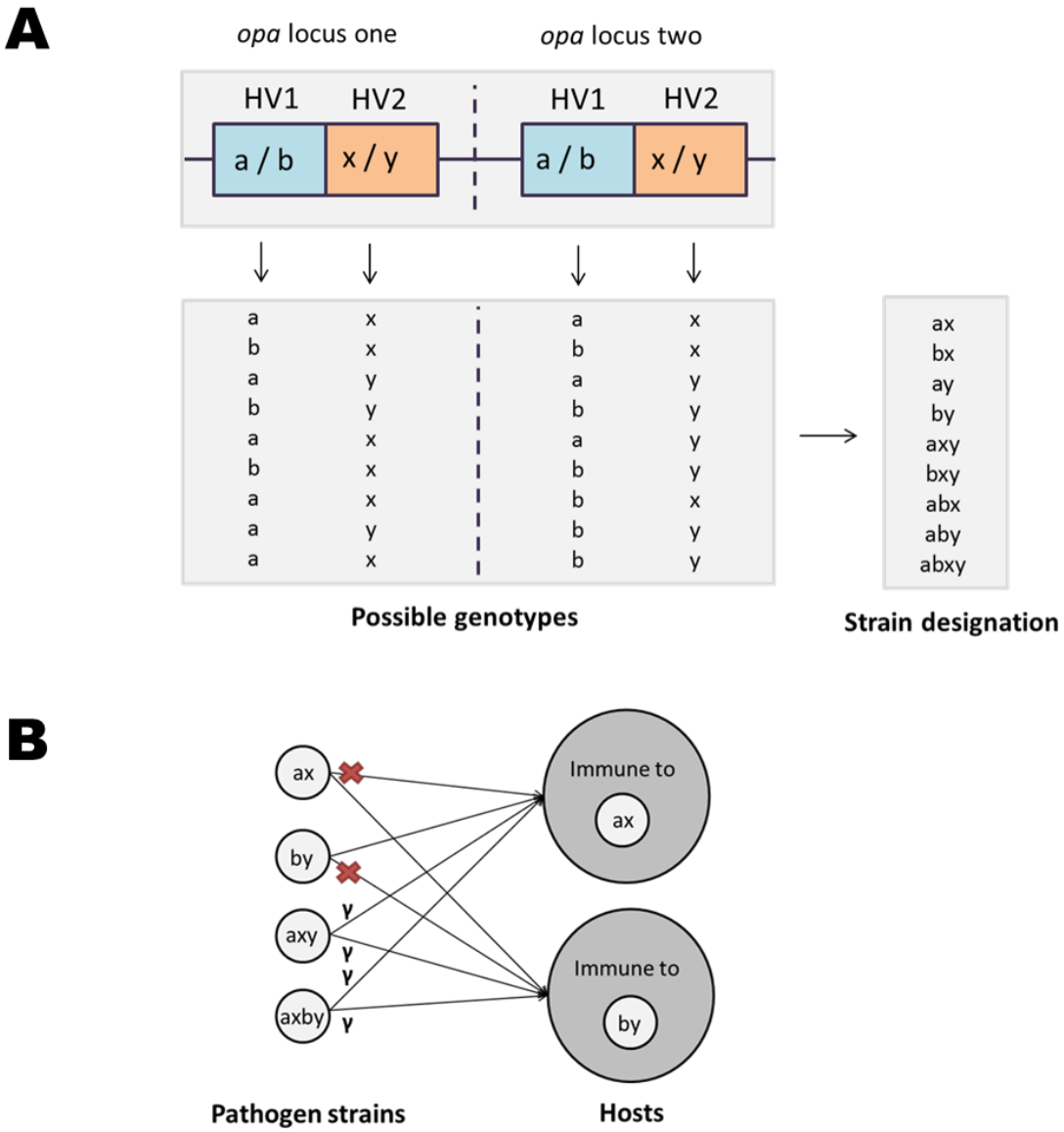


Figure 2.1. Overview of the model. *(A) Schematic diagram of strains included in the model. Each isolate has two opa loci, each with two hypervariable regions (HV1 and HV2). One of two possible alleles can be expressed at each Opa locus (a or b at HV1; x or y at HV2). It is assumed there are no dose response effects; i.e. a strain with the genotype bxbx is considered identical to bx. (Diagram modified from Callaghan et al., 2008). All possible strains can be selected for in the model (i.e. strains do not become extinct, they simply fall to extremely low frequencies), to account for the possibility of recombination between strains. (B) The model assumes that strains are inhibited from infecting hosts co-infected with or immune to strains sharing antigenic determinants by the parameter* γ *(where* $\gamma = 1$ *confers complete cross-immunity).*

The system can be described by three differential equations, based on the model by Callaghan *et al.* (2008):

$$\frac{dz_i}{dt} = \beta y_i(1 - z_i) - \varepsilon z_i - \mu z_i \quad (2.1)$$

$$\frac{dw_i}{dt} = \sum_{i'} \beta y_i(1 - w_i) - \varepsilon w_i - \mu w_i \quad (2.2)$$

$$\frac{dy_i}{dt} = \beta y_i\,[(1 - w_i) + (1 - \gamma)(w_i - z_i)] - \sigma_i y_i \quad (2.3)$$

Once infected by a strain, the host gains partial immunity to any other strain which expresses shared antigenic determinants (the subset of strains *i'* above). The model comprises three compartments: z_i represents the proportion of the host population immune to strain *i*; w_i represents the proportion immune to any strain; and y_i represents the proportion currently infected with strain *i*. The rate of loss of infectiousness of a strain is conferred by σ_i (and therefore the average duration of infection is given by $1/\sigma_i$); β represents transmissibility (a strain's likelihood of being transferred from an infected to a susceptible host); ε confers the rate of loss of immunity; and the average death rate within the host population is conferred by μ. It is assumed that all strains have the same transmissibility. The force of infection (λ) of any strain is determined by the product of its transmissibility and its total prevalence in the infected population (i.e. $\lambda_i = \beta * Y_i$). The basic reproductive number, R_0, can be approximated to β/σ_i. Rather than include explicit recombination in the model, we simply allow all strains to circulate at all times even if they exist at very low frequencies. This can be seen as a way to incorporate recombination, since any combination of alleles can be selected for in the population.

It is assumed that upon infection the host enters the immune and infectious categories concurrently. Cross-immunity is controlled by γ: hosts that have been infected with a strain

sharing antigenic determinants with *i* will become infectious with the probability (1- γ) upon infection with strain *i* (i.e. when γ=1, full cross-immunity operates between strains). The rate of loss of infectiousness is set to differ between strains, and depends on the number of antigenic loci a strain exhibits. Three σ values control the dynamics: the value assigned to strains *ax*, *ay*, *bx* and *by* always exceeds that of strains *axy*, *bxy*, *abx* and *aby*, which always exceeds that of strain *axby*. Three rates of increase in duration of infection ($1/\sigma_i$) as a function of increasing strain antigenic diversity were explored (shown in Figure 2.4), and are as follows.

(A) The infectious period increases by 35% when the strain moves from 2 to 3 antigenic loci, and by a further 10% when it exhibits 4 antigenic loci. (B) The infectious period increases by 22.5% per additional antigenic locus. (C) The infectious period increases by 10% when the strain moves from 2 to 3 antigenic loci, and by a further 35% when it exhibits 4 antigenic loci. (D) The infectious period increases by 7.5% per additional antigenic locus.

Except where noted, the exact parameter values used were as follows: μ=0.02; σ varied between 1 and 3.6 (corresponding to periods of infection of 1 year to approximately 3 months, respectively); β = 4 (in order to investigate R_0 values ranging from 1.1 to 4); ε varied between 0 and 365 (corresponding to periods of immunity of lifelong duration to 1 day, respectively). The periods of duration of infection, rates of loss of immunity, and reproductive number are unfortunately poorly understood for both meningococci and gonococci, but the range of values used here are consistent with the estimates available in the literature (Mueller *et al*. 2007; Yorke *et al*. 1978; Callaghan *et al*. 2011; Noble *et al*. 1977; Hedges *et al*. 1999; Zak *et al*. 1984; Plummer *et al*. 1994; Cole & Jerse 2009; Buchanan *et al*. 1980; Brooks & Ingwer 1978; Trotter *et al*. 2005). The average duration of infection with gonococci is likely to lie closer to the longer asymptomatic estimates, as it is thought that asymptomatic women are responsible for 80-98%

of the transmissions, and asymptomatic men for 60-80% of the transmissions (Hethcote *et al*. 1982).

2.5 Results

The structure and diversity of the Opa repertoires differ markedly between *Neisseria meningitidis* and *Neisseria gonorrhoeae*. Figure 2.2 illustrates a comparison of the within-repertoire allelic diversity among Opa proteins. The meningococci frequently contain identical alleles at multiple Opa loci in comparison to gonococci, which show high levels of within-repertoire diversity and many isolates exhibiting unique Opa alleles across multiple loci (Callaghan *et al*. 2008; Bilek *et al*. 2009). These differences are not simply a function of the number of Opa loci present, but represent a qualitatively different pattern of diversity.

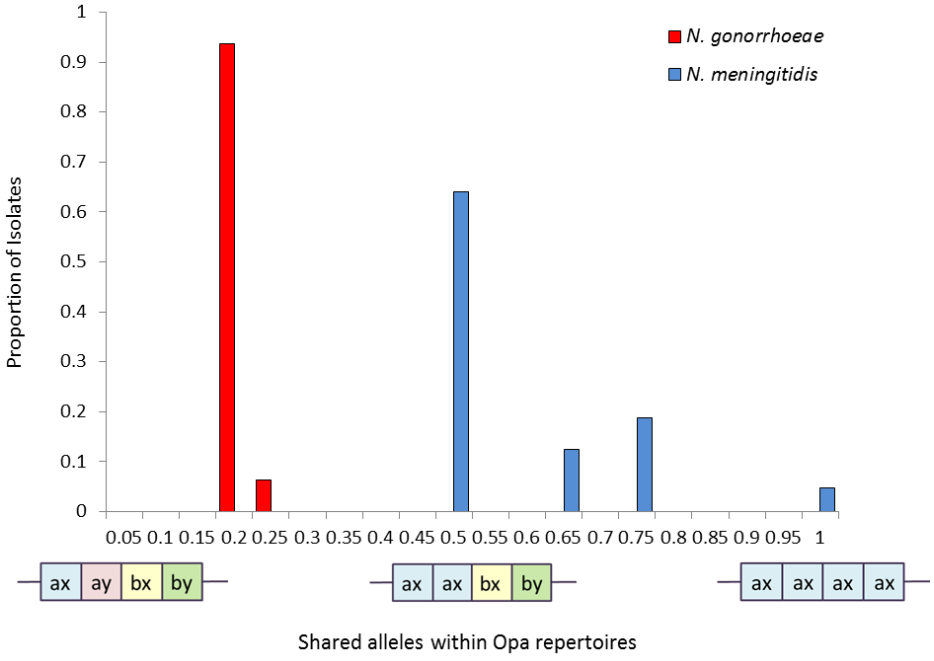


Figure 2.2. The proportion of alleles across individual genome's Opa repertoires which are identical, among isolates expressing shared loci. *Red bars: N. gonorrhoeae (95 isolates); blue bars: N. meningitidis (64 isolates). Calculated using data taken from Bilek et al. (2009) and Callaghan et al. (2008). The schematic underneath the x axis is a conceptual representation of the proportion of alleles within a repertoire of four loci which are identical.*

We can classify our model results broadly into those characterised by high or low reproduction numbers (Figures 2.3a and b, respectively). For pathogen populations with high transmission (R_0) values, strong cross-immunity between hosts led to the stable emergence of strains characterised by a minimal set of non-overlapping alleles (strains *ax* and *by*), whereas low to moderate levels of cross-immunity led to the co-existence of all possible strains, consistent with the results of other theoretical frameworks (Gupta *et al*. 1996; Gupta *et al*. 1998). Strongly non-overlapping patterns among antigenic repertoires have been observed in several bacterial populations, including a number of outer membrane proteins in meningococci as discussed previously (such as PorA, FetA and Opa; Callaghan *et al*. 2006; Callaghan *et al*. 2008; Buckee *et al*. 2010), as well as Group A Streptococci (Johnson *et al*. 2006). Here, strain *axby* cannot emerge since many hosts had cross-immunity to one or more of its antigenic determinants. We envision this scenario to reflect the pattern seen with the meningococcus, and other short-lived and highly transmissible pathogens: low within-repertoire diversity but high diversity at the population level to maximise between-host fitness (Callaghan *et al*. 2008).

In contrast, for the scenario in which the R_0 of the pathogen population is low (Figure 2.3b), strain *axby* had the highest prevalence over all levels of cross-immunity. To investigate the effects of the longer durations of infection facilitated by antigenically complex strains (*axy, bxy, abx, aby, axby*), as well as a larger range of R_0 values, we compared the results of four distinct relationships between the duration of infection and the number of antigenic loci (Figure 2.4; see also Figure 2.6). As the relative fitness advantage to antigenically diverse strains decreased (i.e. moving from *a* to *d* in Figure 2.4), strain *axby* was less likely to dominate the population and the non-overlapping strains *ax* and *by* became more prevalent. These results were not qualitatively altered by waning immunity (see Figure 2.6).

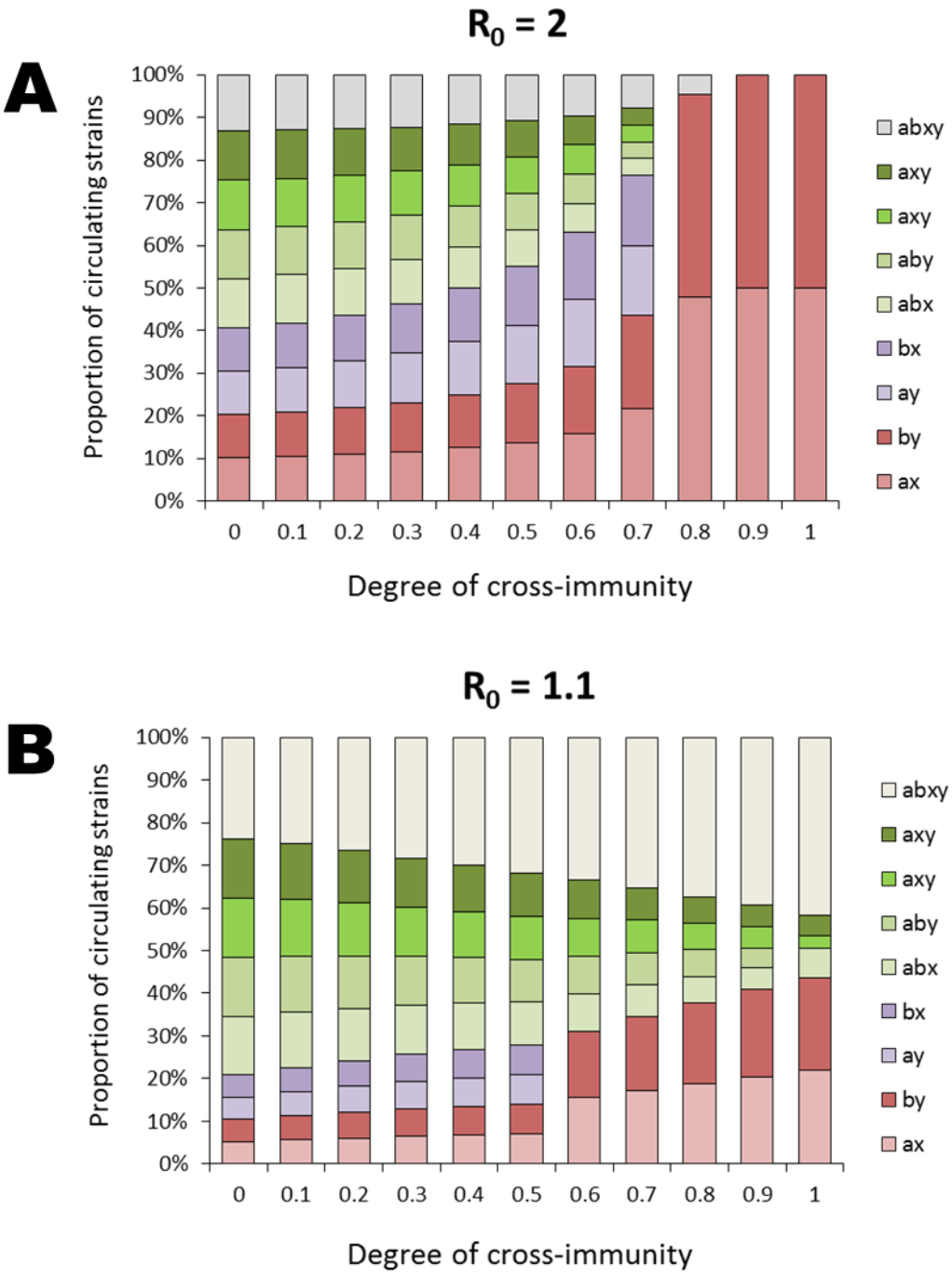


Figure 2.3. The effects of cross-immunity on strain frequencies at different R_0s. *(A) R_0=2 (β=4, σ =2); (B) R0=1.1 (β =4, σ=3.6). In (A), strain axby is lost from the population at higher values of cross-immunity; whereas at the lower R_0 of 1.1, axby dominates the population at all levels of cross-immunity. The rate of increase in infectious period conferred to increasing antigenically diverse strains corresponds to (D) in Figure 2.4.*

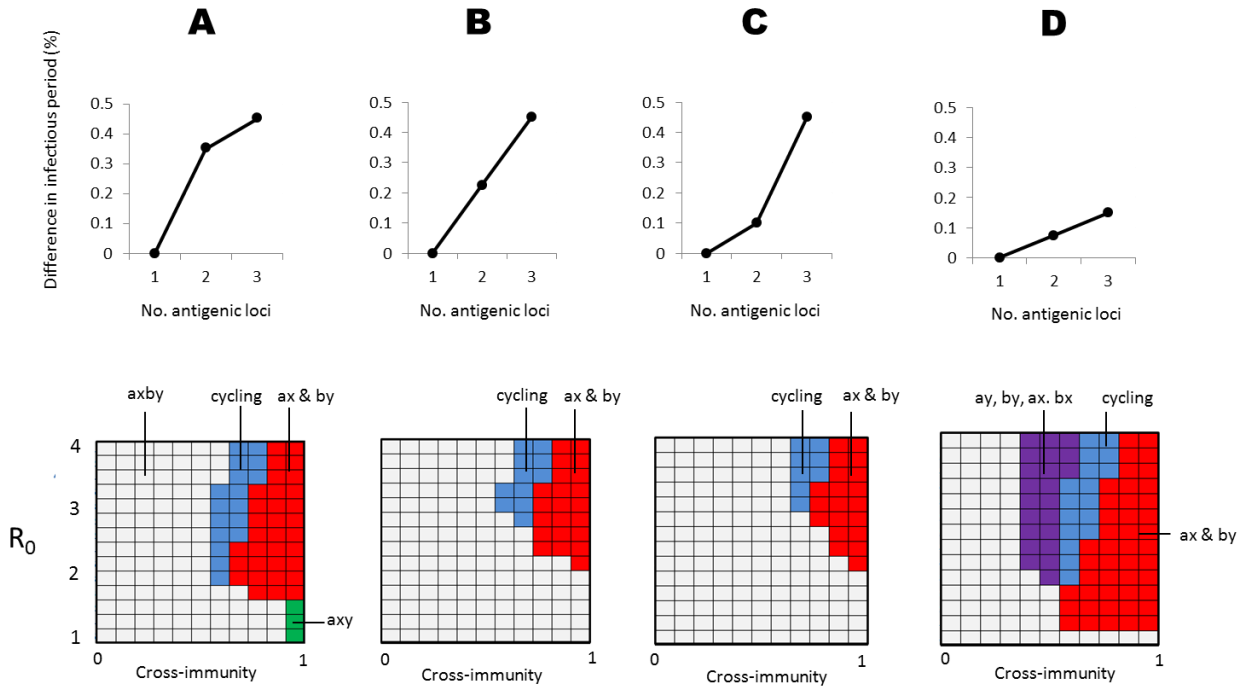


Figure 2.4. The effects of different relationships between increasing antigenic diversity and the increase in infectious period conferred. *As the infectious period advantage conferred to antigenically diverse strains decreases (moving from A to D), strain* axby *(light grey) is less likely to dominate the population. Even when the increase in infectious period is relatively shallow,* axby *can emerge at all values of R_0 when cross-immunity is sufficiently low.*

Oscillations between opposing sets of discordant antigenic types (in which non-overlapping strains *ax* and *by* cycled in prevalence over time with *ay* and *bx*) was observed over a similar amount of parameter space in all simulations with lifelong immunity (see Figure 2.5 for typical dynamics observed during antigenic cycling), but observed less frequently as the period of immunity decreased (see Figure 2.6). Antigenic cycling has been observed in the PorA and FetA outer membrane proteins of *N. meningitidis* (Buckee *et al*. 2008; Buckee *et al*. 2010) but not among gonococci, to our knowledge. Indeed, the emergence of antigenic cycling at longer durations of immunity is consistent with the evidence that meningococci invoke longer-lasting immune responses than gonococci (Noble *et al*. 1977; Hedges *et al*. 1999; Callaghan *et al*. 2011). In summary, when transmission rates and herd immunity were low, strains expressing

the largest and most diverse antigen repertoires dominated the population due to their superior fitness within hosts. We envision that the strains emerging in these areas of parameter space represent the gonococcus and other long-lived pathogens with lower transmission rates.

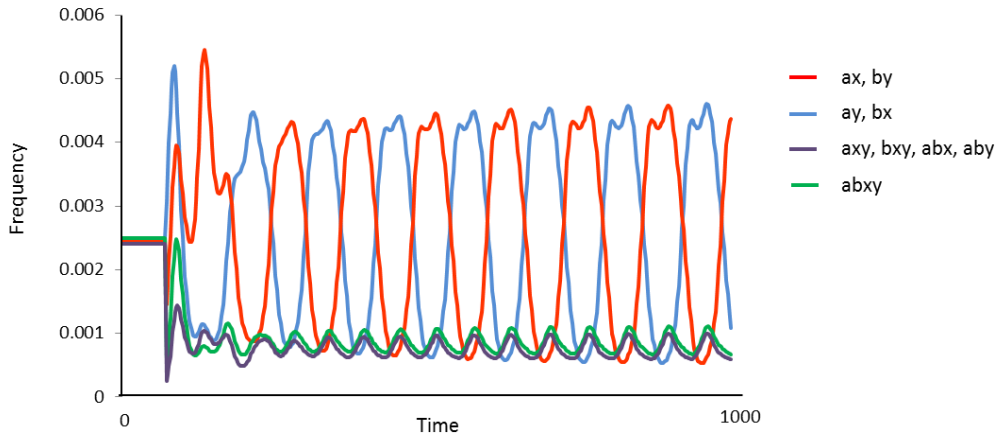


Figure 2.5. The phenomenon of ''antigenic cycling'' between alternating sets of discordant strains (ax, by and ay, bx). *(β=4, σ=2, γ=0.6).*

Across all rates of loss of immunity (ε), the strain *axby* predominates at low levels of cross-immunity, and non-overlapping strains *ax* and *by* are observed at high levels of cross-immunity (Figure 2.6). Increasing the rates of loss of immunity primarily affected the behavior of antigenic cycling, observed predominantly at intermediate values of cross-protection under lifelong immunity (top row, Figure 2.6). When the rate of loss of immunity exceeded a threshold ($\varepsilon > 0.03$; bottom row, Figure 2.6), antigenic cycling was no longer observed, and replaced by different behaviors (with the exception of relationship **e**). Above this threshold of ε, the behavior of the system does not change (data not shown). At intermediate rates of loss of immunity between 0 and 0.03 (middle row, Figure 2.6), an intermediate regime of behaviors can be observed. The overall observation that antigenically complex strains such as *axby* tend to dominate when cross-immunity and R_0 is low, remains unchanged.

We also explored the scenario in which increased antigenic diversity was associated with shorter durations of infection (Figure 2.6E). The strains *ax, by, ay* and *bx* dominated over the majority of parameter space, with non-overlapping strains *ax* and *by* dominating at high levels of cross-immunity. Antigenically complex strains such as *axby* did not emerge at any point, as the penalty of sharing antigenic determinants with more strains – thereby impeding transmission - was not balanced by an increased duration of infection.

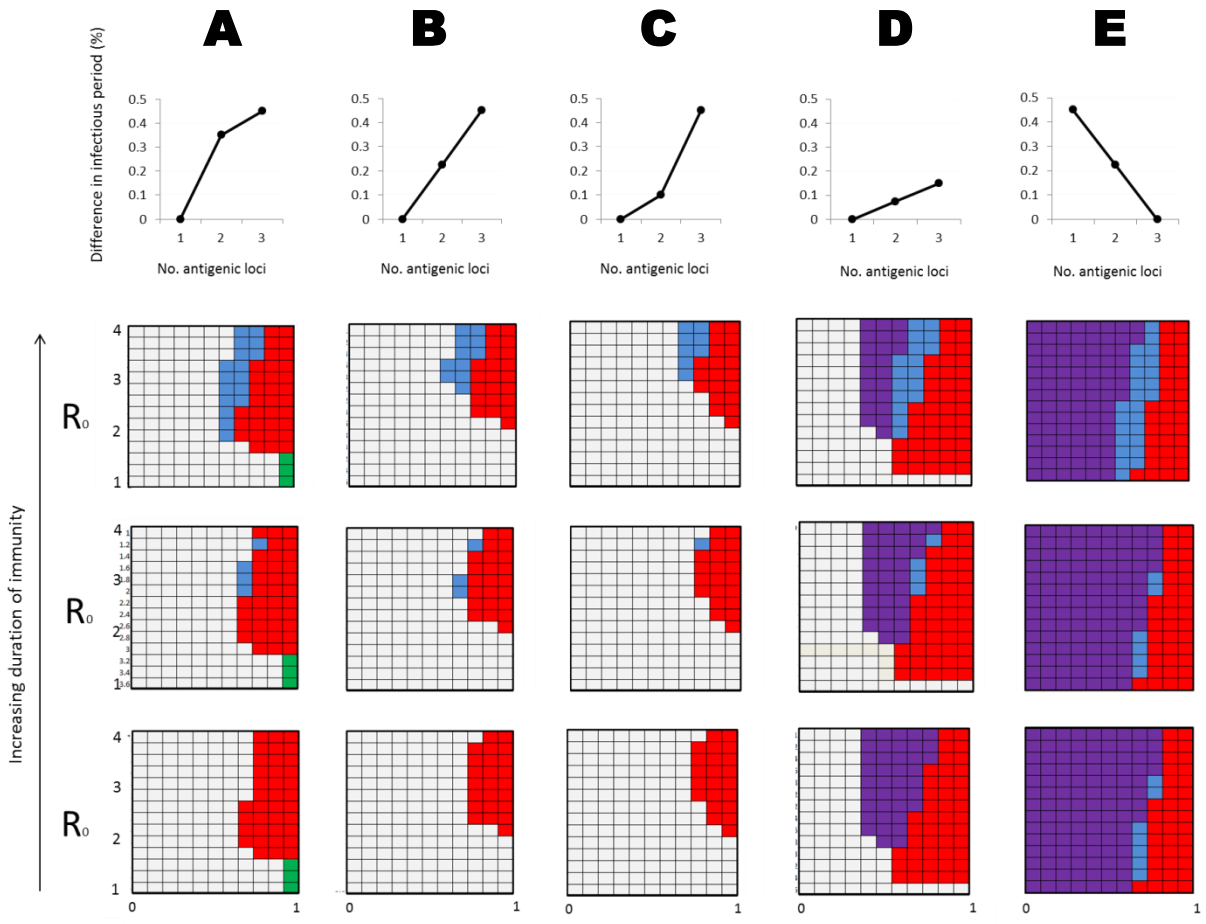


Figure 2.6. The effects of waning immunity on the model outcome of predominating strains observed at equilibrium. *The rates of loss of immunity (ε) are as follows: top row, 0; middle row, 0.025; bottom row, 0.03. Increases in infectious period conferred to antigenically more complex strains are consistent with Figure 2.4 (with the exception of E, which is the reverse of relationship B). The predominant strain(s) observed for each simulation are represented by the following colours: grey, axby; red, ax and by; purple, ax, ay, bx and by; green, axy; blue, continuous cycling of non-overlapping strains ax, by and ay, bx.*

Our model results are therefore consistent with the observed differences in within-repertoire Opa allelic diversity between meningococci and gonococci (Callaghan *et al*. 2008; Bilek *et al*. 2009) (shown in Figure 2.2), despite the simplicity of our assumptions. Meningococcal repertoires contain a much greater proportion of identical alleles in comparison to gonococcal repertoires, corresponding to higher R_0 values and strong cross-immunity between hosts in our model. Callaghan *et al*. (2008) previously suggested that strong cross-immunity between strains on a population level can explain this low within-repertoire diversity in the meningococcus, a finding consistent with our results, and the data available on immune responses to meningococcal Opa antigens (Callaghan *et al*. 2011). The results here suggest that within-host selection for antigenic variation in the meningococcus is not strong enough to counteract the strength of between-host selection. In contrast, the gonococcal repertoires showed few repeated Opa alleles within repertoires, consistent with our model when R_0 is relatively low and/or when cross-immunity is weak and antigenic variation is an important driver of long-lived infections. For the gonococcus, therefore, within-host selection plays a key role in structuring antigen repertoires. Although the specificity and duration of immunity conferred by gonococcal Opa proteins is not well understood, with different studies yielding contrasting results (Brooks & Ingwer 1978; Noble *et al*. 1977; Buchanan *et al*. 1980; Hedges *et al*. 1999; Zak *et al*. 1984; Plummer *et al*. 1994; Cole & Jerse 2009), we explored a wide range of rates of waning immunity to confirm that our model results were robust to this assumption (Figure 2.6).

2.6 Discussion

While the effect of cross-immunity between hosts is well-established as a major selective force structuring the antigenic diversity of pathogen populations (Castillochavez *et al*. 1989; Gupta *et al*. 1998; Ferguson *et al*. 2003; Buckee *et al*. 2011), it is often impossible to directly compare pathogens with different life history strategies because of the differences in the structure, function, and biology of their antigenic determinants. Here we have used the Opa antigen repertoires of ecologically distinct *Neisseria* species as a unique system to simultaneously explore the effects of within- and between-host cross-reactive immunity. Our model provides a simple and testable hypothesis to explain different patterns of diversity observed among meningococcal and gonococcal Opa repertoires: the balance of within-host and between-host immune selection pressures enhances repertoire diversity in the gonococcus and minimises it in the meningococcus. Our framework therefore expands beyond the theoretical frameworks that have investigated either the importance of maximising within-host duration of infection or maximising between-host transmission (Nowak *et al*. 1990; Antia *et al*. 1996; Gupta *et al*. 1998), providing a single mathematical framework accommodating both components of pathogen fitness.

Mathematical models represent tools for generating hypotheses and exploring their implications; here we have purposefully chosen a highly simplified framework to show that these evolutionary forces alone can account for patterns of pathogen diversity. This does not preclude other biological mechanisms playing important roles in Opa diversity, however. Our model does not simulate the exact levels of diversity of Opa loci or alleles in meningococcal and gonococcal populations or the patterns of expression, basing the analysis instead on an abstraction from observations to date. As ongoing population genomic studies and long read sequencing platforms define *opa* diversity for *Neisseria* species, it will be important to

incorporate these data to refine and test our model. Additionally, systematic study of *opa* expression under various contexts of infection – including the role of Opa+ gonococci in urethral infection (James & Swanson 1978; Jerse *et al*. 1994; Bos *et al*. 1997), resistance of Opa- gonococci to proteolytic enzymes and relation to sensitivity to hormonal levels (James & Swanson 1978), and under differing levels of pyruvate (Williams *et al*. 1998) – will present another avenue for future study. Furthermore, Opa proteins may subvert the immune system by means other than antigenic variation. There is evidence that gonococcal Opa proteins inhibit innate immune responses of epithelial cells (Muenzner *et al*. 2005), in addition to adaptive B cell (Pantelic *et al*. 2005) and T cell responses (Boulton & Gray-Owen 2002). The suppression of host immunity by Opa proteins is consistent with the results of our model, and these specific host-pathogen interactions may enhance the selection on Opa repertoires as described in our framework of antigen repertoire evolution.

Functional constraints may also contribute to the specific combinations of hypervariable regions expressed, since these regions affect receptor tropism (Virji *et al*. 1999). As specific Opa variants mediate interactions with distinct host receptors, they may influence the tropism for epithelial, endothelial and phagocytic cells. It has thus been suggested that Opa diversity in gonococci enables the colonisation of a wider range of host sites (Dehio *et al*. 1998). Equally, the greater diversity of host sites colonised by gonococci compared to meningococci may also enhance antigenic repertoire size. Our model results are not inconsistent with this additional benefit of within-host diversity, and suggest that these tropisms would be maintained as long as transmission rates were not sufficiently high to generate strong between-host immune selection. Additionally, this functional constraint is unlikely to account for the non-overlapping structures we observe.

There are a number of parallels that might be drawn with the *opa* genes and the *Neisserial* pilin genes, which code for the pilus. Pili also mediate adherence to host cells and are essential for

establishing infection in both gonococci and meningococci (Robertson *et al*. 1977; Exley *et al*. 2009). The pilus is antigenic and the pilins undergo variation, controlled by recombination between one of the silent *pilS* silent loci and the pilin expression locus *pilE* (Hagblom *et al*. 1985). Variation has been shown to occur over the course of gonococcal infection but not meningococcal infection, as for the Opa proteins. The greater number of storage loci in present in gonococci than meningococci (19 as opposed to 8), as well as the higher antigenic variation frequency recorded (0.13 recombination events per cell as opposed to 0.3), support the hypothesis that gonococci may rely on within-host pilin variation to prolong infection, perhaps to a greater extent that meningococci (Criss *et al*. 2005; Helm & Seifert 2010).

2.6 Conclusion

In this chapter, the effects of immune selection on the diversity of antigenic loci were explored at within- and between- host levels in a unified deterministic framework. Our model raises the testable hypothesis that short-lived pathogens manifest non-overlapping, simple antigenic repertoires to maximise between-host transmission fitness, whereas more complex antigenic repertoires are observed in long-lived infections to facilitate immune evasion. The findings are consistent with the observation of higher within-strain diversity in gonococci relative to meningococci, and bear implications for the evolution of different life history strategies and antigenic diversity in other pathogens. In the remaining chapters of my thesis, I explore mathematical models which incorporate the effects of ecological competition at metabolic genes among strains, in addition to immune selection on antigenic genes.

Chapter 3.

Multi-Strain Dynamics with Competition at Metabolic and Antigenic Loci

3.1 Public Dissemination of Work

These findings have been published in the following article, in association with the results in Chapters 5 and 7:

Eleanor R. Watkins, Bridget S. Penman, José Lourenço, Caroline O. Buckee, Martin C. J. Maiden & Sunetra Gupta (2015). Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. *Plos Pathogens*, 11(7): e1005034.

For the published article, ERW and SG conceived and designed the experiments; ERW performed the mathematical modelling and statistical tests; ERW analysed the genomic data and output from the mathematical models; and SG, ERW, JL, COB, and MCJM wrote the paper.

3.2 Abstract

Several mathematical models have been developed in order to understand the mechanisms which underlie the maintenance of strain diversity in pathogen populations. In the previous chapter, the effects of cross-immunity on antigenic diversity were examined at both within- and between- host levels. In this chapter, we aim to understand the effects of immunological competition operating at the antigenic loci, and ecological competition at metabolic loci using multilocus deterministic mathematical models. We first explore the effects of complete ecological and immunological competition, and compare the results in two different structures of deterministic models. We then explore the effects of varying the strengths of immunological and ecological competition on population structure. Based on the results of Buckee *et al.* (2008), we would hypothesise that strong competition leads to a population of strains

comprising stable non-overlapping associations between antigenic and metabolic alleles, and that these associations are not maintained when competition is relaxed. Our results show that strong ecological competition at metabolic loci combined with strong immunological competition at antigenic loci indeed does lead to the stable co-existence of strains with unique associations between metabolic and antigenic alleles, and that such patterns can emerge even when both ecological and immunological competition are relaxed to intermediate levels.

3.3 Introduction

3.3.1 Diversity in Bacterial Populations

A number of mathematical models of strain competition have been invoked to explain strain structure in recombining populations (reviewed in section 1.3). Although a number of frameworks have examined the effects of immune-mediated and direct ecological competition on the diversity of pathogen populations (e.g. Lipsitch 1997; Recker *et al*. 2004; Gupta *et al*. 1998; Flasche *et al*. 2013), few have examined the effects of both selection pressures in the same framework (Zhang *et al*. 2004). It is likely, however, that both selection pressures are operating in many bacterial pathogens.

Buckee *et al*. (2008) incorporated the effects of immunological competition at antigenic genes, and ecological competition at metabolic genes, within an individual-based stochastic framework, and was able to reproduce a number of features of meningococcal molecular epidemiology (see Section 1.3). Strains in the model comprised two antigenic loci each with two alleles (a/b and x/y), corresponding to the antigenic type (AT), and one locus for metabolic housekeeping genes with two alleles (1/2), corresponding to the sequence type (ST). The antigenic alleles conferred allele-specific immunity, and the ability of a strain to infect a host

already infected with a strain sharing one or more of its antigenic variants was controlled by a cross-immunity parameter. Strains with the same ST were not able to co-infect a host due to ecological competition. At high levels of cross-immunity between strains, stable and fluctuating non-overlapping associations between ATs and STs were observed at small and intermediate differences in transmission differences between STs respectively.

These findings were consistent with observations of discordant associations between housekeeping metabolic genes and outer membrane antigens in *Neisseria meningitidis* from several studies, some of which were observed to fluctuate over time (Caugant *et al*. 1986; Caugant *et al*. 1987; Urwin *et al*. 2004; Buckee *et al*. 2010). As discussed in Section 1.3, non-overlapping associations have also been observed in many studies between serotypes and MLST-defined STs of *Streptococcus pneumoniae* (Table 1.3), which may fluctuate over time. Here we present a series of models which analyse the effects of immunological and ecological competition at antigenic and metabolic loci in combined deterministic frameworks, where the antigenic type is dictated by one locus, and immunological competition is dictated by the strength of variant-specific immunity.

We adopt a general framework with overlapping compartments, in which the infected population are embedded within the recovered population (Gupta *et al*. 1998). This approach is distinct from that used in a number of SIR-based epidemiological models, in which all classes are separate and do not overlap (e.g. Trotter *et al*. 2005; Flasche *et al*. 2013). The overlapping, simplified frameworks are able to avoid the large increase in equations and parameters usually associated with adding extra strains into epidemiological frameworks, but it is not known how these differences from traditional SIR models with separate compartments affect model behaviour and outcome. Thus after showing the derivations of the model formulations from traditional SIR frameworks, we compare the results of the numerical simulations using two analogues of two and four strain models.

In Buckee *et al.*'s framework, non-overlapping associations between ATs and STs were observed at full ecological competition and high levels of cross-immunity; here, we explore the effects of relaxing the strengths of ecological and immunological competition on strain diversity and population structure, and investigate the parameter space under which non-overlapping associations between antigenic and metabolic alleles are observed. Given that the results of Buckee *et al*. showed that strong competition polarises the population into non-overlapping associations, we would hypothesise that such associations are not observed across more relaxed levels of competition, and that overlapping combinations arise. We also explore the effects of attributing R_0 to the AT, and varying the duration of infection between strains while keeping R_0 constant. Our results are discussed in the context of meningococcal and pneumococcal epidemiology.

3.3.2 Model Assumptions

A summary of the assumptions of the models is provided in Table 3.1

	Antigenic Type (AT)	**Metabolic Type (MT)**
Definition	Genes encoding targets of protective immunity	Constellation of genes within a bacterium which function in the uptake of nutrients and metabolic processes
Effect of infection	Inhibits subsequent infection with strains of the same AT	Inhibits coinfection with strains of the same MT
Mechanism	Antigenic variant-specific immunity (e.g. serotype-specific immunity)	Competition for nutrient/receptors
Duration of effect	Lifelong (or shorter if there is waning of immunity)	During carriage
Designation	a, b, c ...	1,2,3...
Example	Capsular serotype (*S. pneumoniae*, *S. agalactiae*) PorA, FetA type (*N. meningitidis*)	Sequence type (ST) defined by MLST (see also Chapter 5)

Table 3.1. Model assumptions in the model frameworks which include ecological and immunological competition.

3.3.2.1 Immunological Competition between Strains with Identical Antigenic Types

Evidence for naturally-acquired serotype-specific immune responses in *S. pneumoniae* has been provided from longitudinal studies, in which the probability of subsequent colonisation with the same serotype has been observed to be reduced for some serotypes (Weinberger *et al*. 2008; Goldblatt *et al*. 2005; Hill *et al*. 2008). The model assumption of a reduced probability of infection with the same serotype is consistent with other mathematical models of pneumococcal colonisation (Cobey & Lipsitch 2012; Bottomley *et al*. 2013; Flasche *et al*. 2013). However, the exact strengths of serotype-specific protection remain unknown for a number of serotypes. Here we explore a wide range of values of serotype-specific immunity to reflect wide variation and uncertainty. Consistent with the assumption of Cobey & Lipsitch (2012) and Bottomley *et al*. (2013), we assume that anticapsular immunity is lifelong. The addition of waning immunity into the model did not qualitatively alter the model outputs.

Variant-specific immune responses resulting in protective immunity has been shown against the outer membrane antigens PorA and FetA of meningococci in a number of studies (Vermont *et al*. 2005; Rosenqvist *et al*. 1995; Wedege *et al*. 2003; Marzoa *et al*. 2012). The exact strengths of immunity against re-infection with the same type, however, are not known.

3.3.2.2 Ecological Competition between Strains with Identical Metabolic Types

A key assumption of the models explored here is that of metabolic interference between strains; that is, that two strains with the same set of metabolic alleles are prevented, to an extent, from co-infecting a host. Carriage of multiple strains is believed to be relatively frequent in individuals colonised with *S. pneumoniae*, with multiple carriage rates of 20-50% recorded (Turner *et al*. 2011; Carvalho *et al*. 2010; Satzke *et al*. 2013). It is believed to be less frequent

with strains of *N. meningitidis*, but stable co-colonisation over several months has been recorded (reviewed by Caugant *et al*. 2007). Several studies show strain-specific differences in metabolism and/or metabolic loci among pneumococci and meningococci, and I have surmised three principal ways through which bacteria with different metabolic types might be able to stably co-colonise a host.

Firstly, given that differences in the transporters and metabolic pathways encoded in the genome dictate specific substrate repertoires for different strains, two strains with different repertoires of transport and metabolic genes may be able to co-infect a host as they are able to make use of different substrates (Figure 3.1a). In contrast, two strains with identical repertoires of transport and metabolic genes may not be able to co-infect a host if they utilise an identical repertoire of substrates. Strain-specific differences in the presence of transporters of particular metabolic substrates have been recorded in a number of studies. The importance of carbohydrate transport was reviewed by Buckwalter & King, who noted strain-specific differences in the absence/presence of 12 distinct carbohydrate transporters across pneumococcal genomes, from 7 independent studies, including those coding for: cellubiose, mannitol, mannose, glucosamine, glucose, sucrose, gentiobiose, N-acetylneuraminic acid, N-acetylmannosamine and arbutin (Buckwalter & King 2012). As well as transporter genes, variable distributions in the presence of metabolic genes have also been noted. The sialidase enzyme encoded by the NanC gene in *S. pneumoniae,* for example, (the structure of which confers a strong preference for the sialosides containing the Neu5Ac form of sialic acid versus those that contain Neu5Gc) is found in only 51% of strains (Pettigrew *et al*. 2006). Linke *et al* recently showed that strain-to-strain variation in the ability to utilise different lengths of fructooligosaccharide chains is determined by diversity at the *sus* transporter locus, with 60-79% of pneumococci able to utilise the fructooligosaccharide inulin (Linke *et al*. 2013).

Secondly, even if two different metabolic types have identical substrate repertoires, they may be able to co-exist within a host through binding different host structures to access such resources (Figure 3.1b). For example, variants of family 98 glycoside hydrolases involved in fucose utilisation in *S. pneumoniae* bind distinct host carbohydrate antigens, showing selectivity for either Lewis or group A/B antigens (Higgins *et al*. 2009). Similarly, meningococci encode several surface receptors to acquire iron or haem from a number of specific iron-binding proteins in the host, including haemoglobin, lactoferrin and transferrin, and there is a variable distribution of such receptors among meningococci. The haemoglobin receptor Hmbr, for example, is significantly over-represented in invasive isolates (Harrison *et al*. 2009), and the iron transporter FetA is deleted from a minority of strains (Marsh *et al*. 2007).

Thirdly, co-colonising bacteria may occupy distinct metabolic niches if they have different rates of uptake for particular resources, and/or different utilisation hierarchies for their energy sources (Figure 3.1c). Even if two strains utilise an identical repertoire of nutritional resources, ecological competition could be reduced if they require different amounts of particular substrates (for example, if metabolic type 1 required a large amount of substrate A but only a minimal amount of substrate B, and metabolic type 2 required only a small amount of substrate A but a lot of substrate B). Bidossi *et al.* used a functional genomics approach to demonstrate strain-specific differences among pneumococci in the ability to absorb a wide range of carbohydrate resources (Bidossi *et al*. 2012). As well as variable distributions in the specific transporters and uptake systems present, the relative intensity at which different substrates were fermented was observed to differ between strains. Furthermore, a transcriptional analysis by Pagliarulo *et al.* identified fourfold differences differences in expression of the *gdhA* gene between different hyperinvasive lineages of *N. meningitidis*, which is involved in ammonia assimilation (Pagliarulo *et al*. 2004).

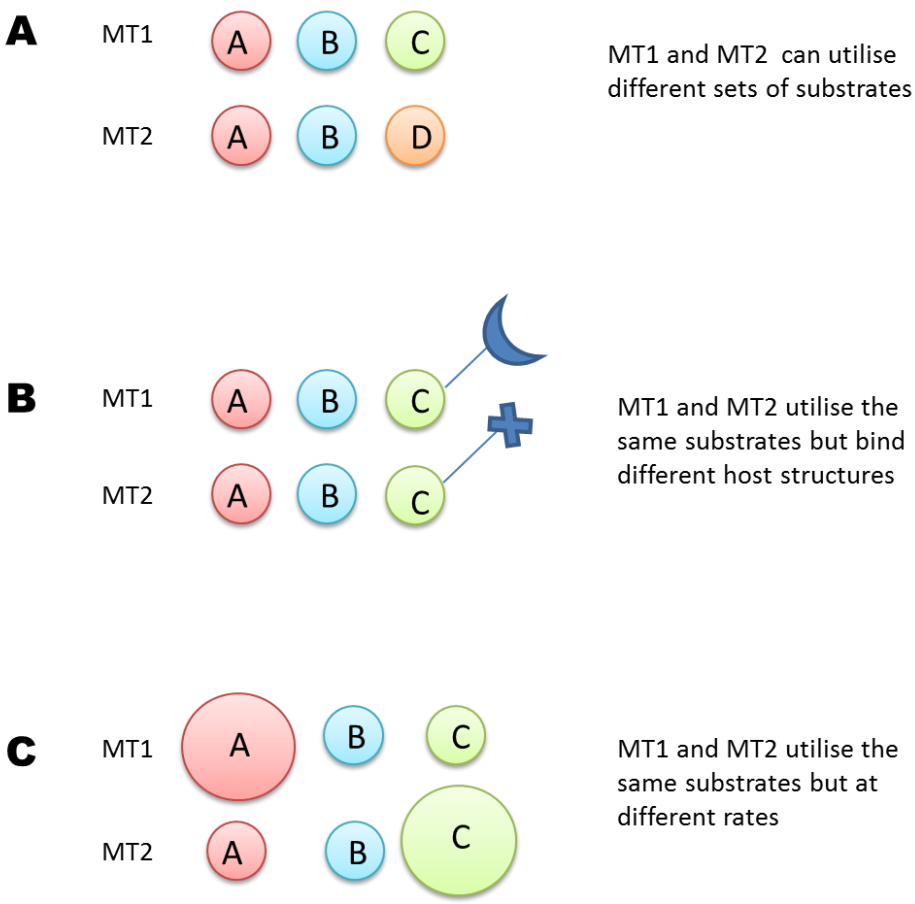


Figure 3.1 Schematic representation of ways through which bacteria with different metabolic profiles (MT1 and MT2) are able to reduce ecological competition and co-infect a host.

The representation in the model of ecological competition - by a reduction in the rate of acquisition in already-colonised hosts - is supported by longitudinal studies of pneumococcal colonisation (Lipsitch *et al*. 2012; Auranen *et al*. 2010). In Denmark, Auranen *et al.* showed that the clearance rate of a given strain was not affected by simultaneous carriage of other strains, but the acquisition of strains in already-colonised hosts was weak. Similarly, the results of a longitudinal study in Kenya showed that the rate of acquisition of pneumococcal strains was higher for non-carriers compared to carriers (Lipsitch *et al*. 2012). Results from experimental data, in which the carriage of vaccine-type strains in mice was shown to inhibit

the subsequent colonisation of non-vaccine type strains, are also consistent with this assumption (Lipsitch *et al*. 2000). To our knowledge, similar studies for *N. meningitidis* have not been carried out.

It should be emphasised that our aim is not to fit the model to data and quantitatively reproduce observed patterns, but to gain a conceptual understanding of the qualitative patterns of diversity observed. We therefore explore a range of parameters of R_0 and σ. The range of values used, though do encompass those of published estimates (e.g. Trotter *et al*. 2005; Sleeman *et al*. 2006).

3.4 Methods

3.4.1 Full SIR Compartmental Frameworks

3.4.1.1 Two-Locus Two-Allele Framework

Alleles of metabolic genes are referred to as metabolic types (MTs), and antigenic genes are referred to as antigenic types (ATs). Strains were defined by one of two ATs (a or b) and one of two MTs (1 or 2), thus yielding a possible four strains in total (a1, a2, b1, b2). A flow diagram of the model is shown in Figure 3.2. The host population is divided into a susceptible class (S), ten infected classes (Y_{a1}, Y_{a2}, Y_{b1}, Y_{b2}; Y_{a1b2}, Y_{a2b1}; V_{a1}, V_{a2}, V_{b1}, V_{b2}), and three immune classes (Z_a, Z_b; Z_{ab}).

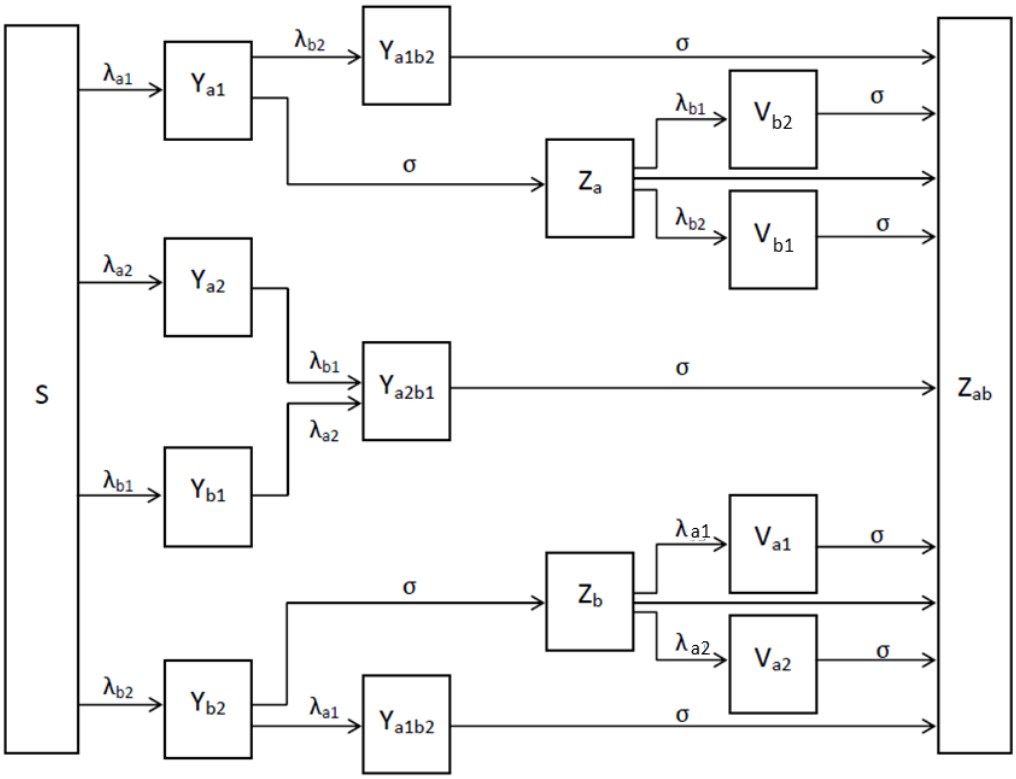


Figure 3.2. Flow diagram of the non-overlapping two-locus two-allele model. *For clarity, the host average death rate is not shown.*

The rate of change in the host population currently infected with strain *a1*, for example, is given by

$$\frac{dY_{a1}}{dt} = \lambda_{a1}S - Y_{a1}(\lambda_{b2} + \sigma + \mu) \qquad (3.1)$$

Where the average host death rate is conferred by μ, and the rate of loss of infectiousness occurs at a rate σ (with the duration of infection therefore corresponding to 1/ σ). The force of infection of strain *a1*, λ_{a1}, is given as the product of the strain's total prevalence in the infected population and its transmission coefficient (its likelihood of being transferred from an infected to a susceptible host, β_i). Thus $\lambda_{a1} = \beta_{a1}(Y_{a1}+Y_{a1b2}+V_{a1})$ where Y and V refer to primary and secondary infections with a1, and Y refers to a co-infection period with strain *b2*. Recovery

leads to antigenic-specific immunity, such that individuals in Z_a can no longer be infected by a strain of antigenic variant *a*, and individuals in Z_{ab} are immune to *a* and *b*.

Additionally, individuals who are currently infected by a particular MT cannot be co-infected by a strain with the same MT due to ecological (rather than immunological) interference. After a fully susceptible host becomes infected, they may become infected with another strain expressing the opposite AT prior to clearance with the current infection, thus entering a co-infected category before gaining immunity to both ATs, or they might clear the infection and become infected at a separate period (V categories, Figure 3.2). Thus a susceptible individual (S) who becomes infected with strain *a1* (Y_{a1}), for example, can either: (i) become co-infected with strain *b2* (Y_{a1b2}) prior to clearance with infection by *a1* and move directly to Z_{ab}, or (ii) clear the infection and gain immunity to antigenic type *a* (Z_a), before later becoming infected with *b1* (V_{b1}) or *b2* (V_{b2}). Through either route, the individual can ultimately gain immunity to both *a* and *b* (Z_{ab}). The basic reproductive number, R_0, of strain *a1* (represented as the average number of secondary cases produced by one infectious case at the beginning of an epidemic, i.e. in a totally susceptible population) is approximated as β_{a1}/σ.

3.4.1.2 Two-Locus Three-Allele Framework

The non-overlapping two-locus two-allele model was extended to a three-allele framework, with three ATs (a, b or c) and three MTs (1, 2 or 3), comprising nine strains in total (a1, a2, a3, b1, b2, b3, c1, c2 and c3). The model consisted of 79 compartments (Figure 3.3). Seven differential equations were required to account for the proportion of the population in immune compartments, and 72 equations required for those in infected compartments.

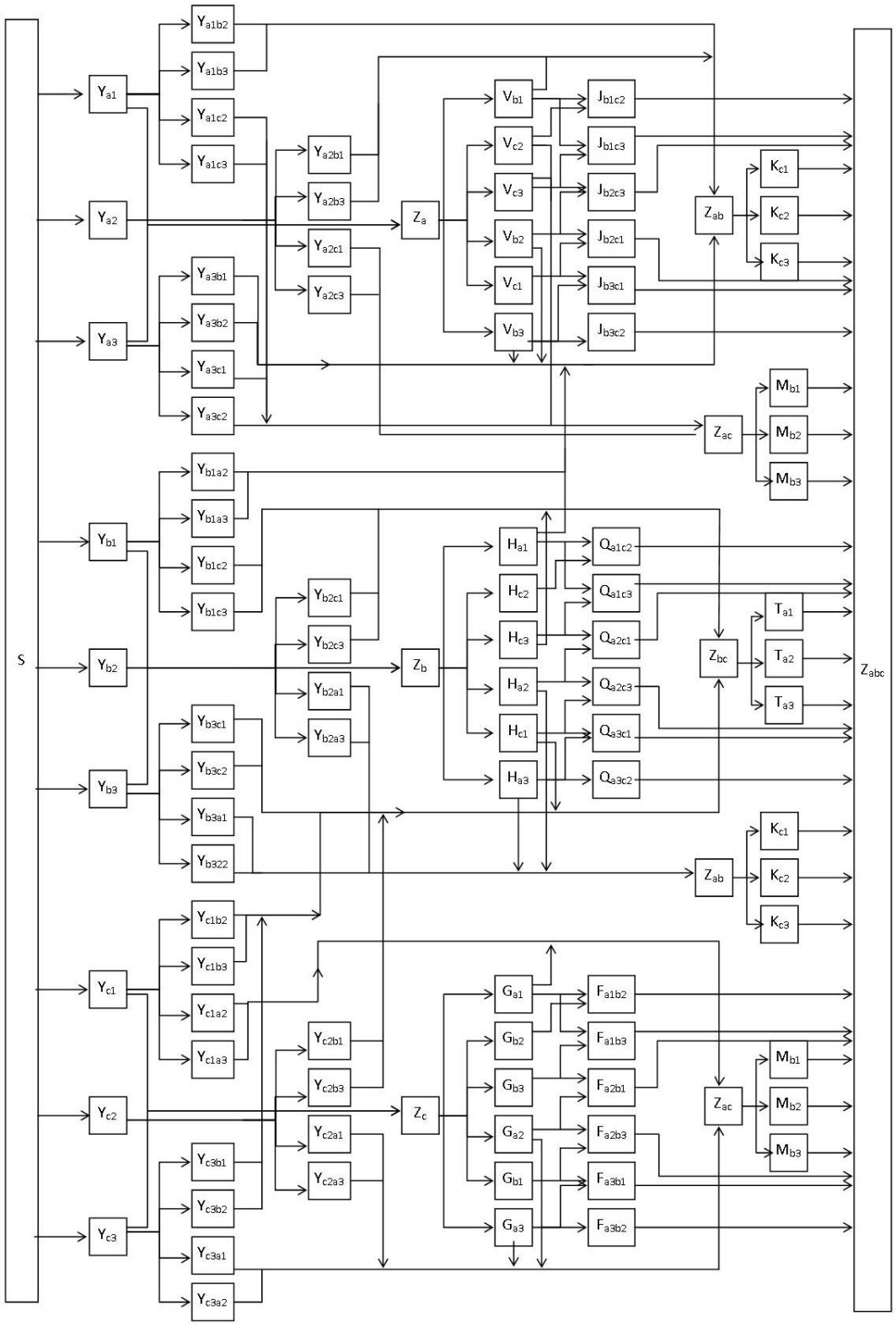


Figure 3.3. Flow diagram of the non-overlapping framework with three ATs and three MTs. *The host population is divided into one susceptible class, 72 infected classes, and 7 immune classes. For clarity, a number of parameters are not shown.*

3.4.2 Simplified Generalised Frameworks

The incorporation of 66 additional equations into the two-locus three-allele from the two-locus two-allele model highlights the challenge in the field of infectious disease modelling of the large increase in dynamic variables required for the addition of extra strains into a network (Gog & Grenfell 2002). Frameworks characterised by overlapping compartments, in which the infected individuals are embedded within the recovered population, and co-infected individuals are not modelled as a distinct compartment frameworks (Gupta *et al*. 1998; Callaghan *et al*. 2008), are able to circumvent this problem to some extent. In this section, we show the derivation of the model formulae from the full, non-overlapping SIR framework, and subsequently explore the effects of collapsing these compartments on the outcome of the model through comparing the results of both frameworks using numerical simulations in four and six strain models.

3.4.2.1 Derivation of the Simplified Two-Locus Two-Allele Framework

Let z_a contain all individuals who have been infected with antigenic type *a*, and y_{a1} contain all individuals currently infected with *a1*. Referring to Figure 3.3, in which the shaded area corresponds to z_a, we can write:

$$\frac{dz_a}{dt} = \lambda_{a1}(1 - z_a - Y_{b1}) + \lambda_{a2}(1 - z_a - Y_{b2}) - \mu z_a \qquad (3.2)$$

$$\frac{dy_{a1}}{dt} = \lambda_{a1}(1 - z_a - Y_{b1}) - \sigma y_{a1} \qquad (3.3)$$

Since $y_{a1} = Y_{a1} + Y_{a1b2} + V_{a1}$, we may obtain equation (3.3) by adding:

$$\frac{dY_{a1}}{dt} = \lambda_{a1}S - (\sigma + \lambda_{b2})Y_{a1}$$

$$\frac{dY_{a1b2}}{dt} = \lambda_{a1}Y_{b2} + \lambda_{b2}Y_{a1} - \sigma Y_{a1b2}$$

$$\frac{dV_{a1}}{dt} = \lambda_{a1}Z_b - \sigma V_{a1}$$

This gives:

$$\frac{dy_{a1}}{dt} = \lambda_{a1}(S + Y_{b2} + Z_b) - \sigma y_{a1}$$

which is equivalent to (3.3) since $S + Y_{b2} + Z_b = 1 - z_a - Y_{b1}$

We can write $(1 - z_a)(1 - y_{b1}) = 1 - z_a - Y_{b1}$ by making the approximation $(1 - z_a)y_{b1} = Y_{b1}$. Therefore, the full SIR system may be represented as:

$$\frac{dz_a}{dt} = (\lambda_{a1}(1 - y_{b1}) + \lambda_{a2}(1 - y_{b2}))(1 - \mathrm{z_a}) - \mu \mathrm{z_a} \tag{3.4}$$

$$\frac{dy_{a1}}{dt} = \lambda_{a1}(1 - y_{b1})(1 - \mathrm{z_a}) - \sigma_{a1}\mathrm{y_{a1}} \tag{3.5}$$

Equations for other strains follow a similar form, and can be generally written for *n* alleles at the antigenic loci (i =*a,b,c*…*n*), and *m* (j =1,2,3…m) alleles at the metabolic locus as:

$$\frac{dz_i}{dt} = \sum_{j=1}^{m}(\lambda_{ij}(1 - z_i)\prod_{w=1}^{n}(1 - \mathrm{y}_{wj})) - \mu \mathrm{z_i} \tag{3.6}$$

$$\frac{dy_{ij}}{dt} = \lambda_{ij}(1 - z_i)\prod_{w=1}^{n}(1 - y_{wj}) - \sigma_{ij}y_{ij} \quad (3.7)$$

The force of infection (λ_{ij}) of strain *ij* is determined by the product of the transmissibility of the MT and the strain's total prevalence in the infected population ($\lambda_{ij} = \beta_j y_{ij}$). The basic reproduction number of a strain, still determined by its MT, is $R_0^{MT} = \beta_j/\sigma_{ij}$. It is assumed that upon infection the host enters the immune and infectious categories concurrently, and all hosts develop lifelong immunity. Therefore individuals are removed from the immune classes at the death rate. The same assumptions of strain-specific immunity and ecological competition apply; strains are not able to co-infect hosts already infected with the same AT and MT. The modified model can be represented by a Venn Diagram, shown in Figure 3.4.

The dynamics of those individuals that are immune to any strain *i* are given by:

$$\frac{dw}{dt} = \sum \lambda_i (1 - w) - \mu w \quad (3.8)$$

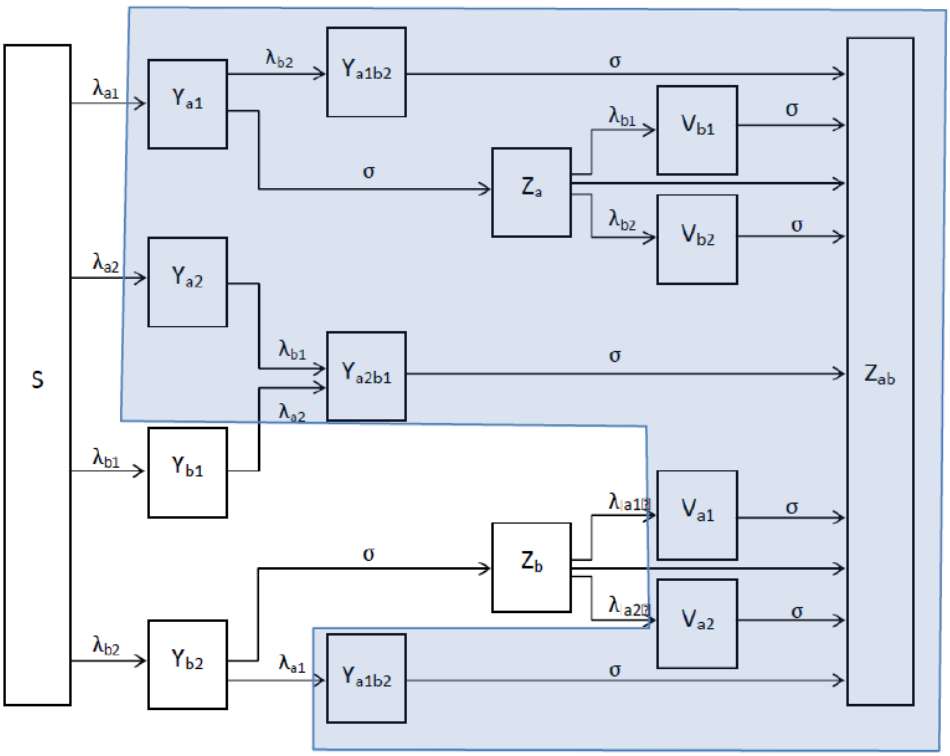


Figure 3.3. The shaded area corresponds to Z_a, i.e. all individuals who have been infected with a.

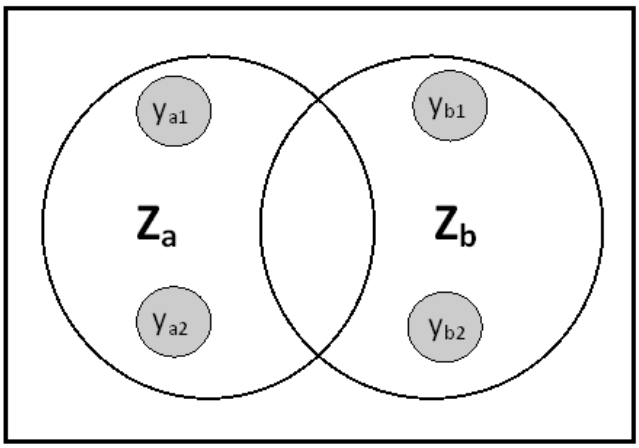


Figure 3.4. Venn Diagram of the overlapping two-locus two-allele model.

3.4.2.2 Two-Locus Three-Allele Framework

The equations governing the dynamics of the system are consistent with those of the simplified two-locus two-allele model (equations 3.6 and 3.7), but incorporate the additional AT and MT. A schematic diagram of the model is shown in Figure 3.5. The differences in transmissibility between the MTs were kept constant (i.e. $\beta_1-\beta_2=\beta_2-\beta_3$).

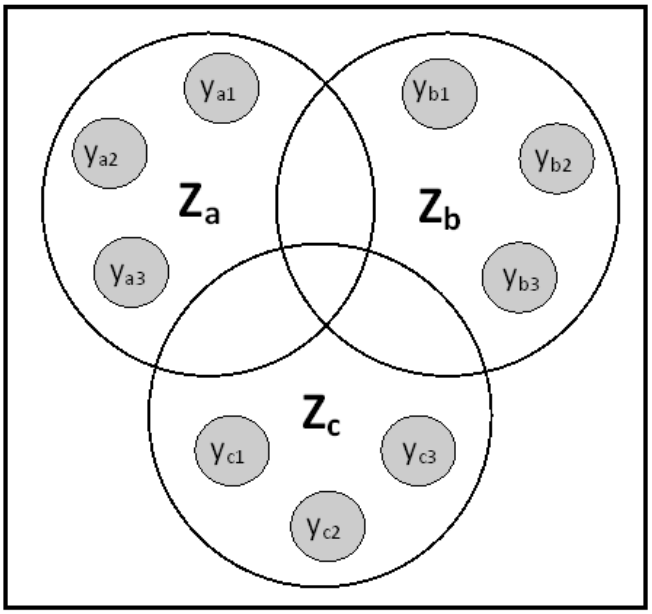


Figure 3.5. Venn Diagram of the overlapping two-locus three-allele model.

3.4.4 Incomplete Ecological Competition

The simplified two-allele model of AT and MT (equations 3.4 – 3.5) may be extended to include the effects of incomplete ecological competition, by introducing a parameter ψ (0≤ψ≤1) which specifies the degree of resistance against co-infection by the same MT:

$$\frac{dy_{a1}}{dt} = \lambda_{a1}(1 - \psi y_{b1})\,(1 - z_a) - \sigma_{a1} y_{a1} \tag{3.9}$$

$$\frac{dz_a}{dt} = (\lambda_{a1}(1 - \psi y_{b1}) + \lambda_{a2}(1 - \psi y_{b2}))(1 - z_a) - \mu z_a \tag{3.10}$$

When ψ = 1, co-infection with strains possessing the same MT is not permitted. Consistent with the models of full antigenic and ecological competition, strains are not able to co-infect hosts already infected with the same AT and are freely able to infect strains with a different MT.

3.4.5 Incomplete Immunological Competition

The assumption of complete strain-specific immunity can be relaxed through adding the variable γ (0≤ γ ≤1). The equations governing the dynamics of the model for strain type *a1* are as follows:

$$\frac{dz_a}{dt} = \big(\lambda_{a1}(1 - \psi y_{b1}) + \lambda_{a2}(1 - \psi y_{b2})\big)(1 - \gamma(z_a - y_{a1}) - y_{a1}) - \mu z_a \tag{3.11}$$

$$\frac{dy_{a1}}{dt} = \lambda_{a1}(1 - \psi y_{b1})\,(1 - \gamma(z_a - y_{a1}) - y_{a1}) - \sigma_{a1} y_{a1} \tag{3.12}$$

The parameter γ ($0 \leq \gamma \leq 1$) is a measure of level of strain-specific immunity. When $\gamma = 1$, an individual who has previously been infected with strain *a1* cannot subsequently be infected with any strain of antigenic type *a*. When $\gamma = 0$, an individual previously infected with strain *a1* remains fully susceptible to all strains of antigenic type *a*. Thus, when $\gamma = 1$, we recover equations 3.9 and 3.10 with full immunity while, at $\gamma = 0$, the only form of competition in the system is ecological (i.e. through sharing of metabolic type). Both γ and ψ can be varied independently.

3.4.6 Attributing R_0 to the Antigenic Type and Modifying the Duration of Infection

The transmissibility (β) of a strain can be attributed to its AT as opposed to its MT. Thus, referring back to the general framework outlined in Section 3.3.2:

$$\frac{dz_i}{dt} = \sum_{j=1}^{m} (\lambda_{ij} (1 - z_i) \prod_{w=1}^{n} (1 - y_{wj})) - \mu z_i \qquad (3.6)$$

$$\frac{dy_{ij}}{dt} = \lambda_{ij} (1 - z_i) \prod_{w=1}^{n} (1 - y_{wj}) - \sigma_{ij} y_{ij} \qquad (3.7)$$

The force of infection can be written as $\lambda_{ij} = \beta_{ij} y_{ij}$ and $R_0 = \beta_{ij}/\sigma_{ij}$ with $\beta_{ij} = f(\beta_i, \beta_j)$ and $\sigma_{ij} = g(\sigma_i, \sigma_j)$, where *i* indicates it is a property associated with AT, and *j* with MT. Thus, both the transmissibility and duration of infection can be specified for a given AT or MT. We use the term R_0^{AT} ($= \beta_i/\sigma_i$) to denote the R_0 of all strains with that AT when all MT are identical in β

and σ, and R_0^{MT} (= β_j/σ_j, or β_j/σ_{ij}) to denote the R_0 of all strains with that MT when all AT are identical in β and σ.

3.5 Results

We confirm by simulation that the simplified, overlapping frameworks deliver the same result as the full SIR models, in both the two-locus two-allele and two-locus three-allele systems (see Appendix). Thus, the simplified frameworks are sufficiently accurate approximations of the full SIR system, and the results described here and in the remaining chapters are based on such frameworks. Unless otherwise specified, the results described in the following sections are derived from model frameworks in which R_0 is defined by the MT of a strain, and the rate of loss of infectiousness is kept constant across all strains.

3.5.1 Full Ecological and Immunological Competition

3.5.1.1 Two-locus two-allele system

The behaviour of the MTs depended upon the differences in their R_0s. At low differences, non-overlapping associations between ATs and MTs are observed, in which individual ATs are associated with unique MTs (Figure 3.6a). The predominant strains under such parameters are a1 and b2, with the remaining strains a2 and b1 declining in prevalence. When the differences in transmissibility between MTs surpass a specific threshold, the more transmissible MT (in this case, MT1) displaces the other from the host population through competitive exclusion (Figure 3.6b). Both ATs are still allowed to co-exist, such that discordant combinations between ATs and MTs are no longer observed, as the most transmissible MT dominates both

ATs. Figure 3.7 demonstrates that the behaviour of the model is influenced by the rate of loss of infectiousness of the strains. Non-overlapping associations between AT and MT are observed less frequently in systems characterised by strains with shorter infectious periods ($1/\sigma$). Systems comprising strains with longer durations of infection show non-overlapping associations between AT and MT at higher differences in R_0 between MTs. This is because strains with lower σ experience longer periods of infection, and thus have an increased window of opportunity for co-infection by a second strain.

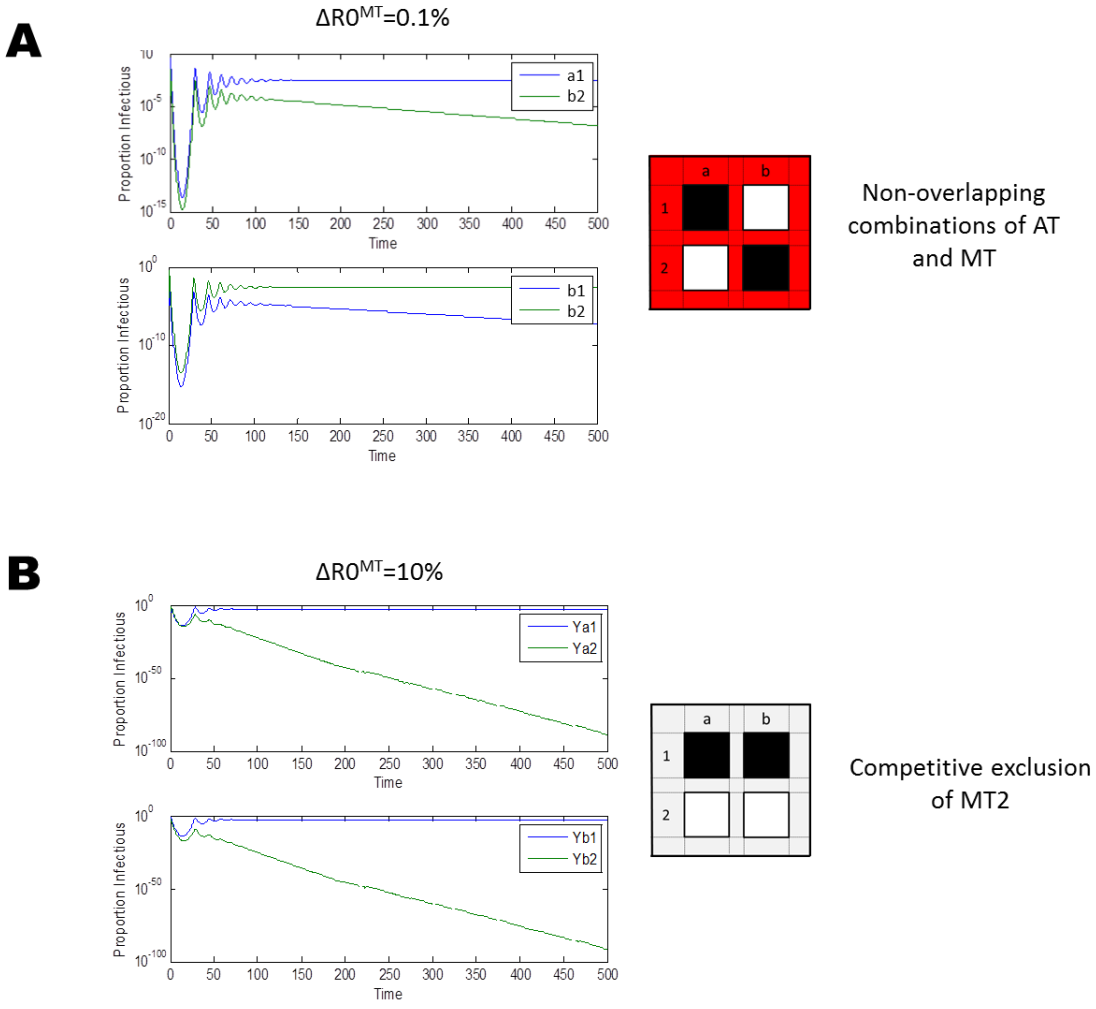


Figure 3.6. Dynamics and population structures generated by the two-locus two-allele (four strain) framework. *(A) Small differences in R_0 between MTs give rise to a population structure comprising non-overlapping associations of ATs and MTs. (B) When the differences in R_0 between MTs are sufficiently large, the fittest MT competitively excludes the other. ($\beta_1 > \beta_2$; $\beta_1=4.5$; $\sigma=3$; $\mu=0.02$). The population structures are represented in terms of associations between antigenic (a, b) and metabolic (1, 2) alleles.*

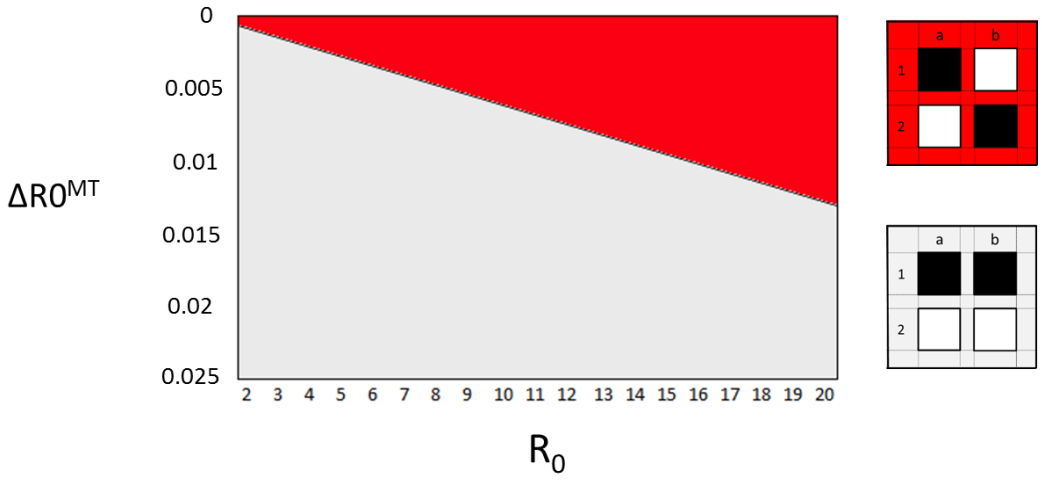


Figure 3.7. Parameter space over which non-overlapping associations between AT and MT (red) and competitive exclusion (grey) occur in a two-locus two-allele system, for R_0 values of 2 to 20. *($\beta_{1>}\beta_2$; β_1=30; μ=0.02).*

3.5.1.3 Two-locus three-allele systems

The effects of differences in R_0 between MTs on population structuring in both two-allele (four strain) and three-allele (nine strain) models are generally consistent. At low differences in R_0^{MT} between strains, all six MTs and ATs are permitted to co-exist, but emerge as three strains characterised by non-overlapping associations of AT and MT (Figure 3.8a). After $\Delta R0^{MT}$ passes a primary threshold, the least transmissible MT (MT3) decays in prevalence to zero (Figure 3.8b). As the differences in R_0 are increased further, the MT with the highest fitness (MT1) displaces both MTs (Figure 3.8c). The behaviour of the model is also influenced by the rate of loss of infectiousness of the strains (Figure 3.9). Consistent with the results of the two-allele model, the exclusion thresholds in ΔR_0^{MT} are higher among strains with shorter infectious periods; strains with longer durations of infection show non-overlapping associations between AT and MT at higher differences in R_0 between MTs.

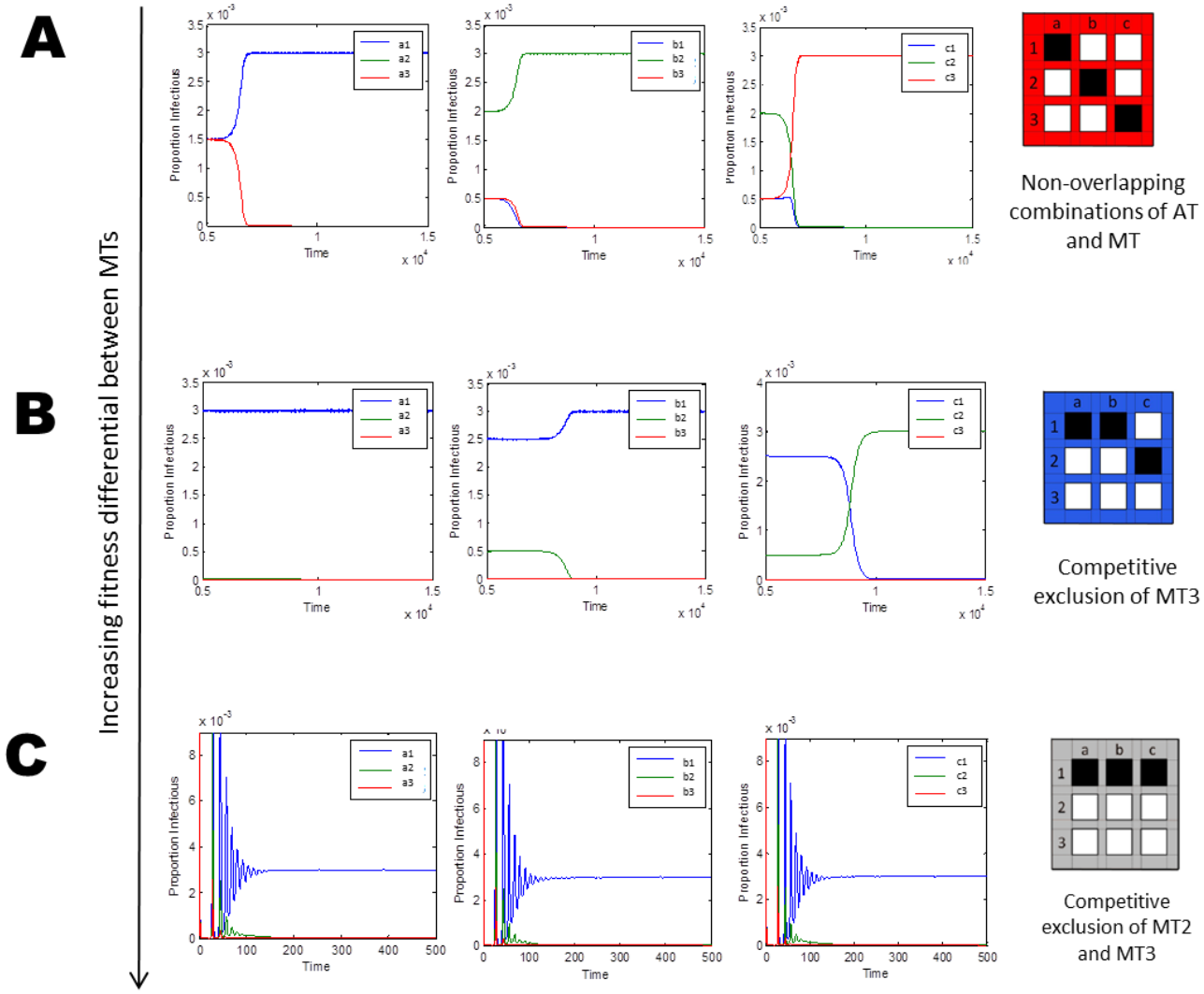


Figure 3.8. Dynamics and population structures generated by the two-locus two-allele (four strain) framework. *(A) When all MTs have equal fitness, the discordant strains a1, b2 and c3 are dominant. ($\beta_1=\beta_2=\beta_3=20$; $\sigma=5$; $\mu=0.02$) (B) At intermediate differences in fitness between MTs ($\Delta R_0^{MT}=0.5\%$), the least transmissible MT is excluded, leaving a1, b1 and c2 ($\beta_1>\beta_2>\beta_3$; $\beta_1=20$; $\sigma=5$; $\mu=0.02$). (C) Under stronger interstrain competition ($\Delta R_0^{MT}=1\%$), only the most transmissible MT remains in the population, in strains a1, b1 and c1 ($\beta_1>\beta_2>\beta_3$; $\beta_1=20$; $\sigma=5$; $\mu=0.02$).*

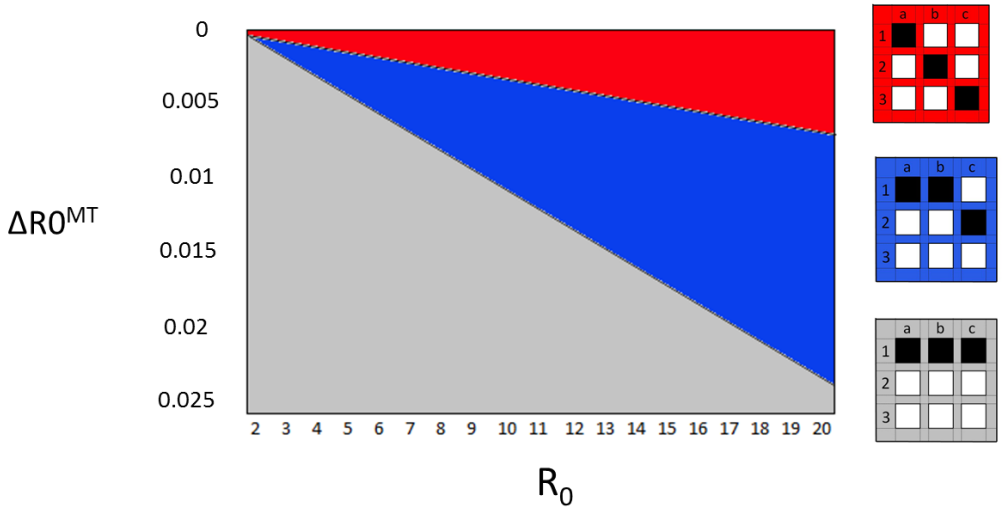


Figure 3.9. Exclusion thresholds of MTs in systems with varying R_0 values. *R_0 was varied by changing the rate of loss of infectiousness (σ). ($\beta_1>\beta_2>\beta_3$; $\beta_1=30$; $\sigma=15$ to 1.5; $\mu=0.02$).*

3.5.2 Incomplete Ecological Competition

Numerical simulations of the model reveal two main classes of strain patterns: (i) non-overlapping associations between AT and MTs; and (ii) the dominance of one MT among both ATs. The particular patterns observed depend on both ΔR_0^{MT} and the strength of ecological competition (ψ) (Figure 3.10).

When MTs are assigned equal transmission fitnesses, non-overlapping associations of MT and AT are observed whenever ecological competition is present (for all values of ψ exceeding 0). When ψ is zero, one MT dominates both ATs. At low differences in R_0 between MTs, the relaxation of ecological competition suppresses the maintenance of strain structuring, and competitive exclusion of the least transmissible MT is more frequently observed. Conversely, increasing ecological competition enhances the ability of strains to form non-overlapping

combinations. At higher levels of interstrain competition, the least transmissible MT is invariably outcompeted regardless of ψ.

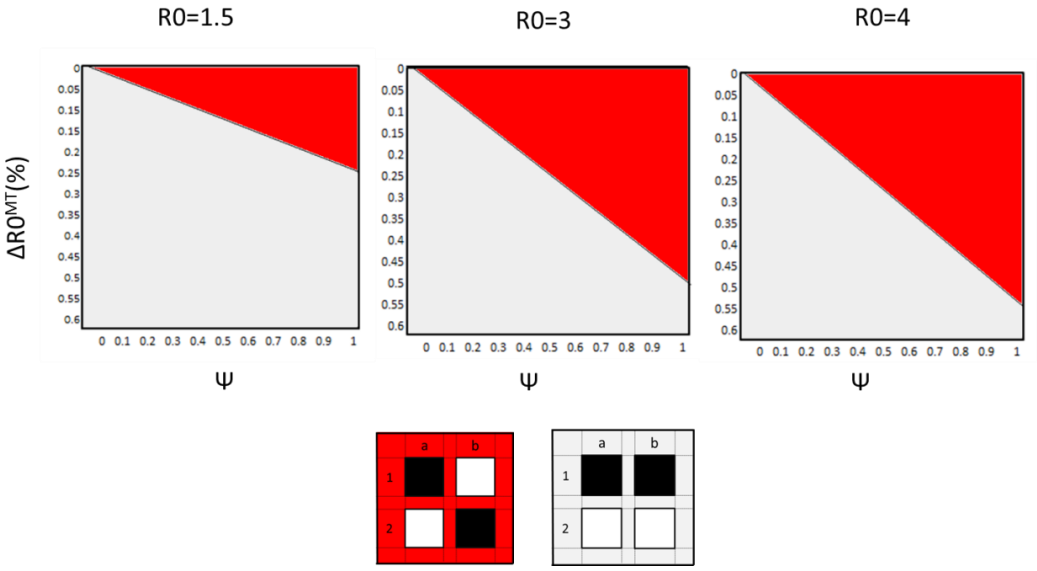


Figure 3.10. The effects of ecological competition (Ψ) and differences in MT transmissibility (ΔR_0^{MT}) in a two-locus two-allele system. *The plots show the predominant strains at equilibrium ($\beta_1>\beta_2$; $\beta_1=4.5$, $\sigma=3$; $\beta_1=9$, $\sigma=3$; $\beta_1=12$, $\sigma=3$; $\mu=0.02$).*

3.5.3 Incomplete immunological competition

Numerical simulations were investigated in a two-locus two-allele (four strain) framework. Predominantly three modes of behaviour can be observed (Figure 3.11): competitive exclusion of the least transmissible MT (grey); non-overlapping associations between AT and MT (red); and an intermediate behaviour, comprising two strains of the fitter MT and one strain of the less fit MT (blue). Additionally, at very low values of γ, co-existence of all strain combinations is observed (white). The behaviour of the model depends on the magnitude of ΔR_0^{MT} and strength of immunological competition, in addition to the R_0 value of the strains.

When the transmission fitness among MTs is equal, non-overlapping associations between AT and MT are observed over all values of γ above 0.05. Beyond a given threshold in ΔR_0^{MT}, competitive exclusion of the least transmissible MT occurs. Between these two areas of parameter space, an intermediate outcome can be observed, in which the least transmissible MT has not yet been excluded. The occurrence of these behaviours depends on γ. Non-overlapping associations between AT and MT are more likely to be observed at lower levels of γ. These outcomes are also influenced by the R_0 of the strains (determined by σ). Systems comprising strains with higher R_0 values are able to maintain non-overlapping associations between AT and MT at higher levels of ΔR_0^{MT} and γ.

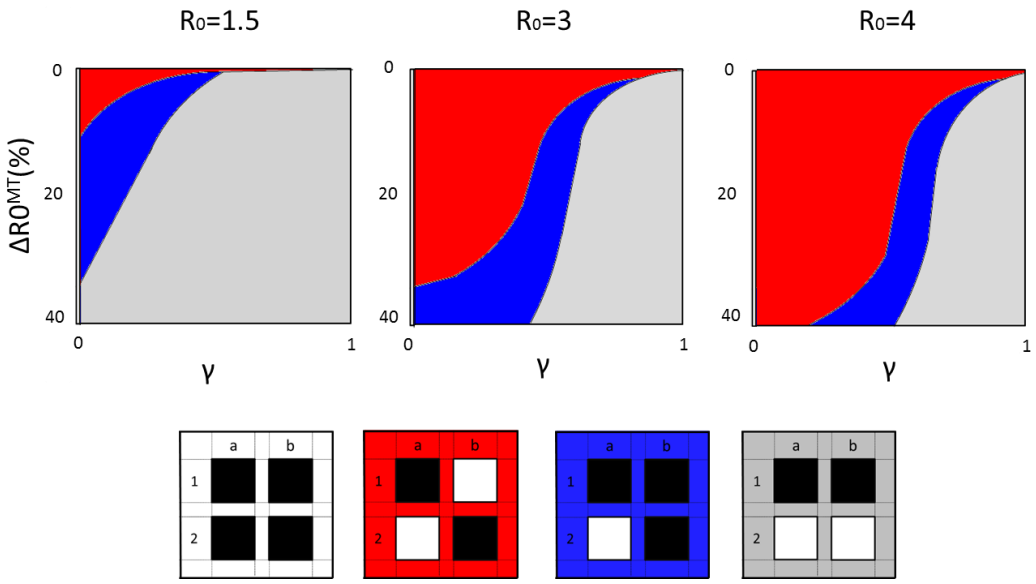


Figure 3.11. The effects of immunological competition (γ) and differences in MT transmissibility ($\Delta R0^{MT}$) in a two-locus two-allele system. *The plots show the predominant strains at equilibrium ($\beta_{1>}\ \beta_{2;}\beta_1$=4.5, σ=3; β_1=9, σ=3; β_1=12, σ=3; μ=0.02).*

3.5.4 Combining Incomplete Ecological and Immunological Competition

3.5.4.1 Two-locus two-allele framework

At equal transmissibilities between MTs (Figure 3.12 – top left panel), all combinations of AT and MT coexist when γ and ψ are low (white box, Figure 3.12). As γ and ψ increase, the least transmissible MT is eliminated and the pathogen population becomes structured into non-overlapping combinations of AT and MT. When there are small fitness differences between MTs, competitive exclusion occurs at low levels of ecological competition (Ψ) and high levels of immunological competition (γ). As ΔR_0^{MT} increases, the strains with the highest fitness become more prevalent over the parameter space.

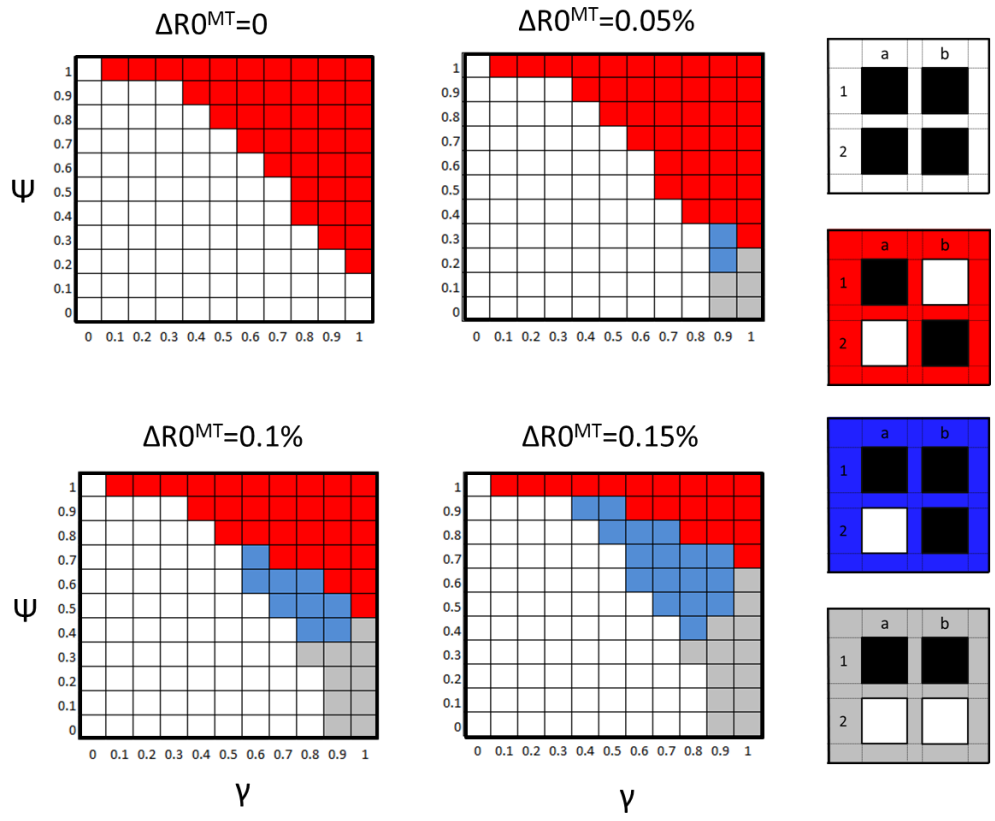


Figure 3.12. The strain combinations (marked in black) within population structures (represented by different colours) which arise under varying strengths of immunological (γ) and ecological (Ψ) competition, at varying differences in R_0 between the MTs of strains ($\Delta R0^{MT}$), in a two-locus two-allele framework. *($\beta_1 > B_2 > \beta_3$; $\beta_3 = 4.5$; $\sigma = 3$; $\mu = 0.02$).*

3.5.4.2 Two-locus three-allele framework

A variety of population structures can be observed in the nine strain framework under a range of strengths of ecological and immunological competition. In parallel to the four strain framework, all combinations of AT and MT co-exist at low values of γ and ψ (white box, Figure 3.13). Exclusion of both MT2 and MT3 (grey box, Figure 3.13) occurs at high γ and low ψ. As immunological competition is relaxed, a larger number of ATs are permitted to circulate: at lower levels of ψ, the grey structure (three strains) progresses to the cyan structure (six strains), to the white structure (representing co-existence of all nine strain combinations). At higher levels of ψ, the blue structure (three strains), progresses to the purple structure (four strains), to the white structure (all strain combinations) as γ is relaxed. A larger diversity of MTs is generally observed at higher levels of ecological competition, and high levels of γ and ψ are needed for non-overlapping associations of AT and MT to occur. As ecological competition decreases, a partially overlapping structure (purple box, Figure 3.13) arises, in which strain c1 - with the fittest MT - becomes prevalent alongside the discordant associations of a1, b2 and c3. As ΔR_0^{MT} increases, the parameter space in which exclusion or partial exclusion occurs becomes more widespread. When ΔR_0^{MT} is sufficiently high (Figure 3.13, bottom right panel), non-overlapping associations between AT and MT occur only at intermediate levels of immunological competition and high levels of ecological competition.

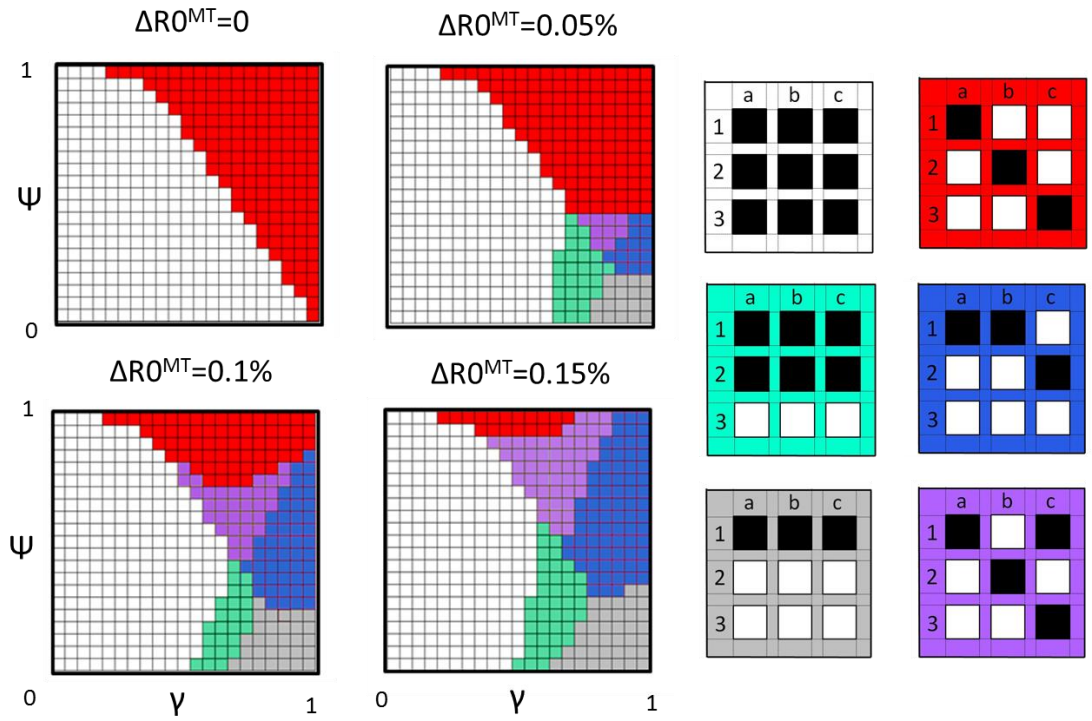


Figure 3.13.Population structures generated by the two-locus three-allele model which arise over a range of strengths of ecological (Ψ) and immunological (γ) competition, at varying differences in R_0 between the MTs of strains (ΔR_0^{MT}). *($\beta_1 > B_2 > \beta_3; \beta_3 = 4.5; \sigma = 3; \mu = 0.02$).*

3.5.5 Attributing R_0 to the Antigenic Type (R_0^{AT})

When differences in fitness between strains are determined by the AT (R_0^{AT}) as opposed to the MT (R_0^{MT}), discordant associations between AT and MT occur at much greater differences in R_0 between strains (Figure 3.14). At full competition, non-overlapping associations can be observed over all differences in R_0 between strains. Decreasing levels of immunological competition resulted in more competitive exclusion. At intermediate levels of γ, non-overlapping associations between ATs and MTs were more likely to occur at lower R_0s. Relaxing ecological competition had no effect on the prevalent strain combinations at equilibrium.

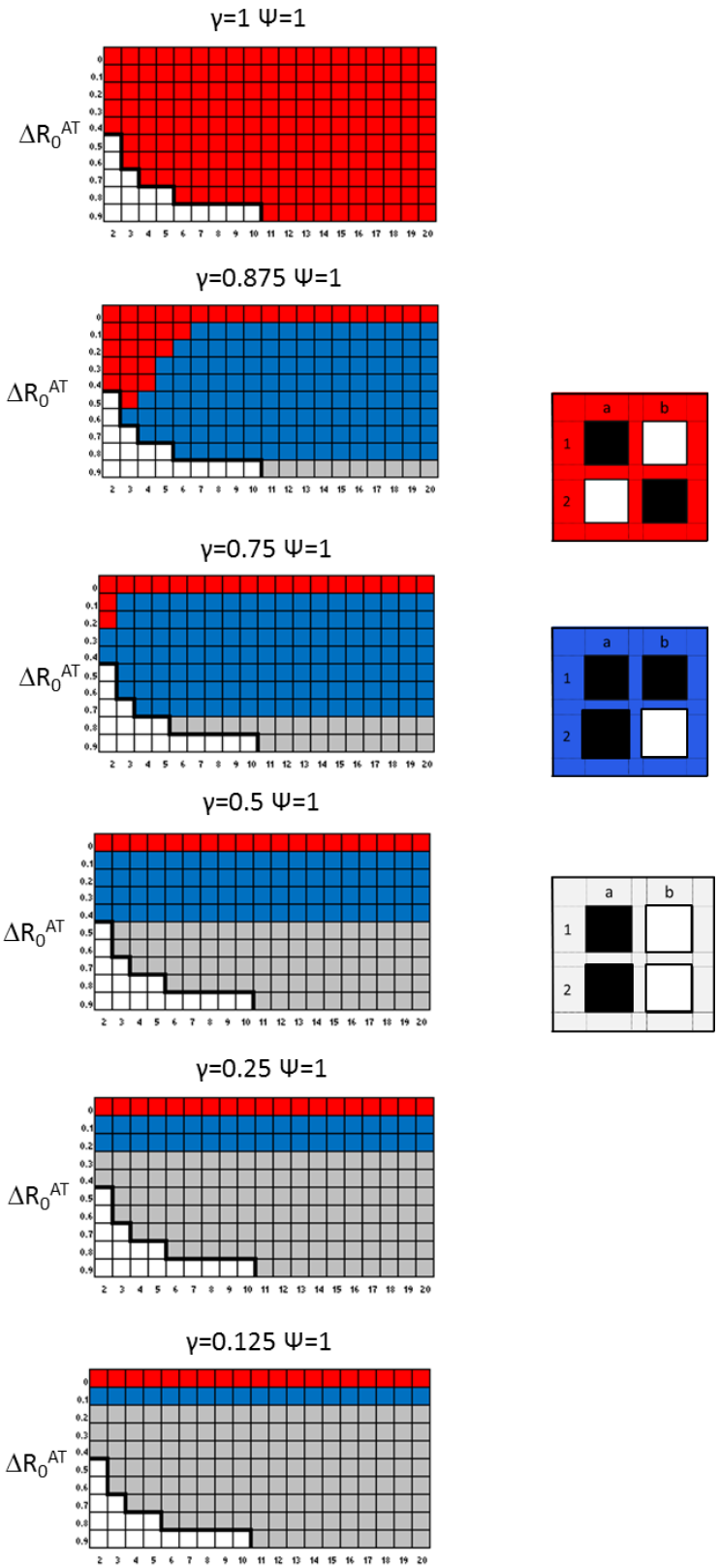


Figure 3.14. The effects of ecological and immunological competition at varying differences in $R0^{AT}$ between strains. *The outcomes yielded when Ψ was relaxed were consistent with that obtained at full ecological competition (the top panel). (β_a = 30; β_b = 3-27; σ is displayed on the x axis).*

3.5.6 Varying the Duration of Infection under Constant R_0

In the results described thus far, the rates of loss of infectiousness have remained constant across all strains in a given simulation. The strains which prevail at equilibrium are at equal frequencies, which is unlikely to mirror the prevalence of pathogen strains observed in real populations. Further, discordant associations between AT and MT can only arise when differences in R_0 between MTs are relatively small. However, it is possible for strains with diverse durations of infection to exhibit discordant associations between AT and MT provided the rates of loss of infectiousness are balanced by their transmissibilities to maintain a similar R_0. Thus, strains with high transmission rates and short durations of infection are able to circulate alongside strains with low transmission rates balanced by longer infection periods. Assigning each AT or MT a distinct duration of infection, balanced by a distinct transmission fitness to maintain similar R_0 values across all strains in the system, results in strains exhibiting different prevalences at equilibrium. The AT or MT with the longest duration of infection exhibits the highest frequency at equilibrium (Figure 3.15).

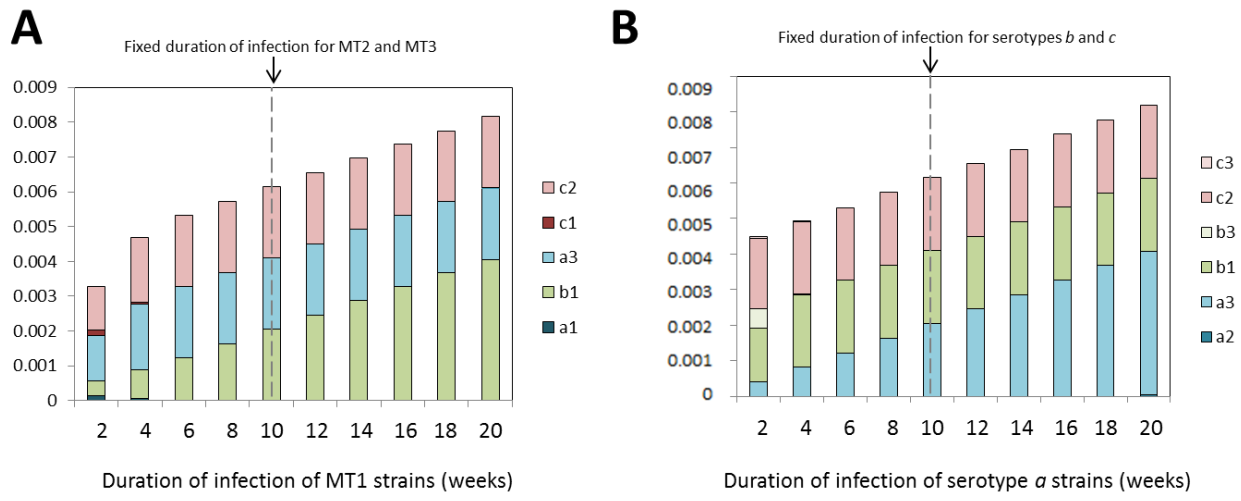


Figure 3.15. Effects of varying the duration of infection under constant R_0. *(A) As the duration of infection of all strains possessing MT1 increases, the relative prevalence of the established strain b1 increases. (B) As the duration of infection of all strains containing serotype a increases, the relative prevalence of strain a3 increases ($\beta_1 > B_2 > \beta_3$; $\beta_3 = 4.5$; $\sigma = 3$; $\mu = 0.02$; $\gamma = 0.75$; $\psi = 0.75$).*

3.6 Discussion

Rather than using a model framework comprising discrete and non-overlapping compartments, as frequently used in epidemiological models (e.g. Trotter *et al*. 2005; Flasche *et al*. 2013; Bottomley *et al*. 2013), we adopted a simplified framework with overlapping compartments, in which the infected population are embedded within the recovered population (Figures 3.4 and 3.5). This approach has been used in a number of published frameworks (Gupta *et al*. 1998; Callaghan *et al*. 2008). Although this type of model framework is able to avoid the large increase in equations and parameters usually associated with adding extra strains into epidemiological frameworks, it was not known how the deviation from the discrete, non-overlapping compartments affects model behaviour and outcome. We constructed full SIR and simplified analogues of both four strain (two-locus two-allele) and nine strain (two-locus three-allele) frameworks, and showed that they produce almost equivalent results, both qualitatively and quantitatively (see Appendix).

We first considered the effects of complete immunological and ecological competition, in which strains with identical ATs and MTs were not able to co-infect the same host. Numerical simulations demonstrated that when differences in transmissibility between MTs (ΔR_0^{MT}) are low, all ATs and MTs co-circulate, but self-assort into discrete strains characterised by unique combinations of AT and MT. Thus, strong immunological competition creates a host population structured into antigenic islands, within which lineages compete ecologically. When the transmission advantage of one MT is sufficiently high, it competitively excludes the other MT(s), and the strains which dominate the population at equilibrium all possess this more transmissible MT, whereas the strains with the less transmissible MTs are at very low frequencies (the dominant strains will be for example, a1, b1 and c1 if MT1 has the highest transmission fitness). These findings are consistent with those of Buckee *et al*. (2008), and

were observed across four and nine strain models, in four separate theoretical frameworks. It is worth noting that cyclical associations between AT and MT, as found by Buckee *et al*., are not observed in the models here because the AT is defined only by one locus (antigenic cycling requires more than one antigenic locus per strain).

As ecological competition between MTs is relaxed, competitive exclusion of the strain with the lowest transmission fitness is more likely to occur; this is perhaps a counter-intuitive result, as one would perhaps expect relaxed levels of competition to promote the co-existence of strains, not their exclusion. In contrast, as immunological competition is relaxed, non-overlapping associations are no longer observed as increasing numbers of ATs are permitted to co-exist. Corresponding results have been concluded from other models of immune selection, which demonstrate the co-existence of many strains at decreased levels of cross-protection (e.g. Gupta *et al*, 1998).

When both ecological and immunological competition are relaxed, non-overlapping associations between AT and MT are still able to arise over a wide range of parameter space when the difference in transmissibility between MTs is low. A range of population structures are observed in the three-allele, nine strain model (Figure 3.13). Indeed, associations between AT and MT in meningococcal, pneumococcal and other bacterial populations are not absolute, and may in some cases correspond to the purple structure, in which the strains are mostly non-overlapping but one AT shares two MTs. Our results therefore suggest that ecological and immunological competition may be important in generating strain structure in pathogen populations, but are not necessarily complete and could be at intermediate or even low levels in some areas of parameter space. A limitation of our framework is the equal strength of strain-specific immunity set across all ATs in the model, as the immune response to particular antigen(s) of some bacterial pathogens may not be uniform across antigenic variants. Strengths of strain-specific immunity have been recorded to vary greatly among different pneumococcal

serotypes, for example. It would be interesting to explore a system with different strengths of strain-specific immunity for different ATs, and the effects such conditions have on population structuring. It may be that the ATs which exhibit strong strain-specific immunity are more likely to form non-overlapping associations with MTs than those exhibiting weak strain-specific immunity.

In contrast, when transmission fitness is dictated by the AT as opposed to the MT, non-overlapping associations are observed even at very large differences in R_0 between strains when ecological and immunological competition are complete. Exact R_0 values are not known for specific strains within pneumococcal or meningococcal populations (an estimate is available for serogroup C meningococci, at 1.36; Trotter *et al*. 2005), so it may be difficult to deduce the differences in R_0 between strains in real populations. Whereas the transmission fitness of strains is difficult to estimate and parameterise from data in real populations, the duration of carriage ($1/\sigma$) is more easily measured and better understood for both *S. pneumoniae* and *N. meningitidis*. A range of durations of colonisation have been recorded for different strains in both pathogens, and it has been suggested that strains may compensate for a particular duration of colonisation by adjusting their transmission fitness (Sleeman et al. 2006; Caugant et al. 2007b). Indeed, when both transmission fitness and duration of infection were simultaneously adjusted for each strain in the model, providing each with an identical R_0, strains with a wide range of duration of infection were able to co-exist in the same population. These results were observed when R_0 was defined by both MT and AT (Figure 3.15). Further, the strains which dominated in each simulation were at different equilibrium frequencies, thus mirroring the differences in prevalence between strains observed in real populations (Figure 3.16).

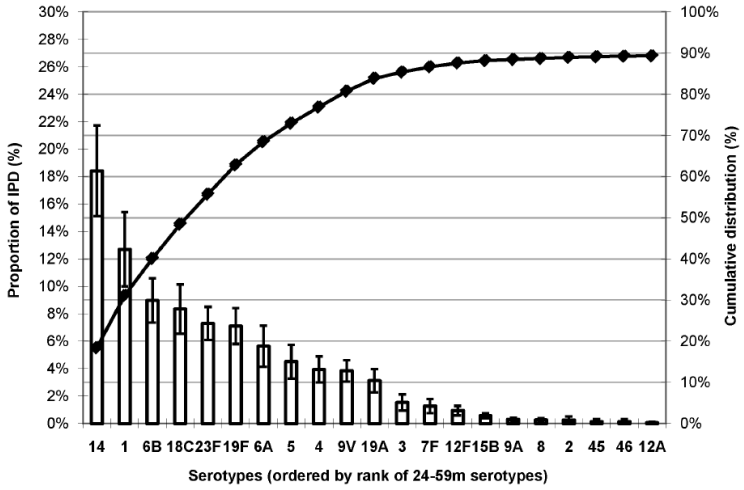


Figure 3.16. Proportion of invasive pneumococcal disease caused by the most common serotypes globally (in children aged 29-54 months) *(taken from Johnson et al. 2010).*

A number of studies have shown differences in colonisation periods between pneumococcal serotypes, consistent with observational and experimental studies (Hogberg *et al*. 2007; Sleeman *et al*. 2006). A possible explanation for the differences in prevalence between serotypes was provided by Weinberger *et al*. (2009), who demonstrated that particular capsule types were more resistant to neutrophil-mediated killing. Those polysaccharide structures which were more metabolically expensive to produce (according to the number of carbons per repeating polysaccharide unit) tended to result in thinner capsules, which were more susceptible to killing by neutrophils. However, differences in the duration of carriage observed among serotypes in different regions suggest that MT may also play a role in determining $1/\sigma$. For example, the average carriage duration of serotype 23F pneumococci was estimated at 9.6 weeks in the Gambia, and 16.7 weeks in the UK (Sleeman *et al*. 2006; Hill *et al*. 2008). In the same longitudinal studies, the carriage duration of serotype 6A pneumococci was estimated at 13.9 weeks and 9.3 weeks in the Gambia and the UK respectively. As for *N. meningitidis*, longitudinal carriage studies have shown that strains with different STs vary widely in their duration of colonisation, thus suggesting that metabolic genes may also influence colonisation period and/or transmissibility (Caugant et al. 1988; Caugant et al. 2007b; Claus et al. 2005).

Although a number of bacterial pathogens show associations between antigenic and metabolic alleles in addition to meningococcal and pneumococcal populations (Table 1.4), it is important to emphasise that the results of the models are only applicable to those pathogens which fulfil a number of conditions, as stipulated by the exact assumptions of the models. As well as showing variation at their metabolic alleles, these pathogens:

(i) must show allelic variation at a minimum of one antigen;
(ii) this antigen(s) must elicit antigenic-specific immune responses;
(iii) these responses must protect to some extent against subsequent infection.

As discussed in Section 3.3.2.1, the antigens PorA and FetA of meningococci, and the capsular polysaccharide of pneumococci seem to fulfil these criteria (to some extent); but whether or not the antigenic variants of the pathogens in Table 1.4 generate protective immune responses is still unclear (see also Section 8.4).

3.7 Summary

The results of the deterministic models presented here suggest that the approximately non-overlapping associations between lineages and antigenic types observed in pneumococcal and meningococcal populations may be ascribed to immunological and ecological competition, in conjunction with relatively small fitness differences between MTs, and higher differences in fitness between ATs. Non-overlapping associations were observed at intermediate to high levels of competition in most simulations, but also emerged at relatively low levels in some areas of parameter space. Whether or not these processes are responsible for the associations in other pathogens depends on a number of specific conditions.

Chapter 4.

Associations between Antigenic and Metabolic Type in *Neisseria meningitidis*

4.1 Public Dissemination of Work

This work was published in the peer-reviewed article:

Eleanor R. Watkins & Martin C. J. Maiden (2012) Persistence of Hyperinvasive Meningococcal Strain Types during Global Spread as Recorded in the PubMLST Database. *Plos One* 7(9): e45349.

For the published article, ERW performed the data analysis; ERW and MCJM wrote the article; and MCJM conceived the study and gave advice.

4.2 Abstract

The models presented in the previous chapter predict that stable, non-overlapping associations between antigenic and metabolic types are found in pathogen populations at a range of strengths of immunological and ecological competition. The stochastic framework by Buckee *et al*. (2008) predicted stable as well as fluctuating associations between metabolic and antigenic type, depending on the exact levels of cross-immunity and differences in R_0 between metabolic types. Buckee *et al*. used a carriage dataset from the Czech Republic to investigate support for the predictions of their model. Here, we aim to ascertain if such patterns are observed across a more extensive sample, using the genotypic information available on the PubMLST database (www.pubmlst.org). We investigate the temporal and geographical patterns of associations among clonal complexes and variants of outer membrane antigens PorA and FetA using a sample of 3460 isolates from 57 countries over a 74 year period. Based on the predictions of the model frameworks presented in this thesis, we would expect to observe

a number of long-lived associations between PorA and FetA variants and clonal complexes. Additionally, according to the results of the stochastic model by Buckee *et al*., we would expect to see a number of shifting associations among other PorA and FetA variants and clonal complexes. The results showed shifting associations between some antigenic variants and clonal complex, as well as a subset of other associations among PorA, FetA and clonal complex, which persisted for decades and proliferated globally. Both FetA and PorA proteins represent candidates for serogroup B substitute vaccines, thus the findings presented here bear implications for vaccine design.

4.3 Introduction

Of 13 serogroups of *N meningitidis*, only five (A, B, C, W, Y) are responsible for the majority of invasive disease worldwide. Although protein-polysaccharide conjugate vaccines which target serogroups A, C, W and Y have been developed, such a vaccine against serogroup B is not available, owing to its similarity to the host antigen NCAM (Jodar *et al*. 2002). There is urgent need for a substitute for a serogroup B vaccine, as this serogroup is the predominant cause of meningococcal disease in many countries. In addition, there is no vaccine available against serogroup X, which has recently caused disease outbreaks in Africa (Boisier *et al*. 2007; Xie *et al*. 2013). Vaccines based on outer membrane proteins (OMPs), such as outer membrane vesicle (OMV) vaccines, provide an alternative to those which are capsule-based. However, the high rates of horizontal genetic exchange and diversifying selection in meningococcal populations results in antigenic diversity at levels which pose problems for designing universal protein-based vaccines (Feil *et al*. 1999; Jolley *et al*. 2000).

As reviewed in Section 1.2, meningococcal genetic diversity is highly structured: the population comprises a number of discrete genetic lineages, recognised as clonal complexes

by multilocus sequence typing (MLST) (Maiden *et al*. 1998). The majority of disease worldwide is caused by a limited number of hyperinvasive clonal complexes. Additional discriminatory power for strain typing and the investigation of disease outbreaks is provided by antigen-encoding genes, such as the two variable regions of the porin PorA (VR1 and VR2) and the iron-regulated OMP FetA, which has one variable region (Thompson *et al*. 2003).

The PubMLST database (www.pubmlst.org) is an internet-based repository of isolate information (Jolley *et al*. 2004) containing publically-available MLST and other typing schemes for a number of species. In addition to MLST data, the *Neisseria* database contains large amounts of genotypic data for the typing antigens PorA and FetA, thereby allowing the investigation of epidemiological patterns among clonal complexes, PorA, and FetA using data collated from a variety of countries over many years. The extensive geographical and temporal sampling frames of the dataset and its large sample size provide unique information for understanding global meningococcal population structure and planning vaccine campaigns. Although in its entirety it is not an epidemiologically defined sample, as it comprises isolates that have been voluntarily submitted by the research and public health community, the database does provide a definitive list of described diversity, as submission to it is a prerequisite of sequence type (ST) assignment. The database also includes data from various studies with coherent sampling frames which can be extracted and analysed. The PubMLST database therefore enables the investigation of certain defined questions into the population biology and evolution of the meningococcus, e.g. the stability of strain type (defined by clonal complex, PorA type and FetA type) can be inferred from the time span over which identical combinations of PorA alleles, FetA alleles and clonal complexes have been recorded.

The models presented in the previous chapter predict that the presence of immunological and ecological competition operating between pathogen strains – even if present at low levels in some areas of parameter space – results in stable, non-overlapping associations between

antigenic and metabolic types. In this chapter, in order to investigate the predictions of the models presented in the previous chapter and those of Buckee *et al*., the PubMLST database was used to examine the temporal and geographical distributions of associations between clonal complex and the antigens PorA and FetA, using 3460 carried and invasive isolates from 57 countries representing a 74 year period. Given that PorA and FetA are potential candidates for OMV vaccines (and indeed, formulations using PorA have been deployed to target specific serogroup B outbreak strains), the results presented here also have important implications for the design of OMV vaccines. The utility of such vaccines will be determined by the extent of diversity at these loci as well as their stability over time. The formulations for several potential protein-based vaccines contain a number of antigens, including fHBP, NadA and NHBA (Pizza *et al*. 2000). These antigens show similar behaviour to PorA and FetA with respect to their associations with clonal complexes (Brehony *et al*. 2009; Bambini *et al*. 2013), so conclusions drawn from the study of PorA and FetA may be relevant to these antigens also.

4.4 Methods

Several sequence typing schemes are embedded in the PubMLST database, including MLST, antigen and antibiotic sequence typing databases. In each case, records of isolates, allele sequences and schemes (groupings of particular loci) are maintained. The database was searched for all *Neisseria* species isolates containing information on PorA VR1, VR2 and FetA. A total of 3460 isolates had allelic information available for both antigens, which date from 1937 to 2010. Internal online database tools were used for data analysis searches, e.g. the "Publication" filter was employed to acquire information from specific published studies.

The nomenclature used was the OMP and clonal complex components recommended by Jolley *et al*. (2007), who suggest: serogroup:PorA type:FetA type:sequence type(clonal complex), thus: B:P1.19,15:F5-1:ST-33(cc32). P1 is a convention maintained from the serosubtyping scheme. Here we use: PorA type:FetA type:clonal complex, thus: P1.19,15:F5-1:cc32. "Minimum lifespan" was defined as the total number of years between the first and final year that a PorA:FetA:clonal complex combination was recorded, including the first and final year. Isolates were not necessarily present in every year over which the life span extends; and equally, isolates present in the database for only a short period may well have been circulating for much longer. Although an approximate measure, it does indicate the minimum period of time over which a specific strain type has been definitively recorded, therefore providing information on minimum strain type longevity. It is highly unlikely that multiple meningococci with identical clonal complexes, PorA alleles and FetA alleles arose independently owing to the large amount of diversity at the antigenic loci.

4.5 Results

4.5.1 Diversity of PorA and FetA Associations

A total of 88 PorA VR1 alleles, 223 VR2 alleles and 212 FetA alleles were present in 1129 PorA:FetA combinations in the dataset, of which 757 (67.1%) were present once. There was an asymmetric distribution among PorA:FetA combinations, with a subset recorded at higher frequencies: 13 PorA:FetA types had frequencies exceeding 30 isolates, such as P1.19,15:F5-1, represented by 270 isolates (Figure 4.1a). Furthermore, several PorA profiles were recorded predominantly with unique FetA profiles not shared by other PorA variants (Figure 4.1b). For example, 66 of 76 (86.8%) P1.7,16 isolates were associated with F3-3.

4.5.2 Diversity of PorA, FetA and Clonal Complex Associations

A total of 1420 PorA:FetA:clonal complex combinations were present in the dataset, of which 1030 (72.5%) appeared once. Several strain types were reported at high frequencies, with 10 PorA:FetA:clonal complex combinations represented by over 30 isolates, such as P1.19,15:F5-1:cc32 (263 isolates) and P1.5,2:F3-6:cc11 (117 isolates) (Table 4.1). Among the more frequent strain types, each PorA:FetA type was recorded predominantly with one clonal complex (Table 4.2), with a mean of 1.33 clonal complexes found with each PorA:FetA variant combination. For example, all of 117 isolates exhibiting P1.5,2:F3-6 were associated with the ST-11 complex. Similarly, each clonal complex was dominated by a unique antigenic type not shared by other complexes (Figure 4.2). The ST-41/44 complex, however, was particularly diverse as reflected by its association with 249 variant combinations. A total of 265 of 375 (70.7%) PorA:FetA:clonal complex associations recorded twice or more were associated with one specific serogroup, such as P1.5-2,10:F3-5:cc1, for which all 80 isolates were serogroup A (Table 4.1). Others were recorded with multiple serogroups, with 14 strain types recorded with three or more different serogroups. For example, P1.5,2:F1-1:cc11 (17 isolates), was recorded with serogroup W-135 (11 isolates), serogroup C (4) and serogroup B (2).

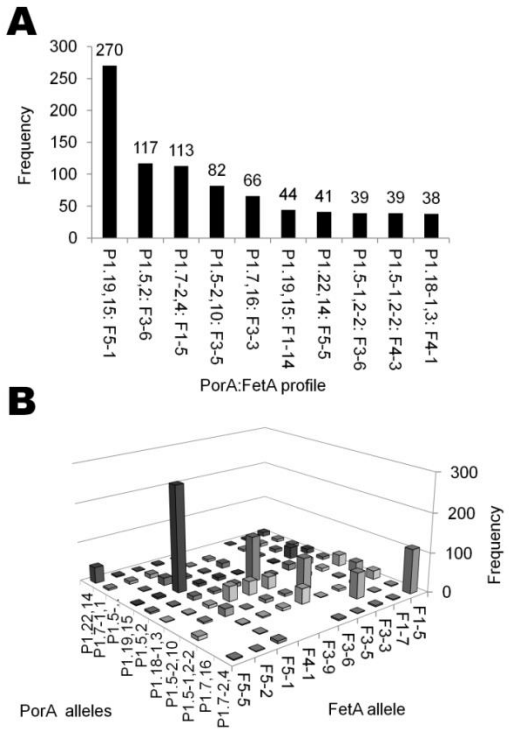


Figure 4.1. Frequently recorded PorA and FetA variant combinations. *The frequencies (A), and asymmetric distribution (B), of the most commonly recorded PorA and FetA allele combinations in the PubMLST database.*

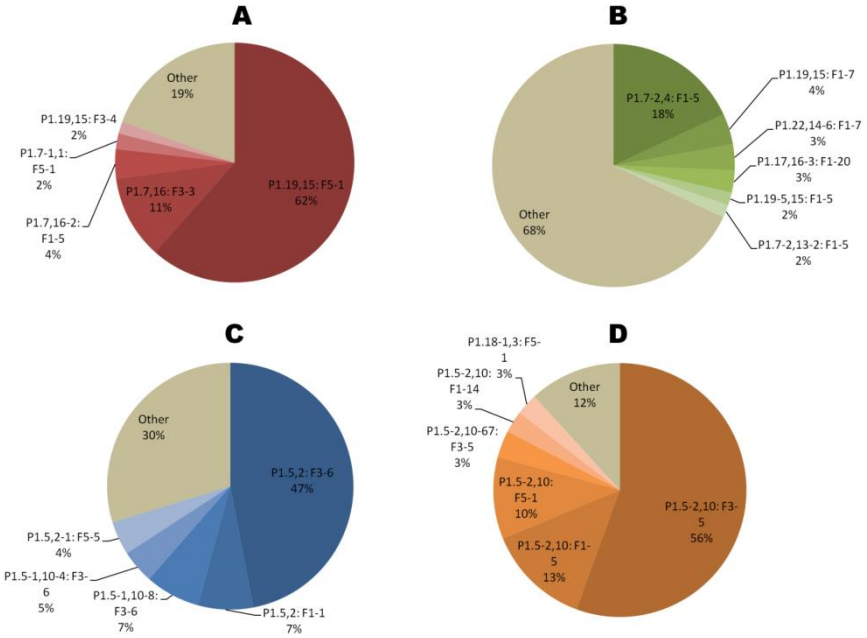


Figure 4.2. Distribution of PorA:FetA types among four hyperinvasive clonal complexes. *(A) ST-32 complex, (B) ST-41-44 complex, (C) ST-11complex and (D) ST-1 complex.*

PorA:FetA:clonal complex	No. isolates	Dominant serogroup(s) (%)	Disease: carriage ratio	Minimum lifespan (years)	Observed time period	Continents found
P1.5,2:F3-6:cc11	117	C (90.6), B (8.5)	1.74	49	1961-2009	Europe, N. & S. America
P1.5-2,10:F1-5:cc1	19	A (94.7)	8.50	37	1971-2007	Asia, Europe
P1.22,14-6:F1-7:cc41/44	20	B (70), NG (20)	0.053	37	1973-2009	Europe
P1.18-1,3: F4-1:cc22	29	W-135 (70), NG (8)	0.33	36	1975-2010	Europe, N. America
P1.7,16: F3-3:cc32	48	B (93.8)	4.63	35	1976-2010	Europe, Oceania, N. & S. America
P1.19,15: F1-7:cc41/44	24	B (58.3), NG (33.3)	0.09	31	1980-2010	Europe, Oceania
P1.19,15: F5-1:cc32	263	B (91.6)	1.19	27	1983-2009	Africa, Europe, N. & S. America
P1.7-2,4: F1-5:cc41/44	105	B (95.2)	32.7	26	1985-2010	Europe, N. & S. America, Oceania
P1.5-1,2-2: F5-8:cc23	22	Y (90.9)	1.13	26	1985-2010	Africa, Europe, N. & S. America
P1.5-2,10-1: F4-1:cc23	23	Y (91.3)	0.38	26	1985-2010	Europe, N.& S. America
P1.18-1,34: F1-9:(-)	22	C (68.2), NG (22.7)	0.16	23	1975-1997	Europe
P1.19,15: F1-14:(-)	40	B (90)	0.67	18	1993-2010	Europe
P1.22-1,14: F4-1:cc35	32	B (84.4), Y (6.2)	0.88	13	1998-2010	Europe, N. & S. America
P1.18,25-15: F5-5:cc198	22	NG (95.5)	0.10	12	1998-2009	Europe, N. America
P1.5-2,10: F3-5:cc1	80	A (100)	2.81	10	2000-2009	Asia
P1.22,14: F5-5:cc213	40	B (82.5) C (10)	2.0	8	2003-2010	Asia, Europe, N. America, Oceania
P1.7-2,30: F1-7:cc53	21	NG (90)	0	8	1991-1998	Europe, N. America

Table 4.1. Characteristics of the most frequent strain types in the PubMLST database: many strain types are long-lived and globally widespread. *NG refers to non-groupable.*

Antigenic profile	Clonal complex	No. isolates (%)
P1.19,15: F5-1	ST-32	263/270 (97.4)
P1.5,2: F3-6	ST-11	117/117 (100)
P1.7-2,4: F1-5	ST-41/44	105/113 (92.9)
P1.5-2,10: F3-5	ST-1	80/82 (97.6)
P1.7,16: F3-3	ST-32	48/66 (72.7)
P1.19,15: F1-14	Unassigned	40/44 (90.9)
P1.22,14: F5-5	ST-213	40/41 (97.6)
P1.5-1,2-2: F3-6	ST-8	33/39 (84.6)
P1.5-1,2-2: F4-3	ST-549	38/39 (97.4)
P1.18-1,3: F4-1	ST-22	29/38 (76.3)

Table 4.2. Association of PorA:FetA types with particular clonal complexes.

4.5.3 Prevalence of Strain Types over Time

Minimum lifespans of strain types, calculated from isolate information in the PubMLST database, ranged from 74 years (of strain type P1.5-2,10:F1-5:cc4) to one year, which was the case for numerous types. A total of 1098 of 1420 (77.3%) PorA:FetA:clonal complex associations had observed lifespans of one year or less; however, these potentially short-lived variants comprised only 1280 out of 3460 isolates (37.0%). The dataset also contained a small number of variants with evidence for intermediate longevity (72 associations persisted for a minimum of 5-9 years), and a larger number of long-lived associations (132) with minimum lifespans in excess of 10 years. Of these, 44 variants had minimum lifespans of 20-39 years and 10 lasted for over 40 years; evidence for longevity was present for only a small minority of strain types.

4.5.4 Global Distribution of Strain Types

Several strain types had representatives in the database from a range of countries across the globe (Table 4.1), such as P1.7,16:F3-3:cc32 and P1.7-2,4:F1-5:cc41/44, each recorded in 11 countries across Europe, the Americas and Oceania (Figure 4.3). Many of these strain types had minimum lifespans greater than 10 years and belonged to known hyperinvasive clonal complexes (Caugant & Maiden 2009): strain types P1.7,16:F3-3:cc32 and P1.7-2,4:F1-5:cc41/44 had minimum lifespans of 35 and 26 years respectively. The location of isolation of these types changed over time, and they were present in different sites for varying lengths of time (Figure 4.3).

4.5.5 Spatio-Temporal Patterns within Countries

Data from coherent population samples from the United Kingdon and Czech Republic stored within the PubMLST database demonstrated that within each country antigenic types showed temporal variation with particular PorA:FetA combinations rising and falling in dominance (Figure 4.4) (Buckee *et al*. 2008; Russell *et al*. 2008). Strain types represented by 2 or more isolates were found in these countries for mean durations of 8.6 and 6.4 years respectively. Each PorA profile was mainly recorded with a particular FetA allele in a given year, which changed over time. For example, F1-5 allele in the United Kingdom was associated with P1.5-1,2-2 in 1975 (8 isolates), P1.7,16-2 in 1985 (11) and P1.7-2,4 in 1995 (16). Fluctuating associations were also seen between clonal complex and PorA:FetA types, e.g. the ST-11 complex in the United Kingdom was associated with P1.5,2:F1-1 in 1975 (6 isolates); P1.5,2-1:F5-5 in 1985 (6); and P1.5,2:F3-6 (8) and P1.5-1,10-4:F3-6 (8) in 1995.

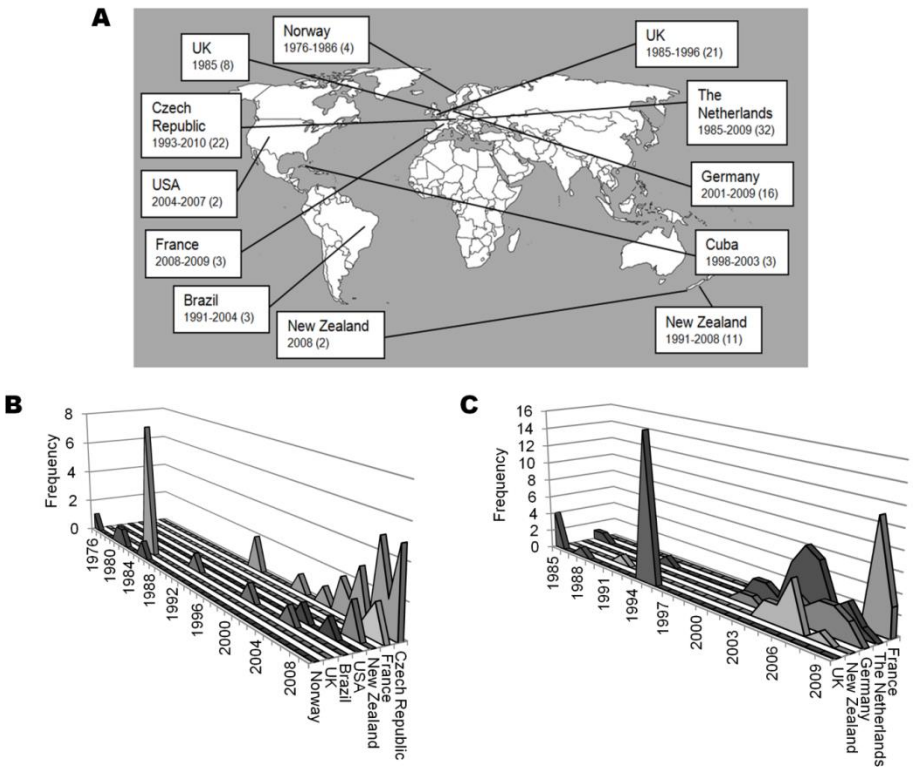


Figure 4.3 Tracking the global distributions of two hyperinvasive strain types over time. *The location, date and frequency of P1.7,16:F3-3:cc32 isolates (A: left side)(B) and P1.7-2,4:F1-5:cc41/44 isolates (A:right side)(C), as reported in the PubMLST database.*

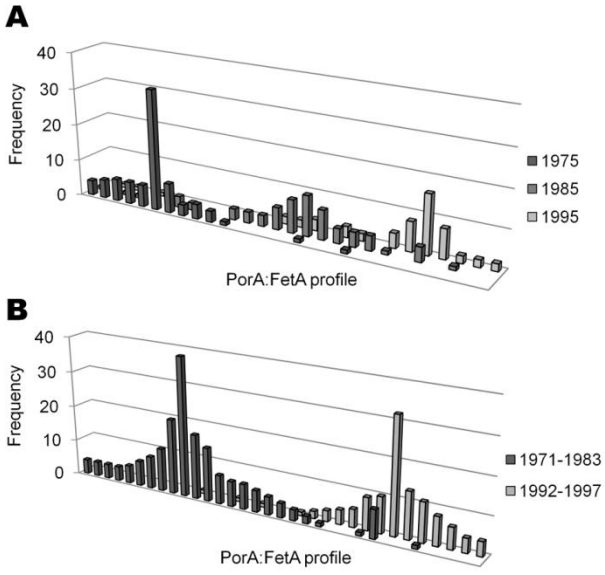


Figure 4.4. Temporal succession of PorA:FetA profiles in 2 countries. (*A) PorA:FetA combinations for 207 disease isolates recorded in the United Kingdom in 1975, 1985 and 1995. (Data taken from Russell et al.,2008). (B) PorA:FetA combinations for 314 carried isolates in the Czech Republic collected from 1971–1983, and 1992–1997. (Data taken from Buckee et al.,2008).*

4.6 Discussion

The PubMLST site contains information on clonal complex, PorA and FetA, collated from a number of sources worldwide over several years for a large number of isolates. Although it is not an epidemiologically representative dataset, as it relies upon the voluntary submission of genotypic data from the scientific community, it still provides information about the location of specific allelic associations at given time points. Data from the whole database demonstrate the stability of strain types and traces their global movements, and are complementary to the published datasets it contains, with detailed information from individual countries. Here, the temporal and geographical distribution of PorA, FetA and clonal complex associations were investigated in the PubMLST database, using 3460 isolates over 74 years from 57 countries. Investigating these associations allows for the predictions of the models presented in this thesis and by Buckee *et al*. (2008), to be assessed within the context of meningococcal typing data using a large dataset spanning several geographical regions over many years, and also presents a number of implications for OMV-based vaccines.

The meningococcus is highly genetically diverse, with 1030 unique strain types defined by PorA, FetA and clonal complex present in the database (corresponding to 72.5% of recorded isolates). These assorted combinations of different alleles are consistent with the large amounts of recombination observed in meningococcal populations. However, this diversity was highly structured, as populations typically comprise a limited number of strain types with non-overlapping combinations of outer membrane antigens and MLST genes (Table 4.1; Table 4.2; Figure 4.1). The PubMLST records demonstrated that a number of these strain types persisted for several decades (Table 4.1): a total of 54 PorA:FetA:clonal complex associations were found in meningococci isolated over periods in excess of 20 years. Although longevity of meningococcal strains has been shown elsewhere (Caugant *et al*. 1986; Urwin *et al*. 2004;

Russell *et al*. 2008; Buckee *et al*. 2010; Bambini *et al*. 2013), to our knowledge this is the first time that extensive minimum lifespans of up to 74 years have been documented. Furthermore, many of the long-lived associations shown here involve meningococci belonging to serogroups B and C. This is especially intriguing for serogroup B meningococci, which are regarded as highly antigenically diverse (Feavers *et al*. 1996). Several of these long-lived strain types were associated with particular hyperinvasive clonal complexes and have spread to a number of countries worldwide (Table 4.1). Their distribution appears to have shifted between continents over the past few decades, perhaps due to changing host immunity (Figure 4.4).

The intercontinental spread of several clones of the ST-32 (ET-5) complex was first described in the 1980s (Caugant *et al*. 1986); this complex comprises several strain types (Figure 4.2), and the global dissemination of the serogroup B strain type P1.7,16:F3-3:cc32 is described here. The PubMLST database included records of isolates from widespread locations from 1976 to 2010 (Figure 4.4a, 4.4b). Consistent with epidemics in Norway from 1974 onwards, which prompted the development of the MenBVac OMV vaccine, the database showed records in Norway from 1976 to 1986 (Bjune *et al*. 1991). This strain type was subsequently reported in the United Kingdom with high frequencies in 1985, in line with a known outbreak in Southampton at that time (Cartwright *et al*. 1986); more recently, isolates have been recorded in the United States and France, corresponding to outbreaks in Oregon and Normandy (Diermayer *et al*. 1999; Rouaud *et al*. 2006). Similarly, the ST-41/44 complex hyperinvasive strain type, B:P1.7-2,4:F1-5:cc41/44, was present in 11 countries worldwide over 26 years (Figure 4.4a, 4.4c). The strain type was recorded at high frequencies in the United Kingdom in 1995, and more recently in the Netherlands, Germany, New Zealand and elsewhere. A subsequent decline in New Zealand accompanied the immunisation programme of the MenZB vaccine in 2004, which has been suggested as a possible means to control outbreaks in Germany (Oster *et al*. 2005; Elias *et al*. 2010).

In addition to global cycling of pandemic meningococci among continents, the PubMLST data suggested a concurrent turnover of strain types within countries. The repertoire of antigenic variants present in each country changed over time with different profiles rising and falling in dominance (Figure 4.3). There are two factors which could help explain this regional instability. First, these patterns could partly be accounted for by antigenic shifts of endemic strains, in which the dominant antigens associated with clonal complexes are observed to vary. Shifting associations between clonal complex and OMP type have been observed in the United States – resulting in increased invasive disease (Harrison *et al*. 2006) – as well as in the Czech Republic (Buckee *et al*. 2008). The second explanation is the movement of pandemic hyperinvasive lineages among countries, with host populations colonised by meningococci expressing previously unseen antigenic repertoires. In both cases, the initial increase in frequency of the novel strain type would be a consequence of the lack of population immunity to the strain type, followed by a decline a few years later as immunity to it increased in the host population. For example, the pandemic strain type P1.5,2:F3-6:cc11 expressed PorA and FetA alleles previously unobserved in the Czech Republic when it was introduced following the Velvet Revolution in 1993. Consequently, immediately following its appearance, this strain type caused elevated levels of carriage that accompanied a disease outbreak (Jolley *et al*. 2005).

Although clonal models of descent have been invoked to explain serogroup A population structure (Bart *et al*. 2001; Olyhoek *et al*. 1987), a number of competing theoretical frameworks have been invoked to explain the population structure of other serogroups, which contain conflicting signals of genetic exchange and clonal descent. The observations presented here, of stable and non-overlapping associations between PorA and FetA, are consistent with the predictions of the framework by Gupta *et al.* (1996; 1998), in which strong-cross immunity segregates pathogen populations into non-overlapping antigenic variants. The shifting associations between PorA:FetA type and clonal complex are consistent with the predictions

of the stochastic model by Buckee *et al*. (2008), which showed fluctuating associations between antigenic type and ST at intermediate fitness differences between STs and high cross-immunity. The persistence of the non-overlapping associations between PorA, FetA and clonal complex lends support to the models presented in this thesis and Buckee *et al*.'s model (at low fitness differences between STs), that non-overlapping associations between antigenic and metabolic type are highly stable. The rationale for such long-lived and unique combinations is that ecological and immunological competition have segregated the meningococcal strains into unique metabolic and antigenic niches. Although most PorA:FetA profiles are not shared among clonal complexes, and most clonal complexes are dominated by one unique PorA:FetA type, the associations are not absolute, and there are a few examples of overlap. For example, the ST-41/44 complex has a large diversity of different PorA:FetA profiles. Thus it could be suggested that meningococcal populations conform to the red and purple structures shown in Figure 3.13.

The antigenic diversity of meningococcal populations presents several challenges for the design of protein-based vaccines based on variable antigens such as PorA and FetA. First, it is apparently difficult to achieve coverage across many strains (Tondella *et al*. 2000): the hypothetical vaccine coverage provided by any of the existing OMV vaccines did not exceed 8% of isolates in the PubMLST dataset (Table 4.3). Further, a vaccine containing the five most frequent PorA and FetA variants would protect against only 46.9% of meningococci represented in the PubMLST dataset. Second, the temporal dynamics raise problems for coverage, and the propensity to exchange genetic material presents the possibility of antigenic shift among vaccine-target antigens, instances of which have been documented (Harrison *et al*. 2006). Third, there is substantial heterogeneity of protein variants among countries at a given point in time (Figure 4.4), thus hindering the development of a comprehensive vaccine appropriate to all countries. Finally, some evidence suggests that existing protein-based OMV

vaccines are poorly immunogenic in infants and may require multiple doses to ensure long-term immunity (Holst *et al*. 2003).

Trial	Vaccine	Antigens	No. isolates (%)	Minimum utility of vaccine (years)*	Reference
Norway	MenBVac	P1.7,16:F3	66 (1.9)	35	(Bjune *et al*. 1991)
New Zealand	MeNZB	P1.7-2,4	147 (4.3)	26	(Oster *et al*. 2005)
Cuba	VA-MENGOC-BC	P1.19,15:F	270 (7.8)	36	(Sierra *et al*. 1991)

Table 4.3. Hypothetical vaccine coverage conferred by OMV vaccines against isolates in the PubMLST sample. **Minimum utility was defined as the total period of time across which a given antigenic combination had been recorded in the PubMLST sample.*

Conversely, a number of opportunities for vaccine design are presented by the structuring of meningococcal populations into non-overlapping antigenic variants associated with particular clonal complexes (Figure 4.2). As a limited number of hyperinvasive strain types circulate at higher frequencies (Table 4.1) (Urwin *et al*. 2004; Caugant & Maiden 2009), immunisation strategies that target these with vaccines containing appropriate combinations of antigen variants can be devised. Importantly, as several hyperinvasive meningococci are associated with more than one serogroup (Table 4.1) (Maiden *et al*. 1998; Jolley *et al*. 2000; Yazdankhah *et al*. 2004), such vaccines would be effective against these irrespective of the serogroup expressed, potentially circumventing the effects of capsule switching. Secondly, although the diversity of meningococci on a global scale might require a large number of variants for comprehensive coverage, geographic structuring could be exploited in order to design simpler vaccines that are tailor-made to target hyperinvasive clones within a given locale, as countries typically experience a limited number of circulating hyperinvasive antigenic types at a given time. For example, data from the PubMLST database suggest that a vaccine containing 5 PorA and FetA variants (P1.7-2,4, P1.22,14, P1.21,16, P1.7-2,13-2, P1.22,9; F1-5, F5-5, F1-7, F3-3 and F5-12) would potentially protect against 70.2% of serogroup B strain types in Germany

(185 isolates). Antigenic types rise and fall in prevalence over time (Figure 4.3, Figure 4.4), but remain present for periods of time sufficient to administer vaccination programmes. This is particularly so in the case of serogroup B outbreaks which can persist for many years (Caugant *et al*. 1986), such as strain type P1.7-2,4:F1-5:cc41/44, isolated from the database in the Netherlands for at least 25 years. Thirdly, as the minimum lifespans of global hyperinvasive strain types span decades as they move among countries, protein-based vaccines directed against them have the potential to be employed for many years in different places (Table 4.3). An example of this is the global serogroup B strain type P1.7,16:F-33:cc32, with a minimum lifespan of 35 years, against which the Norwegian vaccine (MenBVac) was developed. The most recent use of this vaccine was in the Normandy region of France, more than twenty years after it was developed (Caron *et al*. 2011).

4.7. Summary

Notwithstanding frequent recombination in meningococcal populations, the results here suggest that the high level of genetic variation they exhibit is extensively structured. The meningococci analysed in the PubMLST database are dominated by a relatively small number of discrete combinations of clonal complexes, PorA and FetA variants, which persist for decades and proliferate globally. The long lifespans of these strain types are punctuated by their movements among countries. This structuring and dynamic behaviour further support the predictions of the models presented in thesis, and previously by Buckee *et al*. (2008), that invoke selective forces imposed by host immunity and ecological competition, to account for meningococcal population structure in the face of extensive recombination. The longevity of the non-overlapping strain types, combined with their global circulation, suggest that vaccines that target hyperinvasive lineages based on their antigenic repertoires may hold some promise.

Chapter 5.

Genomic Analysis of Metabolic Profiles in *Streptococcus pneumoniae*

5.1 Public Dissemination of Work

These findings have been published in the following article, in association with the results in Chapters 3 and 7:

Eleanor R. Watkins, Bridget S. Penman, José Lourenço, Caroline O. Buckee, Martin C. J. Maiden & Sunetra Gupta (2015). Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. *Plos Pathogens*, 11(7): e1005034.

For the published article, ERW and SG conceived and designed the experiments; ERW performed the mathematical modelling and statistical tests; ERW analysed the genomic data and output from the mathematical models; and SG, ERW, JL, COB, and MCJM wrote the paper.

5.2 Abstract

The genotypic support provided in the previous chapters for non-overlapping associations between antigenic and metabolic alleles encompasses diversity at only a few genes. The patterns observed at the small number of metabolic enzymes used in MLST/MLEE typing schemes may not be representative of the signatures of diversity among the metabolic genes across the whole genome. Therefore, in this chapter, we aim to investigate if the predictions of the multilocus models presented in this thesis and by Buckee *et al*. (2008) incorporating ecological and immune-mediated competition are upheld at the whole genome level, using 616 published whole genomes of *S. pneumoniae*. We first characterise the metabolic/transport genes in the pneumococcal reference genome – corresponding to the metabolic profile – and then assess the diversity of metabolic profiles within and between serotypes. Subsequently, to

investigate if the MLST-defined sequence type (ST) represents an adequate marker of the true genome-wide metabolic type, we extend our analyses to STs. Finally, to investigate the hypothesis that pneumococcal metabolic profiles comprise a set of co-evolved genes, we examine the strength of associations between alleles of metabolic/transport loci across the genome. We would hypothesise that the metabolic profiles are highly uniform within individual serotype and STs, and different to those of other serotypes. We would also hypothesise that associations between alleles of metabolic/transport loci across the genome are non-random, as this would support the notion that pneumococcal metabolic profiles comprise a set of co-evolved genes which have synergistically adapted to exploit a particular metabolic niche. We find that metabolic profiles are highly uniform within individual serotypes and STs, and significantly different to those of others, and that associations between alleles of metabolic/transport loci are highly non-random.

5.3 Introduction

The stochastic model by Buckee *et al*. (2008) predicted non-overlapping associations between antigenic and metabolic genes at high levels of ecological and immunological competition. The deterministic models presented in Chapter 3 also suggest that ecological and immunological competition are important in generating non-overlapping combinations; however, such associations were found even when the strengths of competition were relatively low. Associations between serotype and ST in *S. pneumoniae* have been observed in a number of studies from samples of both carriage and invasive isolates worldwide (Figure 1.9; Table 1.3). Associations are not absolute, however, with some serotypes exhibiting more than one ST in the same locality, and vice versa (Brueggemann *et al*. 2003; Sandgren *et al*. 2004; Hanage *et al*. 2005). Further, the MLST-defined ST, comprising just seven housekeeping genes, may not

be an adequate proxy for the true metabolic type - that is, defined at the whole genome level. To gain a comprehensive understanding of the associations between serotype and metabolic genes in *S. pneumoniae*, and to assess the validity of the ST as a proxy for metabolic type, the correspondence of the metabolic genes of 616 carried pneumococci to serotype and ST groupings are analysed.

The core metabolism of bacterial pathogens comprise highly integrated networks of interacting protein components. Thus, if strains do indeed exploit distinct metabolic niches, different metabolic types may manifest differences in a number of interacting metabolic proteins. It follows that variants of these genes may show high levels of genetic linkage, if they code for interacting and co-adapted protein networks and complexes. In order to gain a more detailed understanding of the genetic structuring of the metabolic gene variants in *S. pneumoniae*, two separate measures of association were investigated across metabolic alleles. Given that the presence of specific transport genes in *S. pneumoniae* determines the exact sugars which a given strain is able to utilise, the transport genes are likely to play an important role in defining the metabolic niche, and are thus also included in the analysis.

5.4 Methods

5.4.1 Genome Assembly, Annotation and Identification of Metabolic Loci

A sample of 616 genomes published previously by Croucher *et al*. (2013), comprising carriage strains isolated from Massachusetts, USA, in 2001, 2004 and 2007 was used for the analysis. Sequencing was carried out at the Wellcome Trust Sanger Institute in Cambridge, UK. Sequence reads were taken from the project ERP000889 on the European Nucleotide Archive (http://www.ebi.ac.uk/) and assembled using an automated pipeline with Velvet (Zerbino &

Birney 2008). Annotation was carried out using the BIGSdb software with an automated BLAST process, and the genomes were analysed using the Genome Comparator tool (Jolley & Maiden 2010). Finished genome ATCC 700669 was used as the reference (accession FM211187). The Genome Comparator algorithm designates alleles which are identical to the reference as "1", and sequences which differ at one or more bases are labelled consecutively. Missing or truncated alleles are also indicated. The 2135 coding sequences in the reference genome were functionally categorised into one of four groups: genes encoding metabolic & transport functions; pseudogenes and mobile genetic elements; genes of unknown function; and other genes. The genes were classified primarily using the KEGG Orthology (KO) groupings of the KEGG database (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/). The genes in the "other" category coded for proteins with characterised functions which did not contribute directly to metabolic or transport processes, and were not related to phages or transposable elements.

5.4.2. Investigating Associations between Serotype and Metabolic Profile

To explore the structuring of metabolic/transport alleles among serotypes, we employed a simple metric summarising the percentages of identical alleles at each locus. The modal allele for each locus was identified and the proportion at which it occurred over all the isolates in a given serotype was calculated. As an example, at the SPN23F00790 locus (encoding Sugar Phosphotransferase System, Fructose Family), allele "1" is present at 54 out of 60 isolates of serotype 19A. Therefore, the modal percentage identity (MPI) for this locus in serotype 19A is 0.9. To compare the MPI for each serotype with randomly selected isolates, 10 isolates were randomly selected (without replacement) from the total sample of 616, and the MPI calculated for the newly generated modal alleles for each locus as above.

To explore if metabolic/transport alleles at each locus were significantly more uniform within individual serotypes than across randomly selected isolates, we performed a Sign Test on the differences obtained by subtracting MPIs within a random sample from MPIs at the same loci within a given serotype. Analyses were carried out using the programme IBM SPSS Statistics (version 22); estimates of median differences were obtained using the programme R (version 2.15.1) and the package "BSDA."

In order to explore the extent to which metabolic/transport alleles differed between serotypes, we employed a second simple metric. The modal metabolic profile for each serotype was identified (the profile comprising the most frequent alleles found at each locus) and the pairwise percentage of alleles which were shared with the modal metabolic profile of other serotypes was calculated.

5.4.3. Investigating Associations between Sequence Type and Metabolic Profile

The MPIs for the most frequent STs in the dataset published by Croucher *et al.* (2013) were calculated, and compared with the MPIs of randomly selected isolates, as described above. To explore if metabolic/transport loci were significantly more uniform within individual STs than across randomly selected isolates, we performed a Sign Test on the differences obtained by subtracting MPIs within a random sample from MPIs at the same loci within a given ST (as described above).

5.4.4 Linkage Disequilibrium and Mutual Information Values for Metabolic and Non-Metabolic Genes

We used the D' measure of Linkage Disequilibrium to investigate the strength of associations between pairs of 100 randomly selected metabolic/transport loci across 300 genomes (which were randomly selected for each comparison). To investigate if these values were significantly higher than among other genes, we carried out the equivalent tests on 100 randomly selected loci with known functions that were not involved in metabolic/transport processes or related to phages or transposons. D' was analysed using the 2LD package (Zhao 2004).

We calculated the Mutual Information (MI) metric (which is used to measure the strength of association between variables) on the same sample (Shannon 1948). The MI values were normalised (by dividing the MI value by the maximum MI value, given as the log of the smallest number of sites at any locus) in order to account for inequalities in the number of alleles at different loci.

5.4.4.1 Linkage Disequilibrium, D'

D' is based on the difference between observed allelic frequencies and those expected from a homogenous, randomly distributed model.

D is the raw difference in frequency between the observed number of AB pairs and the expected number:

$$D = p(AB) - p(A)*p(B)$$

Where

$p(A)$ is defined as the observed probability of allele 'A' for marker 1,

$p(a) = 1-p(A)$ is defined as the observed probability of allele 'a' for marker 1,

p(B) is defined as the observed probability of allele 'B' for marker 2, and

p(b) = 1- p(B) is defined as the observed probability of allele 'b' for marker 2

p(AB) is defined as the probability of the marker allele pair 'AB'.

However, the value of D depends on the frequencies of the alleles. Thus, D can be normalised by dividing it by the theoretical maximum for the observed allele frequencies, to give D'.

Thus $D' = D/D_{max}$

where, if D > 0:

$$Dmax = min(p(A)p(b), p(a)p(B))$$

or if D < 0:

$$Dmax = max(-p(A)p(B), -p(a)p(b))$$

5.4.4.1 Mutual Information

Formally, the mutual information of two variables x and y can be given as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

Where p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively.

5.5 Results

5.5.1 Allelic Diversity

We identified 877 genes which code for proteins involved in metabolic and/or uptake processes (corresponding to the metabolic profile). A total of 177 loci encoded pseudogenes and/or were related to phage/transposons; 570 genes had no known function; and 511 genes encoded proteins involved in other processes not directly related to metabolism or transport. These functional genes were involved in genetic information processes such as translation and transcription, as well as cell signalling systems and cellular processes such as DNA methylation.

There was extensive allelic diversity in the metabolic/transport genes across isolates. The number of alleles present at each locus varied from 1 to 139, with an average of 43.2 alleles. At the 2 loci for which one allele was found, the complete gene was not present in >90% of the isolates; thus no alleles of the 877 metabolic/transport loci were conserved across all 616 isolates. The allelic diversity in metabolic/transport genes was slightly higher than in genes coding for non-metabolic/transport functions, of which there were 38.2 alleles per locus on average.

5.5.2 Associations between Serotype and Metabolic Profile

The allelic differences in the metabolic/transport loci between isolates tended to correlate with different serotype groupings (Figure 5.1a). Although a number of alleles were shared across different serotypes, the overall metabolic/transport profiles were specific to one or a few serotypes.

The percentage of identical alleles at each locus (Modal Percentage Identity, MPI) within each serotype was compared with groups of randomly selected isolates (as described in Section 5.2.2). Loci containing alleles in which one or more isolates were truncated or deleted were discarded from the analysis, giving a total of 782 loci. The violin plots displayed in Figure 5.1b show the distribution of MPIs at those 782 loci for the most frequent serotypes.

To explore if metabolic/transport alleles at each locus were significantly more uniform within individual serotypes than across randomly selected isolates, we performed a Sign Test on the differences obtained by subtracting MPIs within a random sample from MPIs at the same loci within a given serotype. The median difference between within serotype MPIs and MPIs from randomly selected isolates was significantly different from zero for all serotypes included, with the number of positive differences consistently outweighing the number of negative differences (Table 5.1). This supports the hypothesis that the MPI is much higher within individual serotypes than would be expected by chance.

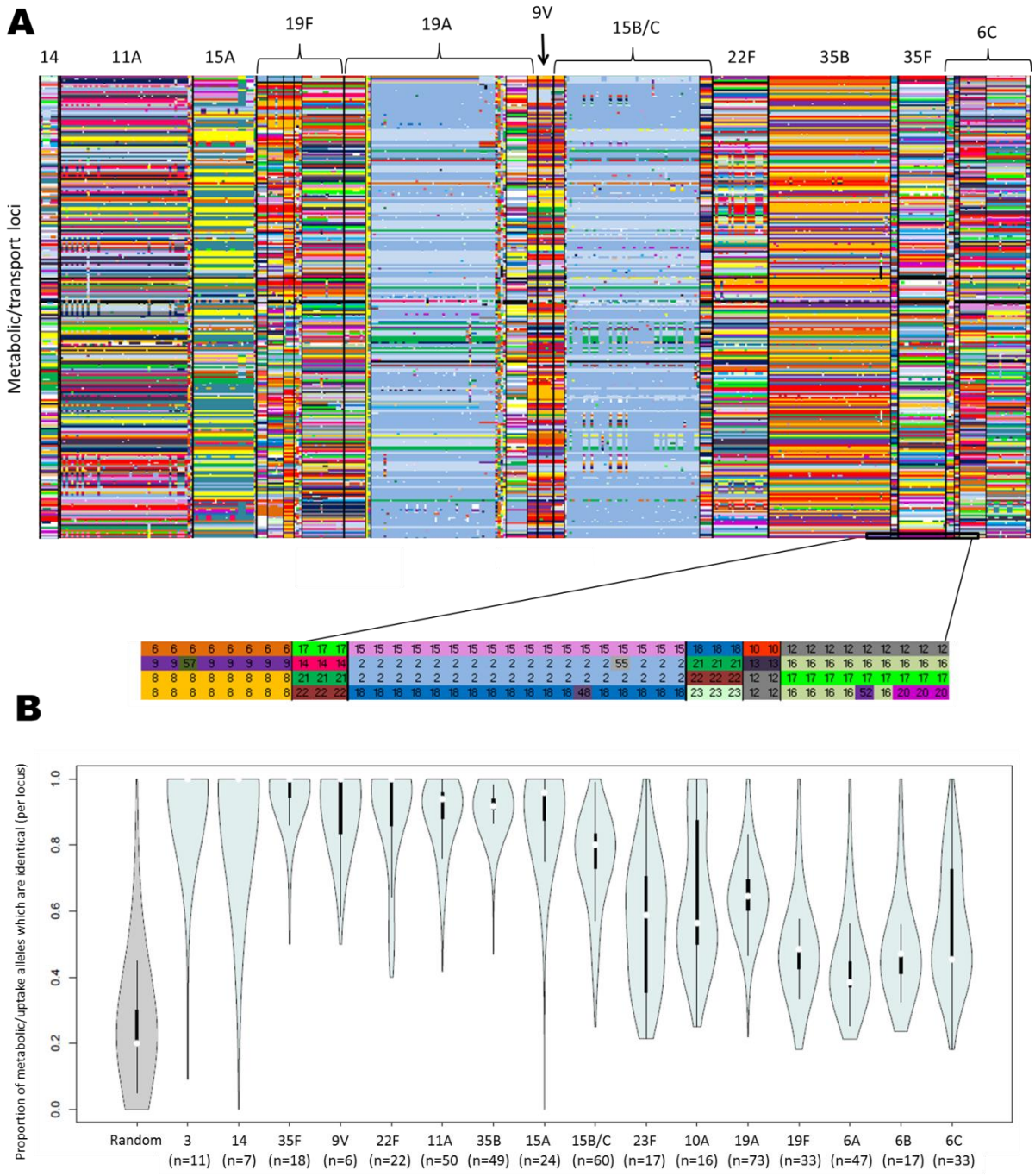


Figure 5.1 Analysis of the metabolic/transport alleles among serotypes. *(A) Partial screenshot of the metabolic profiles of 616 strains, grouped by serotype. Each column contains the allele numbers for a single isolate, of metabolic/transport genes (represented by rows). Allele numbers are represented by different colours. (B) Violin plots showing the distribution of the modal percentage identity (MPI) of metabolic/transport alleles for each serotype (the proportion of alleles at each locus which are identical). The MPI for each serotype (blue) was significantly higher than a distribution of randomly selected isolates (grey). The plots were produced using the "vioplot" package in R (version 2.15.1).The plots were produced using the "vioplot" package in R (version 2.15.1).*

To explore the extent to which metabolic/transport alleles differed between serotypes, the modal metabolic profile for each serotype was identified (the profile comprising the most frequent alleles found at each locus) and the pairwise percentage of alleles which were shared with the modal metabolic profile of other serotypes was calculated (Table 5.2). The percentage of identical metabolic/transport alleles shared across serotypes was relatively low (with any two serotypes sharing 10.1% of their metabolic/transport alleles on average), and significantly less than 1 (Wilcoxon Mann Whitney test, V = 1540, p-value < 1.129e-10).

Test	No. positive differences	No. negative differences	Ties	Test statistic	Standardised test statistic	Sample estimate	Upper achieved confidence interval (0.9509)	P-value (2-sided)
3 - Random	770	10	2	770	27.17	0.71	0.71 - 0.80	< 2.2e-16
14 - Random	777	3	2	777	27.68	0.70	0.70 - 0.78	< 2.2e-16
35F - Random	780	0	2	780	27.89	0.74	0.70 - 0.74	< 2.2e-16
9V - Random	764	15	3	764	26.8	0.70	0.70 - 0.70	< 2.2e-16
22F - Random	778	2	2	778	27.75	0.70	0.70 - 0.70	< 2.2e-16
11A - Random	778	1	3	778	27.8	0.66	0.66 - 0.68	< 2.2e-16
35B - Random	780	0	2	780	27.89	0.70	0.69 - 0.70	< 2.2e-16
15A - Random	775	2	5	775	27.7	0.66	0.66 - 0.70	< 2.2e-16
15B/C - Random	769	10	3	769	27.16	0.53	0.52 - 0.53	< 2.2e-16
23F - Random	756	18	8	756	26.49	0.38	0.35 - 0.40	< 2.2e-16
19A - Random	770	11	1	770	27.12	0.41	0.40 - 0.42	< 2.2e-16
19F - Random	772	58	2	772	23.74	0.25	0.23 - 0.25	< 2.2e-16
10A - Random	748	20	14	748	26.23	0.36	0.31 - 0.36	< 2.2e-16
6B - Random	726	55	1	726	23.97	0.22	0.22 - 0.22	< 2.2e-16
6C - Random	741	39	2	741	25.1	0.25	0.25 - 0.25	< 2.2e-16
6A - Random	687	93	2	687	21.23	0.18	0.18 - 0.18	< 2.2e-16

Table 5.1. Outputs of the two-tailed Sign Tests comparing the average proportion of metabolic/transport alleles which are identical at a given locus (MPI), between isolates of the same serotype and randomly-selected isolates. *The tests were performed for the most common serotypes in the dataset published by Croucher et al.. (2013). The sample estimate refers to the point estimate of the median difference between the two lists of modal percentage identities.*

	3	14	11A	15A	15B/C	19A	22F	23F	35F	9V	35B
3	1	0.077	0.085	0.057	0.070	0.077	0.070	0.085	0.089	0.028	0.078
14	0.077	1	0.120	0.104	0.097	0.131	0.083	0.086	0.079	0.074	0.093
11A	0.085	0.120	1	0.192	0.112	0.110	0.091	0.083	0.122	0.078	0.097
15A	0.057	0.104	0.192	1	0.091	0.102	0.079	0.085	0.094	0.077	0.093
15B/C	0.070	0.097	0.112	0.091	1	0.806	0.106	0.142	0.086	0.103	0.104
19A	0.077	0.131	0.110	0.102	0.806	1	0.107	0.143	0.077	0.120	0.106
22F	0.070	0.083	0.091	0.079	0.106	0.107	1	0.110	0.094	0.075	0.146
23F	0.085	0.086	0.083	0.085	0.142	0.143	0.110	1	0.085	0.090	0.091
35F	0.089	0.079	0.122	0.094	0.086	0.077	0.094	0.085	1	0.086	0.097
9V	0.074	0.074	0.078	0.077	0.103	0.120	0.075	0.090	0.086	1	0.078
35B	0.078	0.093	0.097	0.093	0.104	0.106	0.146	0.091	0.097	0.078	1

Table 5.2 The percentage of metabolic/transport alleles shared between pairs of serotypes. *Loci in which alleles were truncated or absent were excluded from the analysis, providing a total of 756 loci. Serotypes comprising more than one metabolic type were excluded, as the modal metabolic profile is not a legitimate representation of the metabolic traits for these serotypes.*

5.5.3 Associations between Sequence Type and Metabolic Profile

Allelic differences in metabolic/transport genes tended to correlate with the most frequent ST groupings defined by MLST (Figure 5.2a). Although a small number of STs shared highly similar metabolic profiles, they also tended to be closely related according to MLST-typing. For example, ST236, ST320 and ST271 manifest highly similar metabolic alleles (Figure 5.2a), and belong to the same MLST-defined clonal complex (CC271). The modal percentage identities (MPI) for the most frequent STs in the dataset published by Croucher *et al.* (2013) were calculated, and compared with the MPI of randomly selected isolates, as described above. Figure 5.2b shows the MPI for 14 STs and randomly selected isolates across 877 loci.

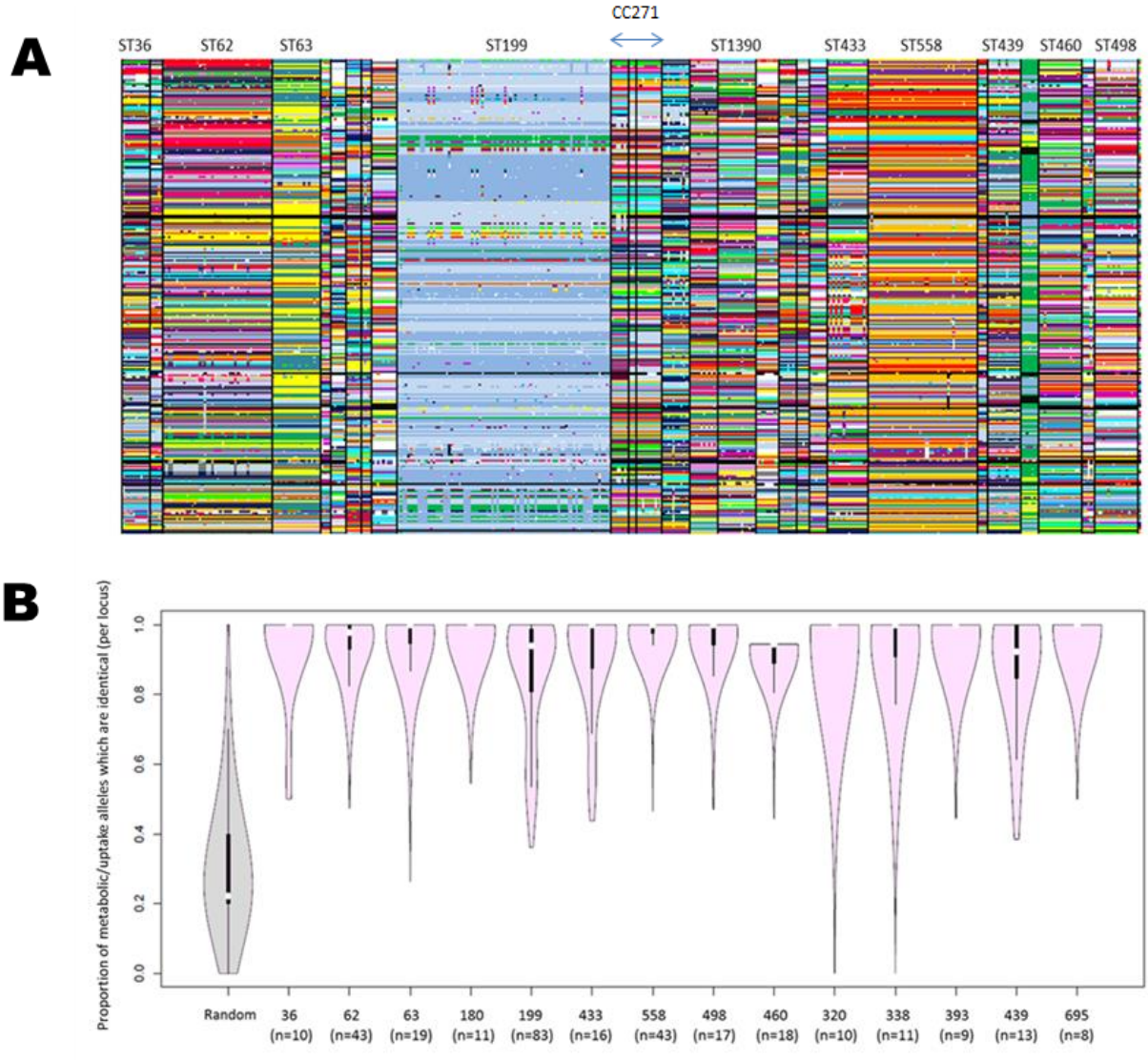


Figure 5.2 Analysis of the metabolic/transport alleles among MLST-defined STs (among 616 pneumococcal isolates sequenced by Croucher et al., 2013). *(A) Partial screenshot of the metabolic profiles of 616 strains, grouped by MLST-defined ST. (B) Violin plots showing the distribution of the modal percentage identity (MPI) of metabolic/transport alleles for the most frequent STs in the Croucher et al. dataset. The MPI for each ST (pink) was significantly higher than a distribution of randomly selected isolates (grey). The plots were produced using the "vioplot" package in R (version 2.15.1).*

To explore if metabolic/transport loci were significantly more uniform within individual STs than across randomly selected isolates, we performed a Sign Test on the differences obtained by subtracting MPIs within a random sample from MPIs at the same loci within a given ST. Loci containing alleles in which one or more isolates were truncated or deleted were discarded from the analysis, giving a total of 769 loci.

The median difference between the percentages of identical alleles within each ST and the percentages across randomly selected isolates was significantly different from zero across all STs included, with the number of positive differences consistently outweighing the number of negative differences (Table 5.3). This supports the hypothesis that the percentages of alleles identical at each locus are higher for within individual STs than for those of the randomly selected isolates.

Test	No. positive differences	No. negative differences	Ties	Test statistic	Standardised test statistic	Sample estimate	Upper achieved confidence interval (0.9566)	P-value (2-sided)
ST36 - Random	758	2	9	758	27.39	0.70	0.70 - 0.70	< 2.2e-16
ST62 - Random	762	3	4	762	27.41	0.68	0.66 - 0.68	< 2.2e-16
ST63 - Random	762	2	5	762	27.46	0.70	0.70 - 0.70	< 2.2e-16
ST180- Random	763	0	6	763	27.59	0.70	0.70 - 0.70	< 2.2e-16
ST199 - Random	754	13	2	754	26.72	0.62	0.60 - 0.64	< 2.2e-16
ST433 - Random	757	6	6	757	27.15	0.68	0.64 - 0.70	< 2.2e-16
ST558 - Random	763	2	4	763	27.48	0.70	0.70 - 0.71	< 2.2e-16
ST498 - Random	763	0	6	763	27.58	0.70	0.70 - 0.70	< 2.2e-16
ST460 - Random	763	6	0	763	27.26	0.64	0.64 - 0.64	< 2.2e-16
ST320 - Random	760	3	6	760	27.37	0.70	0.70 - 0.70	< 2.2e-16
ST338 - Random	758	6	5	758	27.17	0.62	0.62 - 0.70	< 2.2e-16
ST393 - Random	763	1	5	763	27.53	0.70	0.70 - 0.70	< 2.2e-16
ST439 - Random	760	5	5	760	27.26	0.62	0.60 - 0.65	< 2.2e-16
ST695 - Random	763	0	6	763	27.59	0.70	0.70 - 0.78	< 2.2e-16

Table 5.3. Outputs of the two-tailed Sign Tests comparing the average proportion of metabolic/transport alleles which are identical at a given locus (MPI), between isolates of the same ST and randomly-selected isolates. *The tests were performed for the most common STs in the dataset published by Croucher et al.. The sample estimate refers to the point estimate of the median difference between the two lists of modal percentage identities.*

5.5.4 Resolving Metabolic Structure within Serotypes of Lower MPIs

We calculated the MPIs for the most frequent STs and ST groups within six serotypes exhibiting relatively low MPIs (6A, 6B, 6C, 10A, 19F, 23F). STs were assigned to ST groups based on the percentage of identical metabolic/transport alleles shared at each locus (STs belonging to a given ST group share 80% of their metabolic/transport alleles), which were named according to the most frequent ST present. The MPIs of these STs/ST groups were much higher than their respective serotypes (Figure 5.3), suggesting that some serotypes exhibit more than one metabolic profile. Within some of these serotypes, however, a number of isolates did not correspond closely to a given ST (such as serotype 19F, in which only 22 out of 33 isolates mapped to a defined ST group). The large amount of genetic variation reflects the frequent recombination which occurs among pneumococci (Croucher et al. 2011).

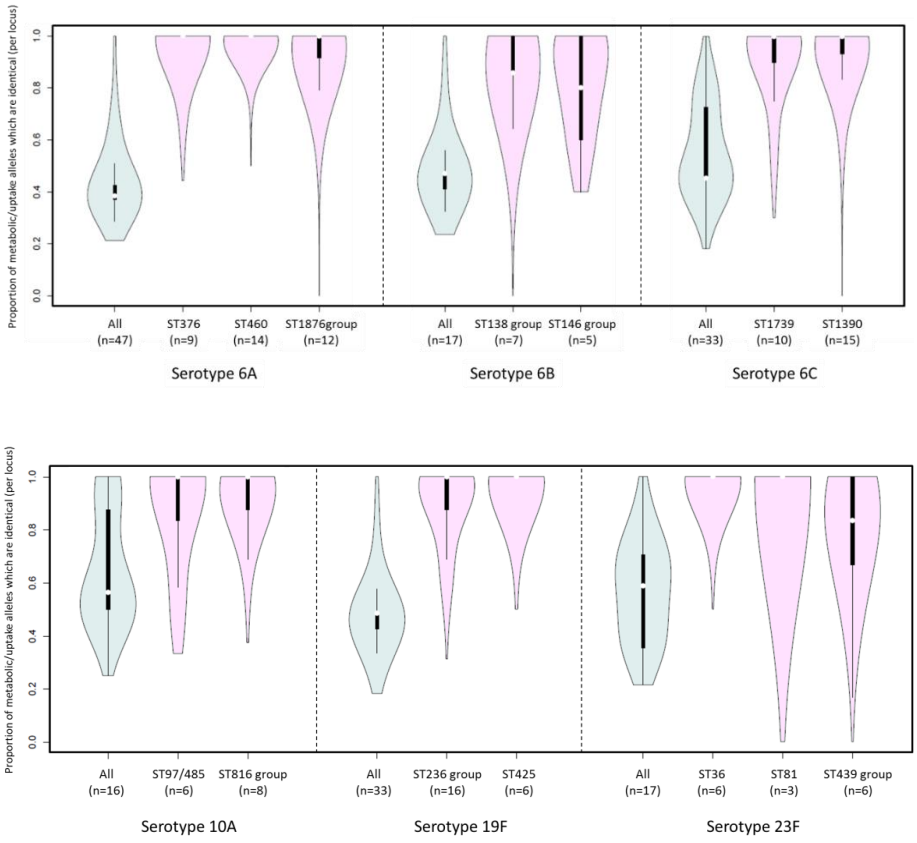


Figure 5.3. Violin plots showing the MPI of metabolic/transport alleles for 6 serotypes showing diverse metabolic profiles (blue), and the MPI of their respective STs (pink).

5.5.5 Linkage Disequilibrium and Mutual Information Values for Metabolic and Non-Metabolic Genes

Values of D' decreased with increasing distance between selected loci, as expected (Figure 5.4), and were significantly higher among metabolic/transport loci than non-metabolic loci (Wilcoxon Test: W=615, p<0.0001). We obtained similar results with the Mutual Information (MI) metric on the same sample (Wilcoxon Test: W=606, p<0.0001) (Figure 5.5).

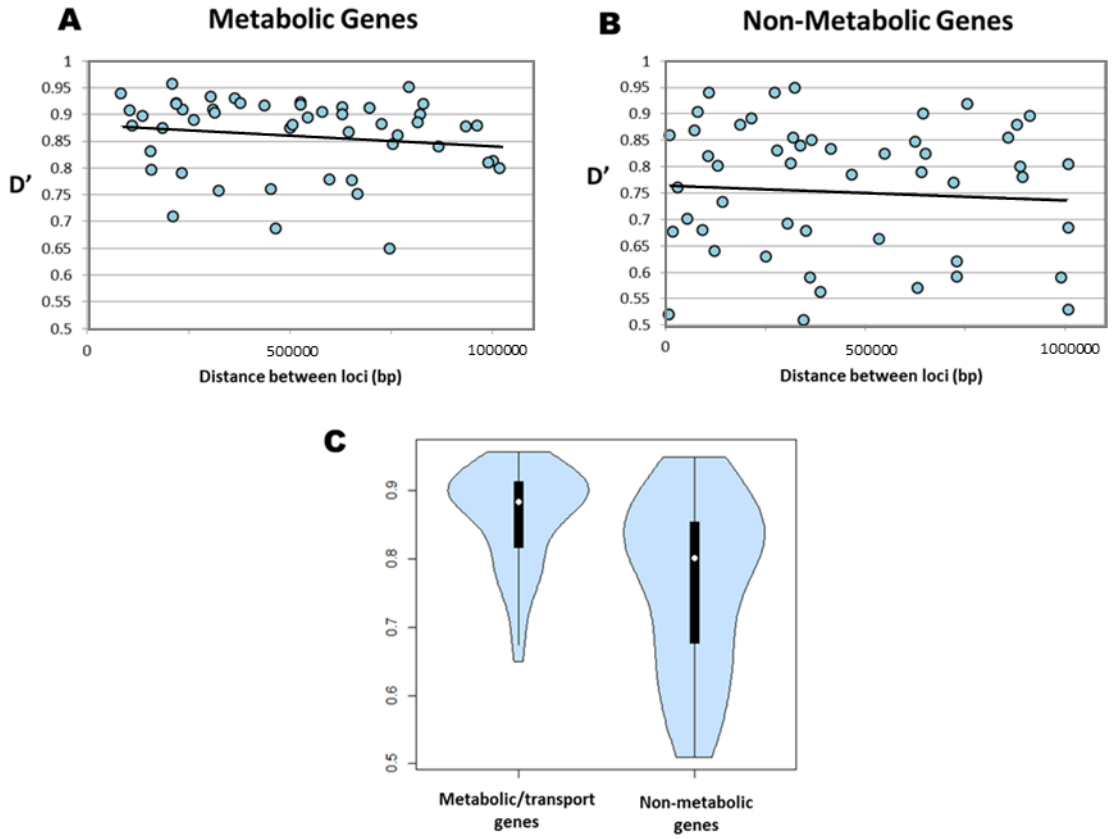


Figure 5.4. Linkage disequilibrium (D') observed between pairs of randomly-selected metabolic and non-metabolic loci. *(A) Values of D' between pairs of metabolic loci, plotted against their genetic distance on the chromosome. (B) Values of D' between pairs of non-metabolic loci, plotted against their genetic distance on the chromosome. (C) Violin plots showing the distribution of D' values among metabolic and non-metabolic pairs of loci.*

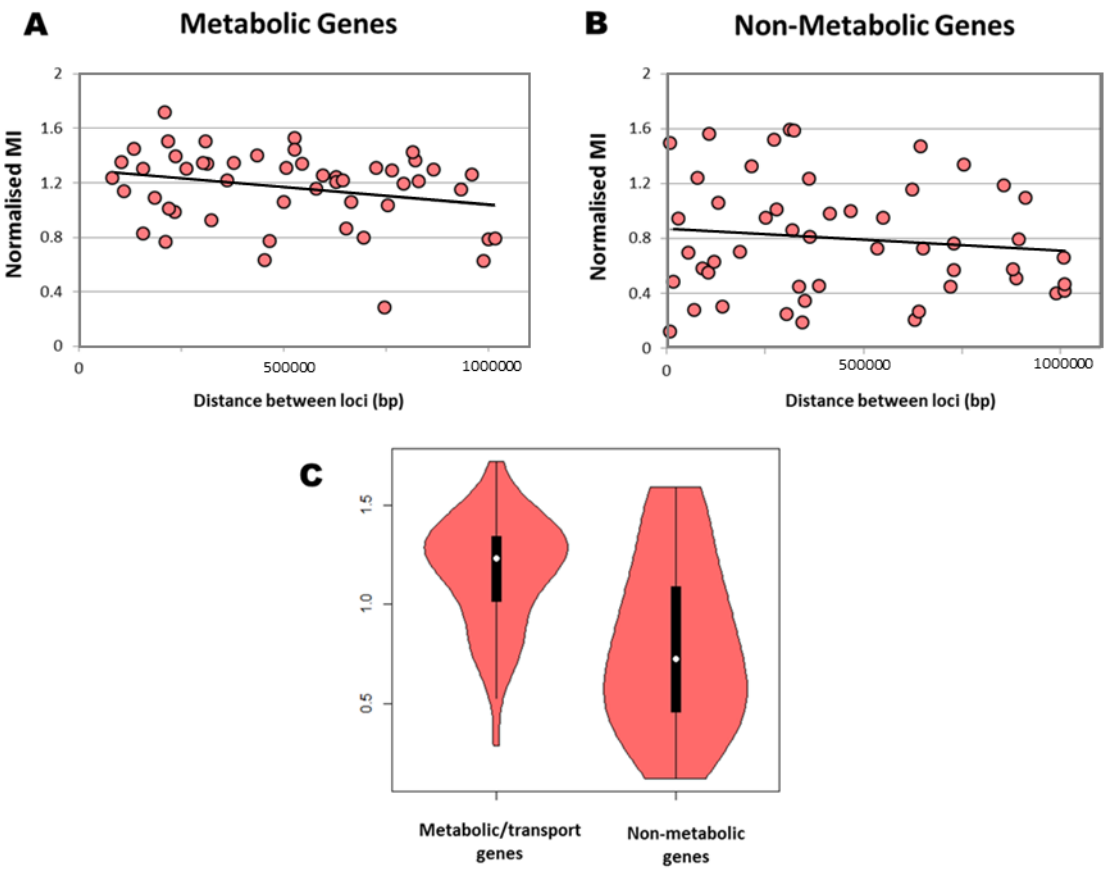


Figure 5.5. The strength of association between pairs of randomly-selected metabolic and non-metabolic loci, as measured by the Mutual Information (MI) metric. *The values of MI were normalised to account for inequalities in the number of alleles at different loci. (A) Values of MI between pairs of metabolic loci, plotted against their genetic distance on the chromosome. (B) Values of MI between pairs of non-metabolic loci, plotted against their genetic distance on the chromosome. (C) Violin plots showing the distribution of MI values among metabolic and non-metabolic pairs of loci.*

5.6 Discussion

Our findings here, of associations between pneumococcal serotype and genomic metabolic profile, support the predictions of the models presented in this thesis of non-overlapping associations between antigenic and metabolic alleles. The correspondence between MLST-defined STs and metabolic alleles also lend confidence to the results of previous studies of associations observed between pneumococcal STs and serotype (Table 1.4). Such findings are also consistent with observations that isolates belonging to the same ST tend to cluster on

phylogenetic trees drawn using whole genome data (Donati *et al*. 2010; Dagerhamn *et al*. 2008).

The metabolic/transport alleles showed large amounts of diversity across the isolates, with not one locus showing an identical allele conserved across all 616 strains. The metabolic profiles also differed extensively between serotypes, with pairs of serotypes sharing approximately 10.1% of their metabolic/transport alleles. Pneumococci seem to have a relatively high reliance on sugar metabolism compared to other respiratory tract organisms, which tend to rely more on carboxylates and other compounds for their carbon needs (Tettelin *et al*. 2001). It has been suggested that specialisation in this way may enable niche differentiation in the nasopharynx between different bacterial species (Tettelin *et al*. 2001). It could also be hypothesised that pneumococci exhibit a large diversity in genes involved with carbohydrate metabolism at the population-level to facilitate niche differentiation among individual strains, as well as among other species. The wide variety of carbohydrate sources available in the nasopharynx (Buckwalter & King 2012), and variable distribution of transporters and other metabolic components observed among pneumococci, suggests that niche differentiation at the strain level may be conceivable (Bidossi *et al*. 2012). The lack of allelic conservation in metabolic/transport genes across the genome suggests that this variation may be relatively old, and that the process of niche differentiation occurred early on in the evolution of distinct lineages/serotypes. This is consistent with the observation in studies of *S. pneumoniae* that metabolic genes tend to be stable over short time-scales while those related to virulence, antibiotic resistance and antigens tend to evolve more rapidly (Loman *et al*. 2013).

The high values of measures of association found here between alleles of metabolic/transport genes (relative to coding functional genes not involved in transport/metabolism) support the hypothesis that the metabolic profile may comprise a set of co-evolved genes, which have synergistically adapted to exploit a particular metabolic niche. Pneumococcal loci, as with all

bacteria, should show high levels of linkage between loci as they reproduce clonally, but we are not aware of an alternative explanation which could account for the higher levels of association observed between metabolic/transport loci than non-metabolic/transport loci. This is the first time, to our knowledge, that measures of association have been investigated among the metabolic alleles of a bacterial pathogen.

It can be observed that a number of serotypes comprise multiple metabolic profiles (as opposed to each serotype showing an association with one genomic profile) (Figure 5.3), and that multiple serotypes can exhibit the same metabolic profile, such as 19A and 15B/C. This population sample may therefore correspond to a roughly non-overlapping population structure (corresponding to a blue or purple structure in Figure 3.13) characterised by intermediate levels of ecological and immunological competition, rather than a strictly non-overlapping structure (corresponding to a red structure in Figure 3.13). It is also possible, because the isolates were sampled following a period of mass vaccination, that the population is in "flux" and hasn't yet settled down to a stable equilibrium. Another putative explanation for the roughly non-overlapping associations may be the asymmetries in the strength of serotype-specific immunity observed among strains. Although, epidemiological support for significant serotype-specific immunity has been provided for two of the serotypes observed with multiple STs here (6A and 23F).

A limitation of our approach is that we have only considered allelic variation and not included analyses of diversity between sequences on a nucleotide level. The number of nucleotide differences between different alleles is not immediately apparent, therefore, and it is not known whether the differences in alleles here translate into differences at the amino acid level. However, the large number of allelic variants observed at such a large number of loci suggests there may be some differences in metabolic function. Previous studies have found that even

small differences in amino acid sequence, including single amino acid changes, can alter protein conformation with knock-on effects for other interacting proteins (Coffey *et al*. 1995).

There are also a number of alternative explanations for the patterns observed here, other than ecological and immunological competition (reviewed in Chapter 8). Particularly pertinent to the results observed here is the high allelic diversity and simultaneous linkage disequilibrium in both metabolic and non-metabolic classes of genes, suggesting that perhaps the divergent patterns among serotypes are driven by within-serotype drift/differentiation across the whole genome. However, the allelic diversity and levels of measured association were higher among metabolic genes, suggesting that different evolutionary pressures may be operating.

5.7 Summary

Associations between serotypes and genome-wide metabolic profiles were observed in a published sample of 616 pneumococcal genomes, as well as between serotypes and MLST-defined STs. These findings lend support to the predictions of the models in this thesis, and also suggest that ST could represent a surrogate of metabolic type. The metabolic profiles of different serotypes tended to be highly distinct (sharing only 10% of metabolic/transport genes on average), and the metabolic/transport genes exhibited high association values, suggesting that pneumococcal strains may exploit specific metabolic niches, and the ability to do so entails a highly integrated network of interacting protein components.

Chapter 6.

Deterministic Frameworks combining Metabolic, Antigenic and Virulence Loci and the Effects of Strain-Targeted Vaccination

6.1 Public Dissemination of Work

These findings have been published in the following article, in association with the results in Chapters 3 and 5:

Eleanor R. Watkins, Bridget S. Penman, José Lourenço, Caroline O. Buckee, Martin C. J. Maiden & Sunetra Gupta (2015). Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. *Plos Pathogens*, 11(7): e1005034.

For the published article, ERW and SG conceived and designed the experiments; ERW performed the mathematical modelling and statistical tests; ERW analysed the genomic data and output from the mathematical models; and SG, ERW, JL, COB, and MCJM wrote the paper.

The findings have also been presented at several conferences as oral presentations and posters.

Oral Presentations:

Multivalent pneumococcal vaccines can increase the transmissibility and virulence of non-vaccine strains. Presented at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK for the meeting "Interdisciplinary Approaches to Understanding Microbial Communities" (October 2014); and at the Jacques Monod Conference, Roscoff, France, on "Infectious Diseases as Drivers of Evolution" (September 2014).

A new paradigm for vaccine replacement? Presented at the International Pathogenic *Neisseria* Conference, Wurzburg, Germany (September 2012).

Posters:

The effects of strain-targeted vaccination in bacterial populations. Presented at the Oxford Interdisciplinary Biosciences Industry Meeting. Oxford, UK. (March 2014).

6.2 Abstract

Bacterial genomes comprise a large number of coding sequences with a variety of functions. To increase the diversity of loci represented in the deterministic models in this thesis, an additional locus conferring virulence-associated properties was incorporated into the framework described in Chapter 3 comprising metabolic and antigenic genes. In this chapter, we aim to: (i) understand the effects of virulence on the population structuring induced by ecological and immunological competition; (ii) understand the effects of strain-targeted vaccination, in frameworks both with and without the virulence locus; and (iii) compare the results of the models to published data. We hypothesise that a variety of different population structures incorporating antigenic, metabolic and virulence loci will arise over a range of parameter space, but do not anticipate the appearance of specific structures. After vaccination, we would expect the dominant strains to exhibit the most transmissible metabolic types. Whether or not the post-vaccine strains manifest virulence determinants we expect will depend on the relative costs and benefits of virulence. We find that, following vaccination, the metabolic and virulence type originally associated with the vaccine-targeted antigenic type are observed in association with strains possessing antigenic types not included in the vaccine, a phenomenon we term "Vaccine Induced Metabolic Shift" (VIMS). We discuss our rationale, methods and findings in the context of pneumococcal conjugate vaccines, for which extensive replacement by non-vaccine serotypes has been reported following mass vaccination. The support for VIMS among *Streptococcus pneumoniae* is examined, using published MLST data and 616 published whole genomes isolated from Massachusetts, USA.

6.3 Introduction

An important challenge to the control of infectious disease by vaccination is the degree of antigenic diversity exhibited by many pathogen populations. One strategy that can be been employed is to target the antigens of only a subset – typically, the most virulent – of the co-circulating strains. Such strain-targeted vaccines are currently in use against *Streptococcus pneumoniae*, *Neisseria meningitidis*, *Haemophilus influenzae*, Human Papilloma Virus, Rotavirus, and veterinary vaccines such as the Foot and Mouth Disease Virus, for example. An important consideration in the deployment of these vaccines is whether they will alter the ecology of the host-pathogen system in a manner that renders them ineffective. Strain-targeted vaccines may lead to an increase in the prevalence of strains in the population which are not targeted by the vaccine, a phenomenon referred to as "Vaccine Induced Strain Replacement" (VISR). Significant increases in non-vaccine strains following vaccination have been reported in a number of pathogens, including *S. pneumoniae* and *H. influenzae* (Weinberger *et al*. 2011; Ribeiro *et al*. 2003); although it is difficult to conclude definitively that increases in non-vaccine strains can be attributed to vaccination, or to other factors, such as natural fluctuations or antibiotics usage (Jefferies *et al*. 2010). One of the most striking examples of VISR is provided by *S. pneumoniae*.

The first pneumococcal conjugate vaccine, PCV7, was licensed in 2000 and targets 7 serotypes (4, 6B, 9V, 14, 18C, 19F, and 23F) (Black *et al.* 2000). Although the widespread use of this vaccine resulted in a decrease in disease caused by the vaccine serotypes, it led to a concomitant increase in carriage and disease caused by serotypes not included in the vaccine (reviewed by Weinberger *et al.* 2011) (see also Section 1.4). Evidence for serotype replacement has arisen in a number of countries on a global scale, including Norway, France, Portugal, Germany, Spain, the United States and Canada (Figure 6.1). Serotype 19A is a particularly

common non-vaccine serotype observed in the post-vaccine era in many countries. Several vaccine escape clones, such as those associated with 19A, are also associated with high levels of antibiotic resistance.

There are a number of frameworks which model the population-level increase of non-vaccine strains following vaccination. Many such frameworks invoke strong competition – mediated either directly through competition within the host (Lipsitch 1997), or indirectly through cross-immunity (White *et al*. 1998; Porco & Blower 1998; Cohen *et al*. 2008; McLean 1995; Gupta *et al*. 1997; Restif & Grenfell 2007; Cobey & Lipsitch 2012), or both (Zhang *et al*. 2004). Vaccination releases competition from the vaccine strains through their removal from the population, thereby allowing non-vaccine strains to fill their niche.

None of the existing theoretical frameworks so far, however, have considered the possibility of constitutive changes occurring among non-vaccine strains as a result of strain-targeted vaccination. Here, we use the frameworks presented in Chapter 3 to explore the effects of strain-targeted vaccination in model systems where strains are defined by antigenic and metabolic loci. We also extend this model framework with the addition of a virulence-associated locus, and, after exploring the various population structures which arise, show that strain-targeted vaccination may lead to an increase in transmissibility and virulence of non-vaccine strains. We discuss our results in the context of pneumococcal conjugate vaccines, and show that our results are consistent with genotypic typing data, as well as analyses of 616 whole genomes published by Croucher *et al*. (2013).

6.3.1 Model Assumptions

Within the model framework, strains with identical alleles at the virulence-associated locus are not able to co-infect a host. We assume that direct ecological competition occurs between

strains expressing identical virulence alleles on the basis that these factors interact directly with host tissues and bind with specific host cell components across a range of pathogens (Pizarro-Cerdá & Cossart 2006). In *S. pneumoniae,* for example, a number of virulence factors bind to specific host target receptors, such as the pili. Pili are long organelles that are able to extend beyond the polysaccharide capsule, and have been shown to promote adhesion in human lung cells, and the development of invasive disease in mouse models (Barocchi *et al*. 2006; Bagnoli *et al*. 2008). There are two types of pili in pneumococcus, type I (PI-1) and type II (PI-2), which are found in approximately 30% and 16% of strains respectively. It has been suggested that pili bind specific components of the host extra- extracellular matrix (Schwarz-Linek *et al*. 2004), and it is thus possible that the co-colonisation of strains sharing such virulence determinants would be inhibited. Significant negative associations have been observed, for example, in co-colonisation of piliated pneumococci and *Staphylococcus aureus* (Regev-Yochay *et al*. 2009); it would be reasonable to assume that the same type of competition occurs between strains of piliated pneumococci. Other assumptions were consistent with those of the model framework in Chapter 3.

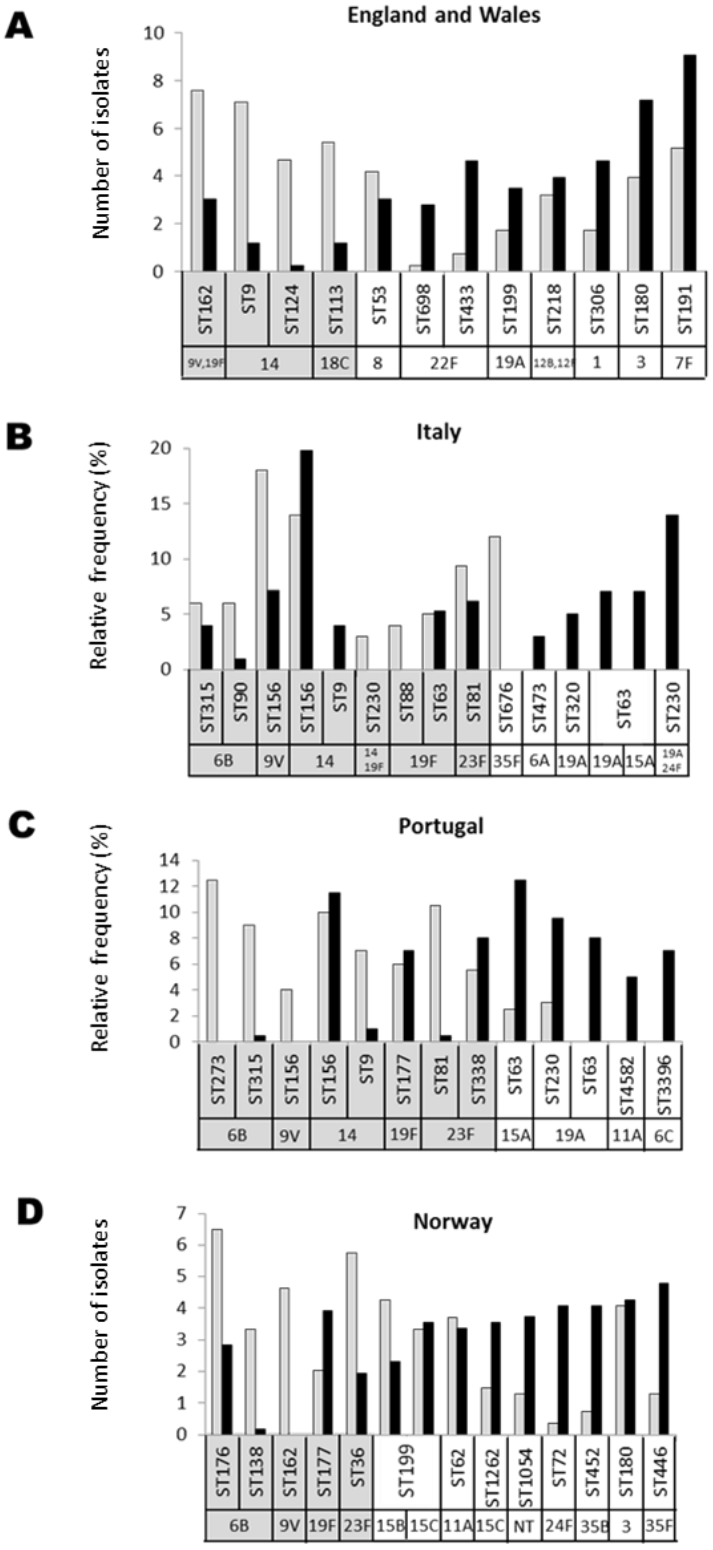


Figure 6.1. Serotype replacement in disease (A-C) and carriage (D) following PCV7 vaccination. *Grey and black bars represent the prevalence before and after vaccination with PCV7, respectively. (Data taken from: (A) Pichon et al. 2013; (B) Gherardi et al. 2012; (C) Simões et al. 2011; (D) Vestrheim et al. 2010). Serotypes shaded in grey are included in PCV7.*

6.4 Methods

6.4.1 Addition of Virulence Locus

The general framework described in Chapter 3, comprising antigenic and metabolic loci, may be extended to accommodate additional non-capsular virulence factors by expanding equations (3.6) and (3.7) to include competition at a third virulence locus. We assume there are two alleles at the virulence locus (+ and -), with (+) strains representing strains with higher virulence propensity than (-). The notation of antigenic type (AT; a, b, c) and metabolic type (MT; 1, 2, 3), is consistent with the models presented in Chapter 3. The system can be described as follows:

$$\frac{dy_i}{dt} = \lambda_i \left(1 - z_j\right) \prod_k (1 - y_k) - \sigma_i y_i \tag{6.1}$$

$$\frac{dz_j}{dt} = \left(1 - z_j\right) \sum_i \lambda_i \prod_k (1 - y_k) - \mu z_j \tag{6.2}$$

Where y_i is the proportion of the population infected by strain *i* (e.g. *a1+*), z_j is the proportion of the population immune to AT *j*, μ is the average host death rate, and σ_i is the rate of loss of infectiousness associated with strain *i*. The rate of loss of infectiousness is determined by the virulence locus, with virulent (+) strains exhibiting a higher rate of loss of infectiousness (and thus shorter duration of infection, 1/σ) than (-) strains. We assume that infection by a particular strain (e.g. *a1+*) can only occur among individuals who are not immune or infected with the same AT (ie 1-z_a), thus *j* represents the AT of strain *i*. We also assume that infection cannot occur among individuals currently infected by other strains with either the same MT or virulence factor (denoted in the equations above by *k*) and encompassing, for example *b1+*,

b2+, *b1-*, *c1+*, *c1-*, *etc* for strain *a1+*. The force of infection (λ) of a strain is given by $\lambda_i = \beta_i y_i$ where β_i is the mean of the transmission coefficients associated with MT and virulence type of the strain. The transmission coefficient for (+) strains always exceeds that of (-) strains. The basic reproductive number, R_0, for strain *i* is given by β_i/σ_i. All possible strains (with all allelic configurations) are initially present at nearly equal frequencies when simulations are performed.

As an example, in the case where there are two ATs (*a* and *b*), two MTs (1 and 2) and two virulence types (+ and -), the rate of change in the host population infected with strain a1+ (y_{a1+}) is given by:

$$\frac{dy_{a1+}}{dt} = \lambda_{a1+}(1 - y_{b1+})(1 - y_{b1-})\,(1 - y_{b2+})(1 - z_a) - \sigma_{a1+} y_{a1+} \qquad (6.3)$$

Thus a host infected with strain a1+ cannot be co-infected with any of the following strains: a1+, a2+, b1+, b1- or b2+.

6.4.2 Simulating Vaccination

Vaccination may be included by allowing the proportion with specific immunity to vaccine strains to increase at a rate v, following Gupta *et al.* (1997):

$$\frac{dz_i}{dt} = \sum_{j=1}^{m} (\lambda_{ij}\,(1 - z_i) \prod_{w=1}^{n} (1 - y_{wj})) - \mu z_i + v(1 - z_i) \qquad (6.4)$$

6.4.3 Incorporating Non-Specific Immunity

Cross-immunity between ATs can be introduced by including a new variable w describing the proportion of the population immune to any AT whose dynamics are given by:

$$\frac{dw}{dt} = \sum_{\mathrm{i,j,k}} \lambda_{\mathrm{ijk}} (1 - w) - \mu w \tag{6.5}$$

The proportion of the host population infected with strain *ijk* (y_{ijk}) now becomes:

$$\frac{dy_{ijk}}{dt} = \lambda_{ijk}((1 - w) + (1 - \theta)(w - z_i)\prod_{m,l}(1 - y_{i} \tag{6.6}$$

$$- \sigma_k y_{ijk} \tag{6.6}$$

The parameter θ represents the strength of cross-immunity: when θ=0, strains are not able to infect hosts who are immune to (or currently infected with) strains possessing the same AT, but can freely infect hosts immune to or infected with a different AT. When θ=1, full cross-immunity operates between strains. Vaccine-induced cross-immunity can be incorporated by increasing w by a small constant.

6.5 Results

6.5.1 Population Structures obtained in a Framework with Antigenic, Metabolic and Virulence-Associated Loci

The baseline model produced a variety of outcomes which correspond to a range of population structures – from obligate virulent pathogens to avirulent commensals, with clonal to more diverse levels of genetic diversity (Figure 6.2). Each population structure can be broadly conceptualised as corresponding to specific pathogen species (Table 6.1). Within each population structure, three strains predominate at equilibrium (black boxes, Figure 6.2), as the others have been competitively excluded. The specific combinations of strains which arise depend on the differences in fitness between metabolic alleles, and the relative transmission benefit and virulence cost conferred to strains with the + virulence allele.

Co-existence of all metabolic alleles occurs only at small differences in transmissibility between metabolic alleles (Figure 6.2: *yellow, pink, red*). When differences in transmissibility between metabolic alleles are too large, the least transmissible alleles are subject to competitive exclusion and the population becomes dominated by only one metabolic allele (Figure 6.2: *grey, green*). When all three metabolic alleles co-exist, they are associated with unique ATs in non-overlapping combinations, a result consistent with the stochastic model of Buckee *et al*. (2008).

Strains with additional virulence factors are able to dominate the population when the benefit of increased transmission significantly outweighs the cost of a decreased infectious period (Figure 6.2: *pink, green*). This corresponds broadly to obligate pathogens (Table 6.1). As the cost of virulence is increased, avirulent strains can emerge (Figure 6.2: *red, blue*) until all virulent strains are outcompeted (Figure 6.2: *grey*). These population structures correspond to

commensal species: when the cost of virulence is too high, it pays for all the strains in a species to be avirulent.

The blue population structure, corresponding to partially overlapping associations between ATs and MTs, represents an intermediate outcome in which the fittest MT is able to outcompete one but not both of the other MTs, and the virulence allele arises in two strains. The non-overlapping associations observed between antigenic and metabolic alleles in a number of pathogens are not absolute: some pneumococcal serotypes are, for example, stably associated with multiple STs. Thus, for certain strains among these populations, the overlapping blue structure is more applicable.

When observed, the high virulence alleles tended to associate with the most transmissible MT, a result also in accordance with the deterministic model of Buckee *et al*. (2008). This supports the hypothesis that only the most transmissible lineages can afford the cost of virulence.

Many bacterial pathogens exhibit the "red" population structure in which unique combinations of ATs and MTs may arise with only some of these exhibiting virulence (including the pneumococcus - Figure 6.2b). Such a structure arises because, for example, although ***b1+*** may be intrinsically more transmissible than ***b2-***, the latter has the competitive advantage while ***a1+*** is in circulation because it differs from ***a1+*** at all three loci. For similar reasons, ***c3-*** outcompetes ***c1+, c1-, c2+, c2-*** and ***c3+***, even though they may all have higher intrinsic transmissibility. Vaccination was therefore simulated within the parameter space mapping to this red population structure. As a number of bacterial populations may also exhibit the blue structure, vaccination was also simulated within this area of parameter space.

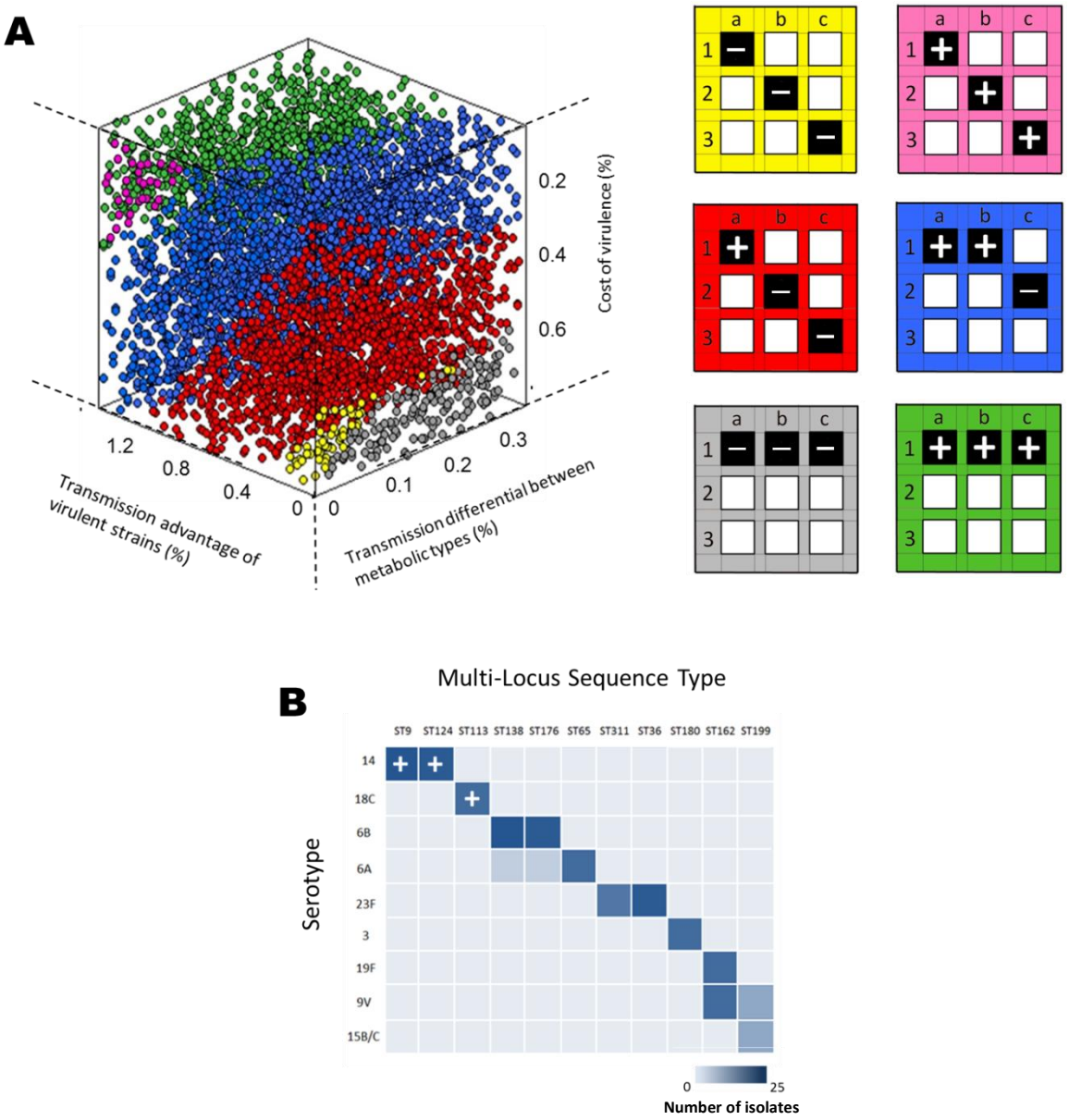


Figure 6.2. Population structures generated by the multilocus model incorporating metabolic, antigenic and virulence-associated loci (A), and the observed structure of* S. pneumoniae *(B).

(A) Left: Distribution of population structures (marked by different colours), with each point corresponding to a deterministic outcome for a random set of parameters (selected using Latin Hypercube Sampling). Right: Population structures are represented in terms of associations between antigenic (a, b, c), metabolic (1, 2, 3) and "excess" virulence (+, -) alleles. Strain combinations which dominate at equilibrium are marked in black. {$\beta_1 = 4.5$, $\beta_2 = \beta_1 (1+\Delta\beta)$, $\beta_3 = \beta_2 (1+\Delta\beta)$ where $\Delta\beta$ is the transmission differential between metabolic types; $\beta_- = 4.5$, $\beta_+ = \beta_-(1+\Delta\beta_v)$ where $\Delta\beta_v$ is the transmission advantage of virulent strains; $\sigma_+ = 2.99$, $\sigma = \sigma_+ (1-\Delta\sigma)$ where $\Delta\sigma$ is the relative cost of virulence; $\mu = 0.02$}.

(B) Associations between the most frequent pneumococcal serotypes and MLST-defined STs, England (data taken from Brueggemann et al. ,2003). (STs with invasive disease odds ratios which are significantly greater than 1 are indicated by the + sign).

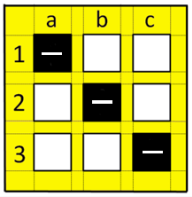
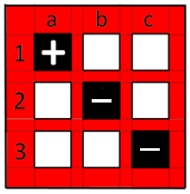
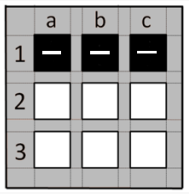
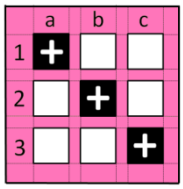
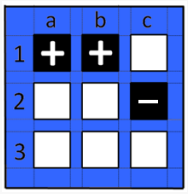
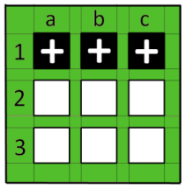
Population Structure	Examples	Details
	*Neisseria lactamica*	Association between sequence type (ST) and FetA epitope (Bennett *et al*. 2009)
	*Streptococcus pneumoniae* *Haemophilus influenzae* *Neisseria meningitidis* *Streptococcus pyogenes*	Association between serotype and ST (see Table 1.3) Association between serotype and clonal complex (eg. Meats *et al* 2003) Association between PorA, FetA and clonal complex (eg. Buckee *et al*. 2010). Associations between ST and emm type (eg. McGregor *et al*. 2004)
	*Streptococcus vestibularis*	Limited variation in housekeeping alleles across serotype (Delorme *et al*. 2007)
	*Neisseria gonorrhoea*	Associations between Opa variants and STs (Bilek *et al*. 2009)
	*Streptococcus agalactiae*	Associations between capsular serotype and ST did not always correspond (Jones *et al*. 2003).
	*Mycobacterium tuberculosis*	No variation in housekeeping alleles but some antigenic variation shown by monoclonal antibodies (Coates *et al*. 1981)

Table 6.1. Examples of pathogens which may exhibit the population structures from the multilocus model incorporating three metabolic alleles, three antigenic alleles and two virulence-associated alleles. *Additional references for the examples of the "red" population structures can be found in Table 1.4.*

6.5.2 Effects of Strain-Targeted Vaccination

6.5.2.1 Framework comprising Antigenic and Metabolic Loci

Figure 6.3 reports the dynamics of the nine strain framework characterised by three antigenic and three metabolic alleles under strain-targeted vaccination. In the pre-vaccination phase, the population adopts the non-overlapping structure {a1,b2,c3} despite all combinations initially being present at nearly identical frequencies and being constantly generated at a low rate thereafter. We then observe that vaccination against antigenic type *a*, which is associated with the most transmissible MT (here MT1), does not result in the elimination of MT1; instead MT1 emerges, after vaccination, in association with the non-vaccine antigenic type *c*. Thus the pre-vaccination profile {*a*1, *b*2, *c*3} is altered to {*b*2, *c*1}. We term this phenomenon Vaccine Induced Metabolic Shift (VIMS). This result holds when ecological competition and variant-specific immunity are relaxed, so long as vaccination is simulated within the "red" areas of parameter space corresponding to non-overlapping associations of AT and MT in Figure 3.13.

It is noteworthy that this result is only observed if the transmissibility of the strains is determined by the MT. If the transmissibility of the strains is determined by the AT, there is no fitness advantage to MT1, and thus, MT1 will not emerge following vaccination (Figure 6.4), with the predominant strains remaining unaltered. It is also of note that VIMS is only observed when the vaccine AT is completely eradicated. Simulations at lower rates of vaccination showed that, when the vaccine AT is still present in the population - albeit at a low prevalence - VIMS will not occur as the metabolic and virulence niche inhabited by the vaccine strain is not vacated (Figure 6.5).

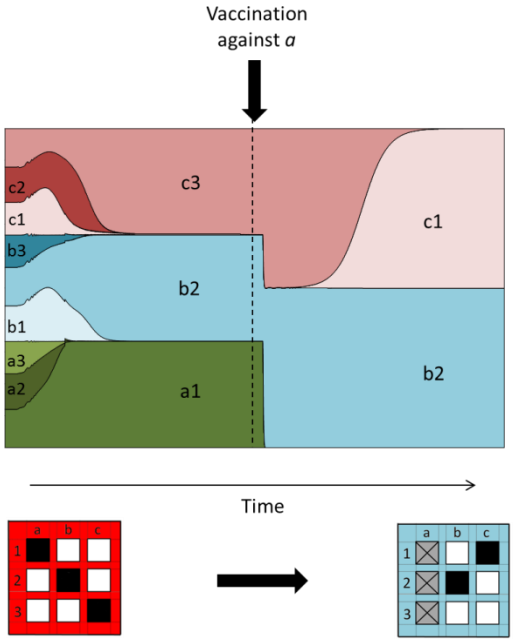


Figure 6.3. Vaccine induced metabolic shift observed in a bacterial population displaying non-overlapping associations between antigenic (a, b, c) and metabolic (1, 2, 3) alleles, when β is determined by the MT of a strain. *Following vaccination against antigenic type* ***a****, the population structure {****a****1,****b****2,****c****3} shifts to {****b****2,****c****1}. (*$\sigma = 3$; $\beta_3 = 4.5$; $\beta_i = \beta_{i-1} + \Delta\beta$, *for* $i=1,2$; $\Delta\beta = 0.001$; $\nu = 0.01$; $\mu = 0.02$*).*

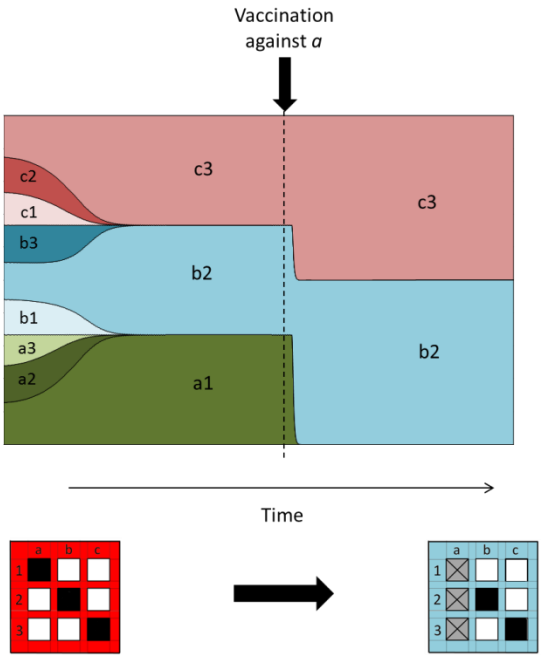


Figure 6.4. The effects of strain-targeted vaccination when β is determined by the AT of a strain. *After* ***a*** *declines following vaccination, the overall population structure remains unchanged. (*$\sigma = 3$; $\beta_A = 4.5$; $\beta_I = \beta_{i-1} + \Delta\beta$, *for* $i=1,2$; $\Delta\beta = 0.001$; $\nu = 0.01$; $\mu = 0.02$*).*

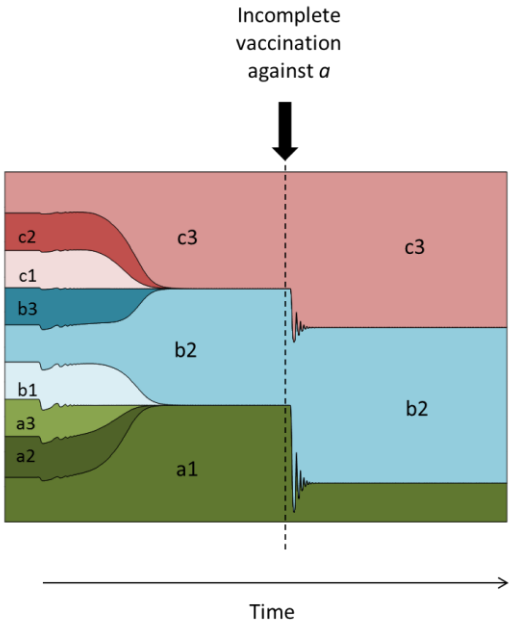


Figure 6.5. Effects of imperfect strain-targeted vaccination. *Strain c1 is not able to emerge as a1 is still prevalent in the population. ($\sigma = 3$; $\beta_3 = 4.5$; $\beta_i = \beta_{i-1} + \Delta\beta$, for i=1,2; $\Delta\beta = 0.001$; $v=0.005$; $\mu = 0.02$). (Strain a1 is removed from the population, and c1 able to emerge, when $v>0.0055$).*

6.5.2.2. Framework comprising Antigenic, Metabolic and Virulence Loci

Following vaccination in the "red" structure against the most virulent antigenic type *a*, strain c1+ increases in prevalence, as it had been hitherto suppressed by competition with *a1+* (Figure 6.6). This is because (i) it does not share alleles at the metabolic or virulence loci with the already co-circulating strain b2- (ii) it outcompetes c3+ (which also does not share alleles with b2-) due to MT1 having a higher transmission efficiency. Thus, we again observe VIMS, but here, vaccination against a virulent AT can lead to a non-vaccine antigenic type bearing the virulence determinants in addition to the metabolic components originally carried by the vaccine AT.

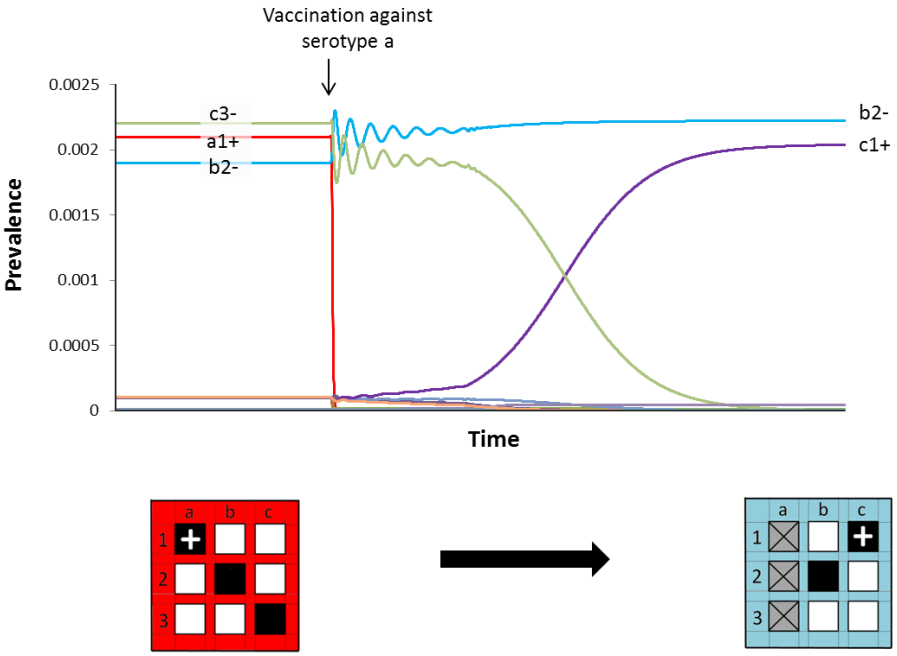


Figure 6.6. Strain dynamics following vaccination against antigenic type a. *The previously suppressed strain* ***c1+*** *expands and competitively excludes strain* ***c3-,*** *thus antigenic type c increases in both transmission efficiency and virulence potential (β_1=4.5, β_2=4.9995, β_3=4.99, σ_+=3, σ_-=2.997, μ=0.02).*

Whether or not the virulence determinant was observed in association with the non-vaccine AT after vaccination depended on the transmission coefficient (β) conferred to (+) strains, relative to the cost (σ). When the fitness of (+) strains surpassed a certain threshold, the vaccine escape strain exhibited the virulence determinant (Figure 6.7a, turquoise). Below this threshold, when the ratio of the R_0s of (+) strains to (-) strains was below 0.995, the vaccine escape strain was not able to bear the fitness of cost of virulence (Figure 6.7a, orange). In contrast, vaccination against both virulent ATs, a and b, resulted in the emergence of virulent strain c1+ across all parameter space (Figure 6.7b). This can be attributed to the lower costs of virulence in this area of parameter space, relative to that of the "red" population structure.

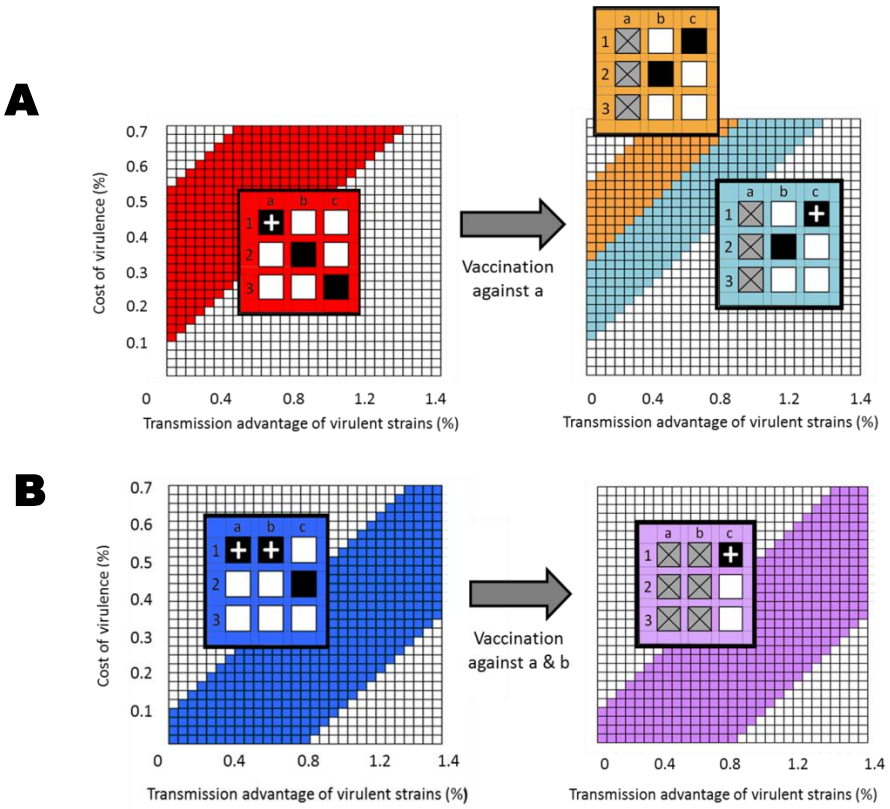


Figure 6.7. Shift in the metabolic type of non-vaccine strains may be accompanied by an increase in virulence. *(A) Whether or not vaccination against antigenic type* a *results in an increase in virulence depends on the relative fitness cost and benefit of bearing the high virulence allele. (B) Vaccination against antigenic types a and b in the blue population structure consistently results in the emergence of a non-vaccine antigenic type bearing the virulence determinant.*

6.5.2.3 Effects of Vaccine-Induced Cross-Immunity

The effects of vaccine-induced and naturally-acquired cross-immunity on VIMS were investigated in a two-allele, four strain framework (a1-, a1+, b1-, b1+). The strains which dominated at equilibrium depended on the strength of naturally-acquired cross-immunity θ. Before vaccination (blue, Figure 6.8), the dominant strains changed from (a1+, b2-) at low levels of θ between strains (A); to (a1+, b1-) over intermediate to high levels of θ (B); and at complete cross-protection, only one strain was able to emerge (that which had the highest starting frequency). Following vaccination against virulent antigenic type *a*, VIMS occurred at low levels of cross-protection (Figure 6.9), with the emergence of strain b1+. At intermediate

levels of θ, b1+ emerges following vaccination. Thus, VIMS occurs when naturally-acquired cross-immunity operates between strains, but only at relatively low levels of θ (<0.33).

The magnitude of the increase in non-vaccine strain b1+ depended, however, on the presence of vaccine-induced cross-immunity relative to naturally-acquired cross-immunity. When the strength of vaccine-acquired cross-immunity is equal to that of naturally-acquired cross-immunity, the post-vaccination prevalence of b1+ cannot exceed the pre-vaccination prevalence of a1+, and complete vaccine-induced strain replacement is thus not possible (Figure 6.10). When vaccine-induced cross-immunity is not present, the prevalence of strain b1+ after vaccination exceeds that of a1+ before vaccination.

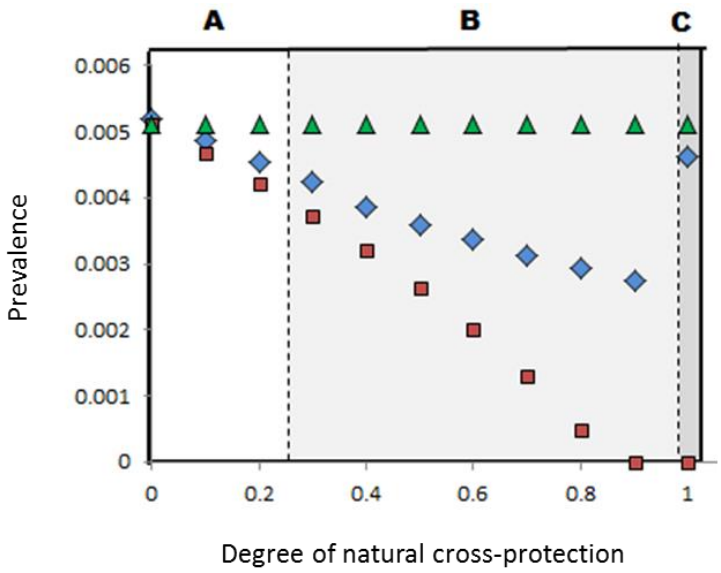


Figure 6.8. The effects of naturally-acquired and vaccine-induced cross-immunity on pre- and post-vaccination strain prevalence. *(Green: Post-vaccination with no cross-immunity in vaccine; blue: pre-vaccination; red: post-vaccination where vaccine-induced cross-immunity equals naturally-acquired cross-immunity). (A) Prevalent strains are a1+, b2-. (B) Prevalent strains are a1+, b1-. (C) One strain dominates at complete cross-protection. ($\beta_1 = 4.001$; $\beta_2 = 4$; $\beta_- = 4$; $\beta_+ = 4.15$; $\sigma_+ = 2$; $\sigma_- = 1.95$; $\mu = 0.02$).*

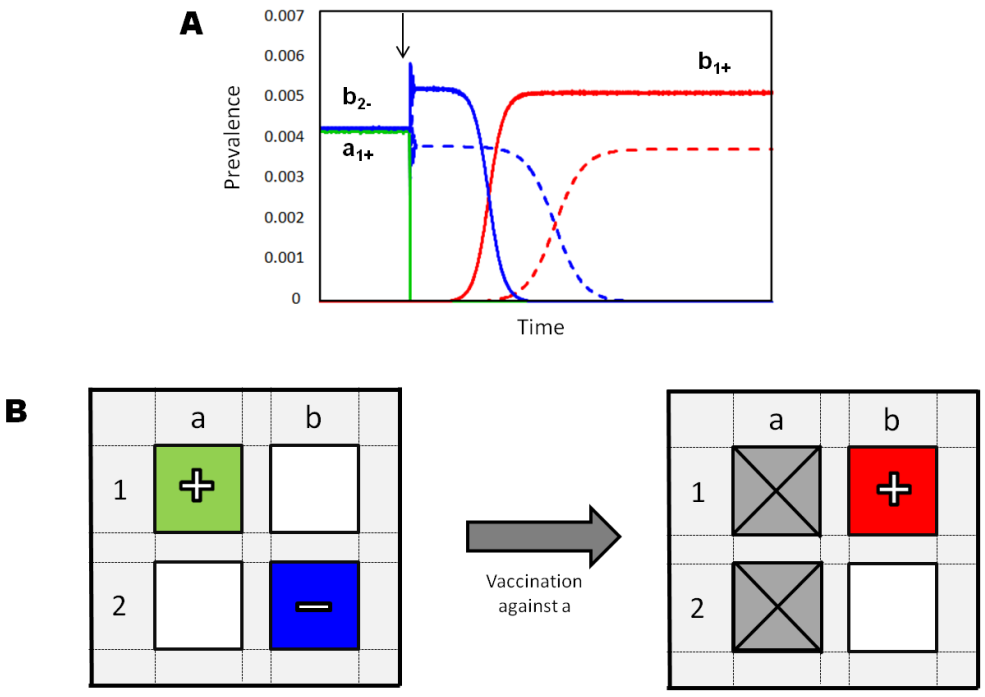


Figure 6.9. Vaccine-induced cross immunity and vaccine-induced metabolic shift in an 8-strain framework. *(A) Dynamics of replacement with natural cross-immunity between serotypes a and b, with (dotted line) and without (solid line) vaccine-induced cross-immunity. In both cases, vaccination against a leads to a metabolic shift and increased virulence in non-vaccine type b (b_{2-} ➔ b_{1+}), but an overall increase in b_{1+} can only occur if the cross-protection provided by natural immunity exceeds that provided by vaccine-induced immunity. (B) Schematic representation of the metabolic shift following vaccination against serotype a. ($\beta_1 = 4.001$; $\beta_2 = 4$; $\beta_- = 4$; $\beta_+ = 4.15$; $\sigma_+ = 2$; $\sigma_- = 1.95$; $\mu = 0.02$; $\gamma = 0.3$).*

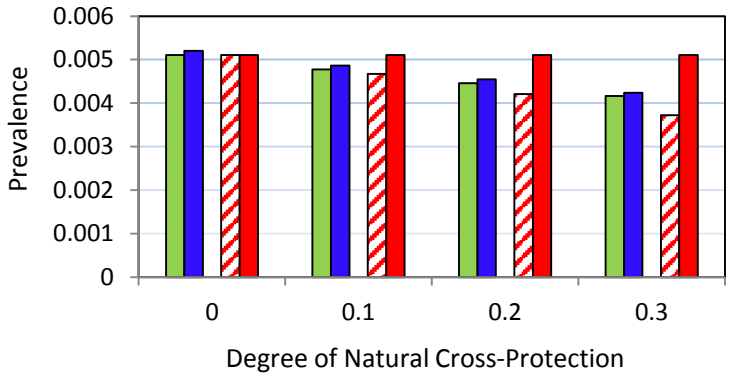


Figure 6.10. Pre-vaccination prevalence of a_{1+} (green) and b_{2-} (blue), and post-vaccination prevalence of b_{1+}, with (striped bars) and without (solid bars) vaccine-induced cross-immunity. *($\beta_1 = 4.001$; $\beta_2 = 4$; $\beta_- = 4$; $\beta_+ = 4.15$; $\sigma_+ = 2$; $\sigma_- = 1.95$; $\mu = 0.02$).*

6.5.3 Genomic Analysis

Although it can be inferred from the results of the mathematical models that the metabolic and virulence types from vaccine strains are observed in association with non-vaccine strains following vaccination, VIMS in real pneumococcal populations would manifest as a "capsule switch", in which the genetic background of the vaccine-targeted strain remains unchanged (including the metabolic and extra-capsular virulence genes), and these pneumococci inherit genetic material which code for a non-vaccine capsular polysaccharide.

Two putative cases of VIMS can be identified from the genomic data of the 616 published pneumococcal whole genomes (Croucher *et al*. 2013), previously analysed in Chapter 5. The metabolic profile corresponding to CC320 (comprising ST320, ST236 and ST271 strains) was observed primarily in association with vaccine serotype 19F in 2001, and with non-vaccine serotype 19A in 2004 and 2007. Similarly, the metabolic profile of vaccine serotype 9V is observed in association with non-vaccine serotypes 19A and 15B/C in 2007 (Figures 6.11 & 6.12). For the VIMS observed between vaccine serotype 9V and non-vaccine serotypes 15B/C and 19A, modal metabolic/uptake alleles were shared at 84.6% (694 out of 820) and 90.0% (737 out of 819) of loci respectively. For the VIMS observed between vaccine serotype 19F and non-vaccine serotype 19A, modal alleles were shared at 77.8% (633 out of 814) of loci. However; when only the isolates of 19F ST320 and ST271 were included in the analysis (excluding ST236), the percentage identity was 89.8% (729 out of 812 loci). The number of isolates from which VIMS can putatively be deduced is small, however. A total of 16 serotype 19F isolates were observed with the metabolic profile corresponding to CC320 in 2001, and 9 serotype 19A isolates in 2004 and 2007. Similarly, a total of six serotype 9V isolates were observed with the metabolic profile corresponding to ST156 in 2001, 2004 and 2007, and eight isolates belonging to serotypes 19A and 15B/C in 2004 and 2007.

We also observe that a number of alleles coding for virulence factors observed among vaccine serotype 19F strains in 2001 are observed with non-vaccine serotype 19A strains in 2004 and 2007; and similarly that virulence factors among vaccine serotype 9V strains are observed with non-vaccine serotypes 15B/C and 19A in 2004 and 2007 (Table 6.2). Strains of $CC320^{19F}$ shared alleles with 81% (17 out of 21) of virulence-associated loci of $CC320^{19A}$ strains isolated in 2007. The virulence-associated loci exhibited high levels of allelic diversity: the number of alleles present at each locus varied from 28 to 139, with an average of 73.8 alleles across the 616 genomes. Further, serotype 19A strains of ST199 – the most prevalent 19A strain prior to vaccination – shared no virulence-associated alleles with ST320 strains. Non-vaccine strains 19A and 15B/C isolated in 2007 shared identical alleles with 81% (17 out of 21) and 76% (16 out of 21) of the virulence-associated loci of serotype 9V strains, respectively. None of these loci were observed in the predominant 19A (ST199) and 15B/C (ST199) strains. It should be noted that Table 6.2 does not represent a comprehensive list of loci involved in virulence/pathogenesis.

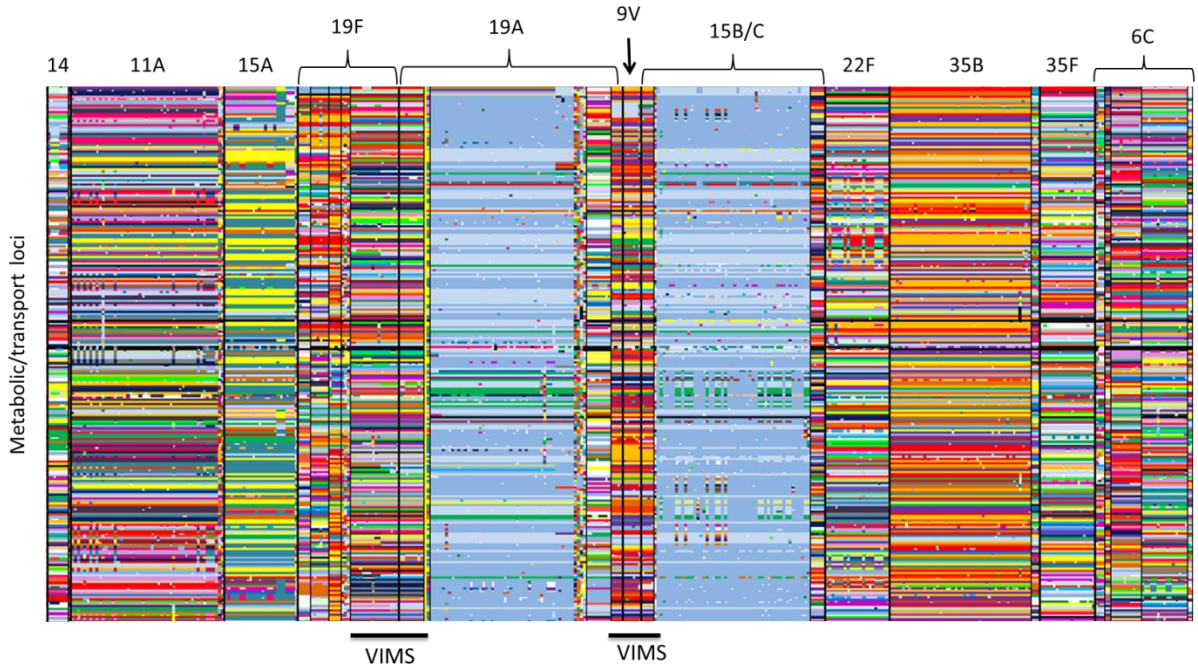


Figure 6.11. Partial screenshot of the metabolic profiles of 616 pneumococcal isolates (sequenced by Croucher et al. (2013)), grouped by serotype (see also Chapter 5). *Each column contains the allele numbers for a single isolate, of metabolic/transport genes (represented by rows). Allele numbers are represented by different colours. Black bars indicate where VIMS (Vaccine-Induced Metabolic Shift) has occurred.*

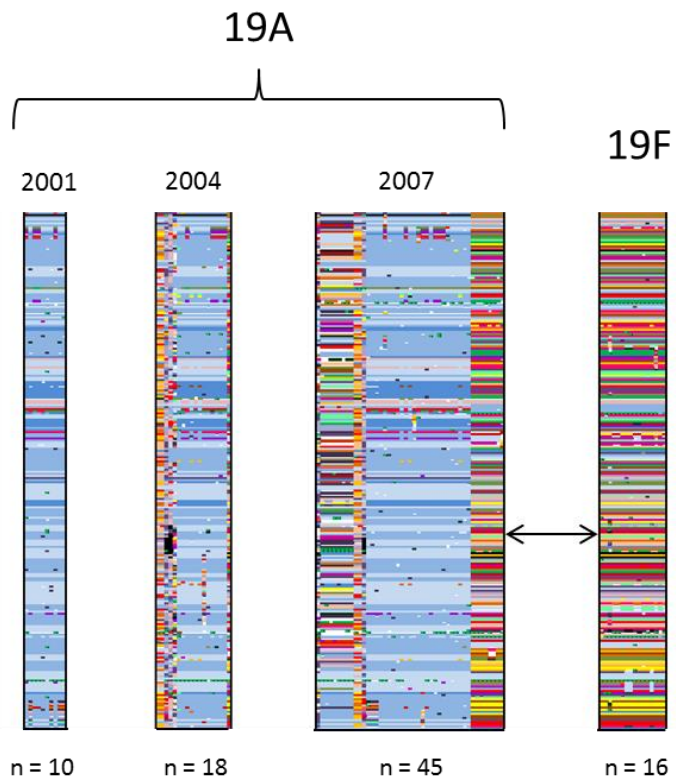


Figure 6.12. VIMS from 19F to 19A strains*. A snapshot of the metabolic alleles from all 19A strains collected in 2001, 2004 and 2007 (left), and all 19F CC320 strains (right)(sequenced by Croucher et al.,2013).*

Locus	Genome position	Vaccine strain	Non-vaccine strains		Vaccine strain	Non-vaccine strains		
		19F ST320	19A ST320	19A ST199	9V ST156	19A ST1925	15/AB ST162 or 3275	15A/B ST199
nanA | SPN23F12210	1189320	29 (0.9)	29 (100)	3 (0.91)	8 (100)	8 (100)	2 (100)	2 (0.81)
bgaA | SPN23F05830	571752	21 (100)	21 (100)	2 (0.91)	9 (100)	9 (100)	9 (100)	2 (0.65)
strH | SPN23F00730	69102	45 (0.9)	45 (100)	1 (0.8)	23 (0.6)	23 (100)	23 (0.75)	1 (0.83)
SPN23F02890	274732	19 (100)	81 (100)	1 (0.77)	1 (0.8)	1 (100)	42 (100)	2 (0.6)
fbpS | pavA | SPN23F08910	864927	17 (100)	17 (100)	2 (100)	4 (100)	4 (100)	4 (100)	2 (0.81)
eno | SPN23F10490	1012087	13 (100)	13 (100)	1 (100)	8 (100)	8 (100)	8 (100)	1 (100)
ply | SPN23F19470	1889503	12 (100)	12 (100)	2 (100)	8 (0.8)	8 (100)	8 (100)	2 (0.96)
lytA | SPN23F19600	1898038	1 (100)	1 (100)	3 (0.86)	9 (100)	3 (100)	9 (50)	3 (0.75)
nanB | SPN23F16870	1626581	37 (100)	37 (100)	3 (0.91)	10 (100)	10 (100)	10 (100)	3 (0.65)
SPN23F06940	681761	12 (100)	12 (100)	1 (100)	8 (100)	8 (100)	8 (100)	1 (0.92)
lytB | SPN23F08900	862890	59 (100)	59 (100)	1 (100)	11 (100)	11 (100)	11 (100)	1 (0.63)
cbpE | SPN23F08530	834686	30 (100)	30 (100)	1 (0.94)	9 (100)	9 (100)	9 (100)	1 (100)
phtE | SPN23F09300	898830	52 (100)	23 (100)	2 (0.94)	3 (100)	3 (100)	3 (100)	2 (0.94)
gapN | SPN23F10400	1002438	17 (100)	17 (100)	1 (0.94)	9 (100)	9 (100)	9 (75)	1 (0.98)
SPN23F04520	435720	23 (100)	23 (100)	26 (0.91)	12 (100)	47 (100)	52 (100)	2 (0.6)
ccpA | SPN23F20200	1959921	4 (100)	4 (100)	1 (100)	1 (100)	1 (100)	1 (100)	1 (0.98)
pcpA | SPN23F21690	2108466	23 (0.22)	15 (50)	28 (0.14)	T (0.8)	T (50)	62 (50)	3 (0.1)
cbpD | SPN23F22340	2181850	X (100)	29 (100)	1 (0.97)	8 (0.8)	8 (0.75)	8 (100)	1 (0.71)
htrA | SPN23F22720	2219291	4 (100)	4 (100)	2 (100)	8 (100)	8 (100)	8 (75)	2 (100)
cbpJ | SPN23F03490	340640	33 (0.67)	33(100)	7 (0.91)	1 (100)	7 (100)	20 (100)	2 (0.21)
cbpG | SPN23F03640	352443	39 (0.67)	39(100)	21 (0.97)	10 (100)	21 (100)	10 (100)	2 (0.94)

Table 6.2. Predominant alleles (%) of 21 virulence-associated loci present in selected vaccine and non-vaccine strains, according to the output from the Genome Comparator Tool. *T refers to alleles which have been truncated; X refers to alleles not present in the genomic assembly.*

6.6 Discussion

Here, we explored the effects of strain-targeted vaccination using multilocus models comprising antigenic, metabolic and subsequently, virulence-associated components. The results of the models comprising antigenic and metabolic loci showed that the MT originally associated with the vaccine AT was not eliminated by the vaccine, but instead emerged in association with a different AT. We refer to this result as Vaccine Induced Metabolic Shift (VIMS).

These findings are consistent with data recorded from previous pneumococcal studies. Capsular switching events have been documented to occur regularly throughout the past seven decades (Wyres *et al*. 2013), and since the introduction of the PCV7 vaccine, it has been noted that many non-vaccine serotypes have acquired MLST-defined sequence types (STs) that were previously associated with vaccine serotypes. For example, since the introduction of the PCV7 vaccine in North America, there has been a significant increase in the prevalence of the non-vaccine serotype 19A: initially this was due to clonal expansion of the pre-existing genotype ST199, but ST320 (previously associated with the vaccine serotype 19F) has now replaced ST199 as the most common genotype in 19A invasive disease and carriage in several US regions (Beall *et al*. 2011; Hanage *et al*. 2011; Sharma *et al*. 2013) (Figure 6.13) and in Canada (Pillai *et al*. 2009). Increases have also occurred in the prevalence of the ST695^{19A} strain in which the vaccine serotype 4 capsule has been switched for a 19A capsule (Brueggemann *et al*. 2007; Golubchik *et al*. 2012). Similarly, in Korea, where PCV7 caused a drop in vaccine serotypes 23F and 19F (but notably not 6B), ST81 (previously associated with serotypes 23F and 19F) is now the primary MLST type of serotype 6A (Baek *et al*. 2011). Similar findings were also yielded by Croucher *et al*. (2013), who noted examples of Sequence Clusters (lineages defined by clustering analysis of core COGs, or Clusters of Orthologous Genes)

which were originally linked to vaccine strains, but increased in prevalence following vaccination in association with non-vaccine serotypes. We subsequently found support for VIMS within the 616 genomes published by Croucher *et al*. (2013): the metabolic profile corresponding to ST320 was observed in association with vaccine serotype 19F and subsequently non-vaccine serotype 19A; and that of vaccine serotype 9V in 2004 was observed in association with non-vaccine serotypes 19A and 15B/C in 2004 and 2007 (Figures 6.11 & 6.12). Thus, the strains corresponding to serotype 19F CC320 (ST320, ST236, ST271) underwent capsular switches to 19F, and serotype 9V underwent capsular switches to 19A and 15B/C.

A number of theoretical frameworks of VISR assume strong competition across ATs in the form of cross-immunity, but the expansion of non-vaccine strains in the form of VIMS was observed here with no cross-immunity operating between strains. Further, VIMS was observed when vaccine-induced cross-immunity was equal to naturally-acquired cross-immunity (at low levels, where $0 < \gamma < 0.3$). The levels of naturally-acquired cross-immunity and cross-protection provided by PCV7 are not known, but these low levels are not inconsistent with the weak responses suggested from previous studies (Weinberger *et al*. 2008; Hausdorff *et al*. 2010).

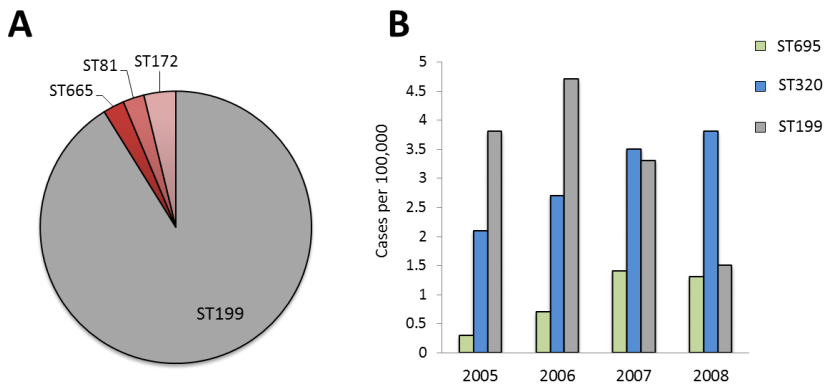


Figure 6.13. Capsular switching associated with serotype 19A in the United States. *(A) STs associated with serotype 19A in 1999 (n=82 isolates; adapted from Beall et al. (2006). (B) Disease prevalence among children <5 years, following vaccination in 2000, from Beall et al. (2011).*

The model frameworks presented here and in Chapter 3 can thus explain both how non-overlapping associations between serotype and ST may arise despite frequent recombination, and how vaccination can cause the STs of vaccine strains to become associated with non-vaccine serotypes through capsular switching. It is important to note that these results do not imply that vaccines play a mechanistic role in inducing capsular switching variants; on the contrary, capsular switch variants can be present at low frequencies prior to vaccination but will only expand after vaccination due to altered ecological and immunological host landscapes (Feikin & Klugman 2002; Ihekweazu *et al*. 2008; Jefferies *et al*. 2010; Croucher *et al*. 2011). Our model is thus able to resolve why - for example - although the ST695^{19A} vaccine escape variant in the US was first reported in 2003 (Pai *et al*. 2005; Beall *et al*. 2006), evolutionary analyses by Croucher *et al*. (2013) indicate that the capsule switch may have taken place prior to 1997. Overall, our model predicts that rare genotypes (which may or may not have been there prior to vaccination) may increase in frequency after vaccination due to the removal of ecological competition with vaccine serotypes possessing very similar metabolic profiles. For example, as noted by Croucher *et al*. (2013), there was an expansion within Sequence Cluster 15 of non-vaccine serotypes 23A and 23B following vaccination, as the previously dominant vaccine serotype 23F declined in prevalence. The analyses presented in Chapter 5, suggesting that the metabolic genes form a tightly linked and highly specialised metabolic profile, are consistent with the hypothesis that pneumococcal strains exploit distinct metabolic niches, and that ecological competition operates between strains with similar metabolic profiles. The result that VIMS was only observed when strain transmissibility is determined by the MT (as opposed to the AT) also reinforces the hypothesis that the relative fitness of pneumococcal strains is determined by a combination of the metabolic genes as well as the serotype (Figure 6.4).

VIMS may have consequences for virulence among non-vaccine strains if the factors determining MT are also associated with propensity to cause invasive disease (reviewed by

Shelburne *et al*. 2008; Schoen *et al*. 2014; see also Section 1.4 & 1.5). For example, MT1 may be intrinsically more virulent than MT2, thus justifying the choice of the associated serotype as a vaccine candidate. Under these circumstances, VIMS will always be accompanied by an increase in virulence among non-vaccine serotypes acquiring the MTs of vaccine serotypes. However, virulence factors may also be extrinsic to MT, and thereby require an additional locus to represent them within our framework.

We extended the model framework presented in Chapter 3 to explore the population structures which arise in an 18 strain multilocus model comprising antigenic, metabolic and virulence-associated loci. Figure 6.2 presents the most common population structures that arise within this extended model, which may putatively correspond to recorded data for other bacterial pathogens, listed in Table 6.1 (although see discussion on page 228) . The red population structure, in which non-overlapping associations occur between AT and MT but with only a subset of strains having a high propensity to cause invasive disease, most likely corresponds to recorded data for *S. pneumoniae* (Figure 6.2b).

An important feature of this model is that high virulence alleles tend to associate with the most transmissible MTs, consistent with the results of the deterministic model by Buckee *et al*. (2008). There is no evidence that this is the case in *S. pneumoniae* but a comparative whole-genome hybridisation study of *N. meningitidis* (Joseph *et al*. 2011) supports the hypothesis that only certain combinations of metabolic genes can "afford" virulence. Under these circumstances, vaccination against the most virulent strain will not only cause VIMS but may now also be accompanied by an increase in virulence, if the transmission benefit of the virulence alleles is offset by the cost in terms of a shortened infectious period (Figure 6.7a).

A possible example of this phenomenon is the large increase in piliated strains observed in the US since PCV7 vaccination. In Massachusetts, PI-1 was associated primarily with vaccine-

type serotypes before vaccination in 2000. PI-1 subsequently decreased in prevalence with the declining vaccine serotypes, but re-emerged in 2004-2007 in association with non-vaccine serotypes, in particular serotype 19A (Regev-Yochay *et al*. 2010). Similarly, there has been a 40% increase in PI-2 in serotype 19A following the introduction of PCV7 in Atlanta, Georgia (Zahner *et al*. 2010). In line with our model predictions, the increase in PI-1 and PI-2 in serotype 19A are linked to the acquisition of ST695 and ST320 in the post-vaccine era. It is notable that no significant association was found between the presence/absence of Type I pili and whether the isolate was obtained from carriage or invasive disease in sample of pneumococci from the United States (Basset *et al*. 2007). This finding contrasts with those of experimental studies which suggest that Type I pili contribute to increased virulence (Barocchi *et al*. 2006).

The non-vaccine strain is only observed in association with the virulence determinant if the cost of virulence is offset by a transmission benefit (Figure 7.7a). Vaccination was also simulated in the "blue" population structure, in which partially non-overlapping associations between ATs and MTs emerge. As associations between serotype and ST are not perfect in *S. pneumoniae* (and overlapping associations between AT and ST/ET are observed in other pathogens), we also simulated vaccination against the two ATs associated with virulence alleles in this population structure. We found that the non-vaccine AT was consistently observed in association with the virulence determinant after vaccination (Figure 7.7b).

The model may also explain why drug-resistance would increase among non-vaccine serotypes following vaccination, since the "superior" MTs associated with vaccine serotypes are also more likely to be able to bear the fitness costs associated with antimicrobial resistance. Indeed, in North America, the majority of penicillin-resistant 19A isolates are linked with STs (such as ST320) previously associated with vaccine serotypes (Beall *et al*. 2011). In Italy, the highly prevalent antibiotic resistant ST230 clone, previously associated with vaccine serotypes 14 and

19F, is now predominantly observed with non-vaccine serotypes 19A and 24F (Gherardi *et al*. 2012). The acquisition of antibiotic resistance may increase the rapidity with which VIMS occurs, but it is not a necessary condition as shown by the example of $ST695^{19A}$ which is taking over from $ST199^{19A}$ in spite of an identical resistance profile (Beall *et al*. 2011). We believe it is likely that both antibiotic resistance in addition to vaccine-induced strain replacement have played important roles in the emergence of non-vaccine serotypes in the post-vaccine era.

The model framework may also offer a putative explanation for the inconsistent observation of VISR across bacterial pathogens. For example, although the meningococcal conjugate C vaccine was introduced in the UK over 15 years ago - indeed before the pneumococcal conjugate vaccine - we have not yet observed vaccine replacement among meningococcal populations by other serogroups (and notably, serogroup B, which was highly prevalent at the time of vaccination) (Maiden *et al*. 2008). The polysaccharide of serogroup B meningococci is not immunogenic, and thus does not represent a "true" antigenic type in the model framework. VISR or indeed VIMS is therefore not possible with serogroup B meningococci. However, a number of meningococcal vaccines are based on several immunogenic outer membrane antigenic variants distributed across a wide variety of strains (e.g. fHBP, NadA, NHBA, PorA, FetA). The results of our model suggest that VIMS may be possible following the use of these vaccines.

It is noteworthy that this mode of increase in virulence is mechanistically different from that proposed by Gandon *et al*. (2001) where the costs of virulence are directly mitigated by the vaccine (by preventing disease but not transmission); in our model it is the intrinsic transmissibility of non-vaccine antigenic types that undergoes an increase, thereby supporting the costs of increased virulence.

The results of the model suggest that VIMS will only occur if the vaccine AT is completely eradicated (Figure 7.5). Such a result is consistent with the anticapsular immunity provided by the vaccine which is strongly protective against carriage, but vaccine efficacy is likely to be much less than 1 (Rinta-Kokko *et al*. 2009). It also assumes comprehensive vaccine coverage across populations. Although vaccine coverage is high in many countries, it is heterogeneous in others, such as Spain. The effects of vaccination will also depend on vaccine schedules and rates of introduction, as well as levels of vaccine-induced and vaccine acquired cross-immunity (Cobey & Lipsitch 2012). However, our aim here is not to reproduce the precise dynamics of replacement in real populations, or to predict the precise conditions under which VISR may occur, but to provide an insight into the effects of the selective pressures conferred by strain-targeted vaccines on the population structure of bacterial pathogens.

The framework here predicts that the associations between AT and MT are arbitrary; it would be interesting to investigate the effects of vaccination where epistasis operates in such a system, for example by attributing fitness costs and benefits to particular combinations of AT, MT and virulence type. Multilocus models allow one to attribute particular characteristics to particular loci. Correlations between pneumococcal serotype and virulence are well-recognised, and a number of sub-capsular antigens are also recognised as playing important roles in virulence, such as PspC and PspA. Further, a number of metabolic genes are also recognised as playing important roles in virulence, and some structures may play roles in all three functional categories, such as NanA. Thus, it would thus be interesting to explore model frameworks in which virulence is a function of the AT and/or MT.

6.7 Summary

Several theoretical models have demonstrated the potential of strain-targeted vaccines to increase the prevalence of non-vaccine serotypes due to the removal of cross-immunity or direct resource competition, but few have considered the effects of vaccination on non-vaccine strains. Further, these frameworks do not explain why changes have occurred in the MLST composition of non-vaccine serotypes of *S. pneumoniae*. Here, we use models incorporating both metabolic and antigenic loci to show that vaccination against particular antigenic types can cause their metabolic components to be observed in association with non-vaccine antigenic types: a phenomenon we term Vaccine-Induced Metabolic Shift. The theoretical frameworks comprising metabolic and antigenic loci were subsequently expanded to include virulence-associated loci, and resulting population structures explored. Simulating strain-targeted vaccination in this framework suggested that, in addition to the metabolic alleles, the virulence-associated alleles of vaccine serotypes may become associated with non-vaccine serotypes following vaccination. Our results provide a novel explanation for changes observed in the population structure of the pneumococcus following vaccination, and our results are supported by observations at the whole genome level.

Chapter 7.

Genetic Exchange of Metabolic Loci among Pandemic Meningococcal Lineages

7.1 Public Dissemination of Work

This work was presented as part of a poster at a number of conferences:

"Interdisciplinary Approaches to Understanding Microbial Communities," Isaac Newton Institute for Mathematical Sciences, Cambridge, UK (October 2014).

International Pathogenic *Neisseria* Conference, North Carolina, US (September 2014). (Poster prize awarded).

European Molecular Biology Organisation "Genomes 2014: Microbiology after the Genomics Revolution", Pasteur Institute, Paris (June 2014).

7.2 Abstract

Serogroup A meningococci from the ST-5 complex have given rise to three pandemics over the past 50 years, and more recently, the ST-2859 clone from this lineage has emerged as a major cause of meningococcal disease in sub-Saharan Africa. The reasons for the emergence of the three pandemics and ST-2859 remain unclear, especially as meningococci from this lineage are highly conserved antigenically. In this chapter, we aim to understand the allelic differences among these meningococci at a whole genome level, and we analyse the diversification of 153 isolates from the three pandemics and ST-2859. As originally proposed by Buckee *et al*. (2008), and according to the model frameworks presented in this thesis, differences in the metabolic genes may translate into differences in transmissibility and virulence. We therefore hypothesise that differences in the metabolic genes may provide a mechanism by which each successive strain out-competes its predecessor and emerges as a

new epidemic clone. Our results showed that the majority of the allelic changes among strains from the three pandemics and ST-2859 indeed pertained to metabolic functions. We also found evidence of eleven introgression events in the emergence of two pandemic waves and ST-2859, all of which originated in other hyperinvasive clonal complexes, suggesting that multiple recombination events among different meningococci from several virulent lineages has led to the emergence of novel epidemic strains.

7.3 Introduction

Since 1905, the largest known epidemics of meningococcal disease have occurred in sub-Saharan Africa, in an area described as the 'Meningitis Belt'. These epidemics occur approximately every 10 years, and have mostly been caused by serogroup A meningococci (Greenwood 1999). In addition to the epidemics across Africa, Serogroup A meningococci were also responsible for a number of meningococcal pandemics throughout the 20th century (Olyhoek *et al*. 1987), caused by three major clonal complexes (ST-1, 4 and 5 complexes; Figure 7.1).

Meningococci belonging to the ST-1 complex (also known as Subgroup I) were responsible for epidemics in Niger, Nigeria, Chad, North Africa and the Mediterranean in the 1960s, and subsequently in Nigeria and Rwanda as well as among native Americans and homeless people in the US and Canada in the 1970s. During the 1970s, ST-1 complex meningococci were also responsible for endemic disease globally. Outbreaks were recorded in the 1980s among indigenous peoples from New Zealand and Australia. In contrast, ST-4 complex meningococci have caused epidemics almost solely in West Africa over a period of 40 years or so. During the early 1960s to 1980s, almost all meningococci disease in West Africa could be attributed to the ST-4 complex (Olyhoek *et al*. 1987).

The ST-1 and ST-4 complexes were gradually replaced by the ST-5 complex in the 1980s (also known as Subgroup III), which gave rise to three pandemic waves (Olyhoek *et al*. 1987; Morelli *et al*. 1997) (Figure 7.2). The first pandemic of the ST-5 complex – caused by meningococci belonging to ST-5 – originated in China during the mid-1960s, and subsequently spread to cause outbreaks in Moscow, Norway, Finland and Brazil. Meningococci of the second pandemic wave – also belonging to ST-5 – were first recorded in China in the early 1980s. An epidemic in Mecca during the Hajj pilgrimage in 1987 resulted in large numbers of pilgrims carrying these meningococci back to their countries of origin, including the USA, UK, Israel, France and East Africa. The ST-5 complex epidemics in most of these countries declined within a few years, except in East Africa, from where the epidemics spread throughout Central and West Africa. The majority of cases of meningococcal disease in Africa since that time have been caused by the ST-5 complex. The Ps-ATT conjugate vaccine (MenAfriVac™) was developed and introduced recently in response to these epidemics (Frasch *et al*. 2012). Meningococci of the ST-5 complex emerged at a global scale for the third time as a different sequence type, ST-7, and spread from China in the early 1990s to Mongolia, Russia, Europe and the US, as well as to many countries in the meningitis belt. These meningococci of the third pandemic wave gave rise to large and severe epidemics across sub-Saharan Africa, and indeed, the largest recorded meningococcal epidemic to date in 1998-1997, with more than 250,000 cases (Roberts 2008).

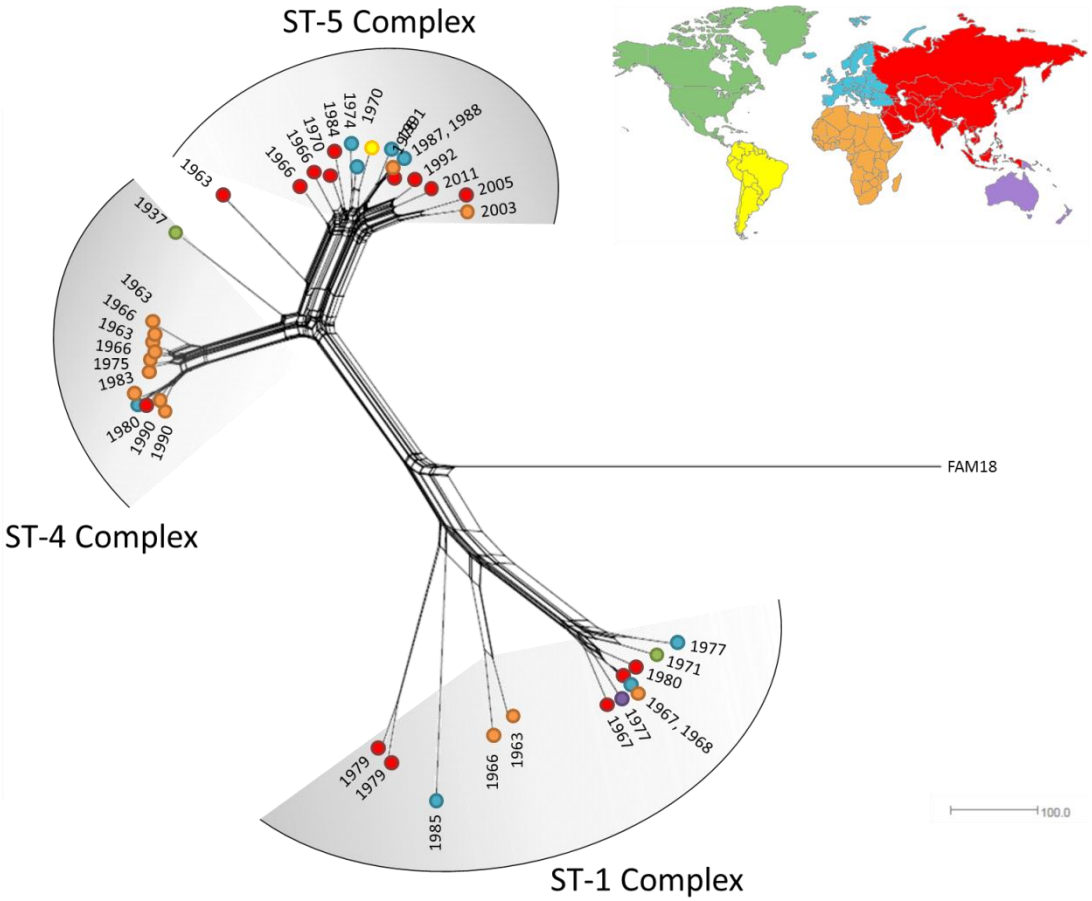


Figure 7.1. Allele-based phylogenetic network of 41 published serogroup A whole genomes from the ST-1, 4 and 5 complexes. *The date of isolation is shown for each strain, and the colour of the tips refers to the continents in which they were isolated. The network was constructed using the NeighbourNet algorithm, based on the alleles of 1993 coding sequences.*

A new sequence type of the ST-5 complex, ST-2859, was identified as a major cause of disease in several countries across the meningitis belt in the mid-2000s. First isolated in Burkina Faso in 2003, the hyperinvasive strain ST-2859 has given rise to several epidemics across the belt, and currently represents the predominant cause of disease in a number of African countries, including Burkina Faso, Mali and Togo (Sié *et al*. 2008; Caugant *et al*. 2012).

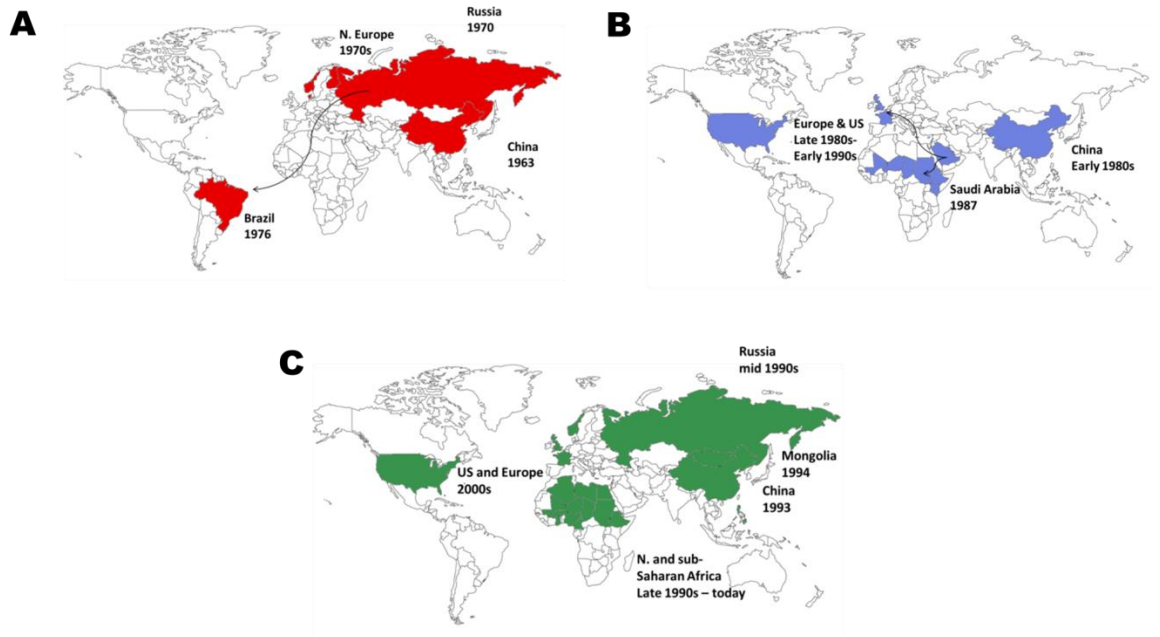


Figure 7.2. Global spread of the ST-5 complex meningococci belonging to the first (A), second (B) and third (C) pandemic waves.

The rise and fall of specific meningococcal clones across the meningitis belt has been referred to as clonal waves. Such waves are usually characterised by an increase in carriage of a particular epidemic clone as well as an increase in invasive disease. Longitudinal studies in Ghana have provided insight into the dynamics and diversity of a series of clonal waves (Leimkugel *et al*. 2007). Between 1998 and 2005, three major waves were observed: meningococci belonging to ST-5 were followed by serogroup X ST-751, and succeeded by serogroup A ST-7 strains. These studies demonstrated the low diversity of circulating strains within carriage relative to industrialised countries, as carriage strains within a wave were dominated by a single epidemic clone. Subsequent studies in Burkina Faso and Ghana showed a further wave of ST-2859 following that of ST-7 (Sié *et al*. 2008; Lamelas *et al*. 2014).

The reasons for the emergence of clonal waves across Africa, as well as the pandemic waves, are not well understood. Typing methods have shown meningococci from the ST-5 complex to

be highly uniform at their housekeeping alleles, as isolates from the first and second pandemics belong to the same sequence type (ST-5), and isolates from the third pandemic and ST-2859 differ only at the *pgm* MLST locus (Nicolas *et al*. 2005).

The "genocloud" hypothesis postulates that meningococci constantly evolve over time, through microevolution, and the resulting genoclouds/clonal complexes comprise a dominating, highly fit genotype as well as its closely related variants (Morelli *et al*. 1997; Zhu *et al*. 2001). Eventually, a new variant will arise with heightened fitness, and will thus out-compete its parents' strains within the genocloud. If such a fit variant also has distinct antigenic characteristics, thus allowing for antigenic escape, this allows the spread of a new wave to occur. Thus, successive waves are predicted to comprise genoclouds of a number of variants, each dominated by a particularly fit, central genotype.

Differences in a number of genes pertaining to immune escape have indeed been observed between isolates of the second and third pandemic waves, including transferrin-binding protein B (tbpB), IgA1 protease, OpaB, OpaD and the *lgt* gene (involved in lipooligopolysaccharide synthesis) (Olyhoek *et al*. 1987; Achtman *et al*. 1992; Zhu *et al*. 2001) . Immune escape at the sub-capsular antigens has also been suggested as a reason for the emergence of new clonal waves by Lamelas *et al*. (2014), who noted diversification at genes involving pilus glycosylation and retraction, as well as the *maf* adhesin locus, between ST-7 and ST-2859 strains.

However, there are a number of factors which suggest that, although potentially important, immune escape at the sub-capsular antigens may not be the primary factor driving the emergence of these epidemics. Firstly, if it were to be assumed that the sub-capsular antigens are able to induce long-lasting and protective responses, immune escape would be hypothetically possible if meningococci changed their variants at a large number of antigens.

However, previous studies have shown that meningococci are highly stable at several of the sub-capsular antigens across the pandemic waves, with identical variants of PorA VR1 and VR2, PorB, and the variable and immune-exposed region of the FetA gene (Norheim *et al*. 2006; Huber *et al*. 2012).

Secondly, we understand relatively little about the development and duration of naturally-acquired immunity from carriage among meningococci. Although it has been shown that polysaccharide conjugate vaccines are able to protect against carriage and induce strong herd immunity (Maiden *et al*. 2008), the immune response induced by natural carriage is less well understood. A number of studies have shown that carriage is able to induce anti-capsular immune responses against some – but not all – serogroups, but it is not known whether natural immunisation leads to immunological memory, and indeed herd immunity (e.g. Blakebrough *et al*. 1982; Amir *et al*. 2005; Mueller *et al*. 2006). Further, although a number of studies have shown antibody responses directed against the non-capsular antigens (Amir *et al*. 2005), the extent to which these are protective against carriage is not well understood; indeed, vaccination against the sub-capsular antigens does not always protect against carriage (Read *et al*. 2014). Thus, it is not known whether long-lasting herd immunity is generated against the sub-capsular antigens, or indeed even the capsule.

If it is the case, then, that herd immunity generated by natural-acquired immune responses against the sub-capsular antigens may not be the primary reason for the emergence of the clonal waves, there must be other reasons for the repeated re-emergence of these antigenically and genetically stable clones. If sub-populations of competing meningococci were not able to gain a competitive transmission advantage through changing their dominant antigens, perhaps they were able to out-compete their predecessors through increasing their inherent transmissibilities. The results of the models presented in this thesis suggest that if the transmission fitness of a given strain exceeds that of a competing strain just slightly, it is able to replace the other

circulating strains. As originally proposed by Buckee *et al*. (2008), it is possible that small differences in transmission fitness may be dictated by differences in metabolic efficiency between strains, resulting in a differential ability to grow within and transmit between hosts.

Here, we analyse 153 whole genomes of the ST-5 complex to investigate the diversification of this lineage over time and the possible genetic factors driving the three separate pandemics and the evolution of ST-2859 in Africa. We investigate the diversity of additional sub-capsular antigens not previously surveyed by typing studies; however, in line with the results of previous models, we hypothesise that differences in the metabolic genes may be observed between meningococci of the three pandemic waves and ST-2859, and contribute to their spread and emergence.

7.4 Methods

The sequence reads for all ST-5 complex meningococcal genomes available as of 17/10/13 were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/). The genomes were sequenced at the Wellcome Trust Sanger Institute in Cambridge, UK and the Institute for Genomic Sciences in Maryland, US (the Accession Numbers are listed in the Appendix). Reads were assembled using an automated pipeline based on the Velv*et al*gorithm, version 1.2.01. Annotation was carried out using the "autotagger" feature of the Bacterial Isolates Genome Sequence Database (BIGSdb) software (Jolley & Maiden 2010), which scans deposited sequences against defined loci in an automated BLAST process. The whole genome sequence data were compared using the BIGSdb Genome Comparator tool, which is implemented on the PubMLST website (www.pubmlst.org). The coding sequences within the annotation were extracted and compared against the reference strain Z2491 (accession number AL157959) using default parameters. Through Genome Comparator, unique allele sequences

were labelled consecutively, allowing the identification of shared and unique alleles between isolates. Loci with alleles specific to each pandemic wave were identified, and functionally characterised according to the KEGG Orthology (KO) groupings of the KEGG database (www.kegg.jp). Genes with uncharacterised functions which did not fall into a KO category were blasted in the PFAM database (www.pfam.sanger.ac.uk), and functionally characterised accordingly. Genes without significant hits in the PFAM search were designated with an "unknown" function. Genome Comparator was run, in addition, with the ST-5 strain WUE 2594 (accession number FR774048) as a reference, to identify alleles specific to the ST-2859 strains.

Allele sequences were blasted against the PubMLST database to find the appropriate NEIS number and allele, which were then scanned in the database for sequence matches in other meningococci. The PubMLST database represents a large repository of whole genome data, with approximately 5000 whole genome sequences of *Neisseria meningitidis*, from a variety of clonal complexes spanning a 78 year period (4998 whole genome sequences when searched on 22/07/15). The Neighbour Net networks in Figures 7.1 and 7.3 were created using SplitsTree v4. Annotated plots of the genome (Figure 7.5b) were created using the programme "CG view." The Artemis Comparison Tool was used to create Figure 7.5c (Carver et al. 2005).

7.5 Results

We refer to the three pandemics and the recent outbreaks caused by ST-2859 as four separate "epidemic waves". There was allelic variation at 999 out of 1993 loci (50.1%) among the 153 isolates (excluding incomplete alleles). Alleles within a given epidemic wave were highly uniform, however, with an average of 67.2% of the 999 variable loci exhibiting identical alleles within each wave, and 87.2% exhibiting alleles conserved across >90% of alleles within a given

wave (excluding incomplete alleles). A number of differences were found between the epidemic waves, with alleles at 39, 69 and 95 loci specific to isolates from the first, second and third pandemic waves respectively (Tables A8-A10). There were also 73 alleles specific to ST-2859 strains, and 14 deletions (Figure 7.3b; Tables A11-A12).

Upon functional characterisation of the allelic differences between the epidemic waves, however, we found that many major antigens were highly conserved (Table A13; Figure 7.4), and only a small minority of allelic changes across the epidemic waves pertained to the antigenic genes (3.7%). The majority of allelic differences among pandemics were characterised as metabolic (40.0%), followed by genetic information processing (16.2%), environmental information processing (12.3%), other functions (6.6%), cellular processes (2.2%), and 18.3% had no characterised function (Figure 7.5a). These loci were dispersed around the chromosome, and when plotted consecutively against their position in the genome, large groups of allelic changes at contiguous loci could be observed (Figure 7.5b), suggesting possible recombination of large genomic areas.

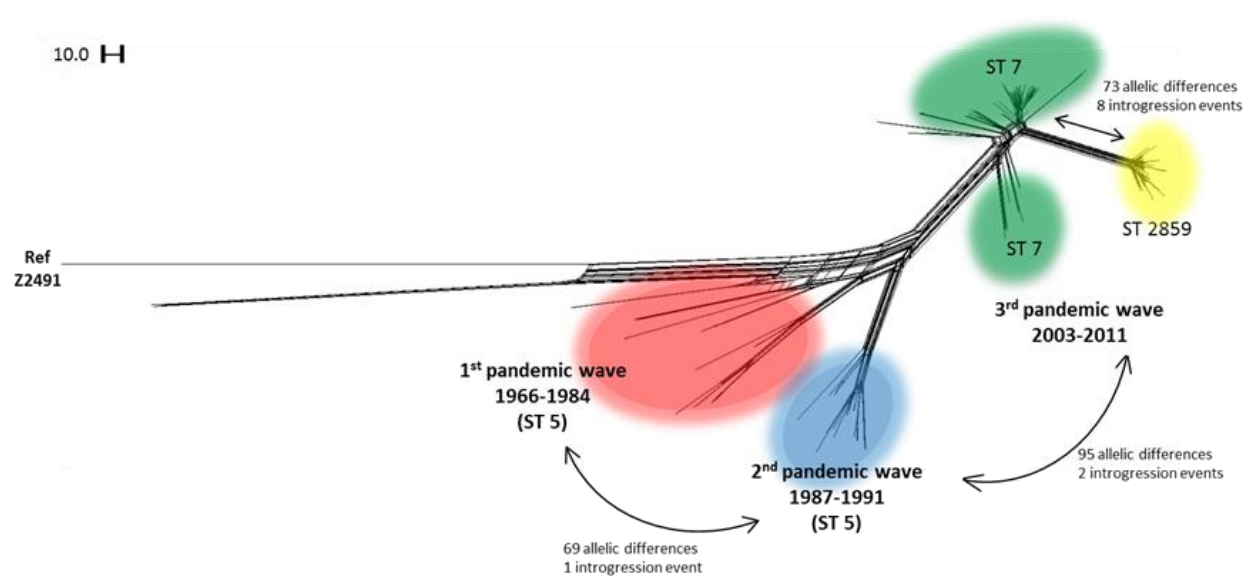


Figure 7.3. Allele-based wgMLST NeighbourNet network of 153 whole genomes of the ST-5 complex.

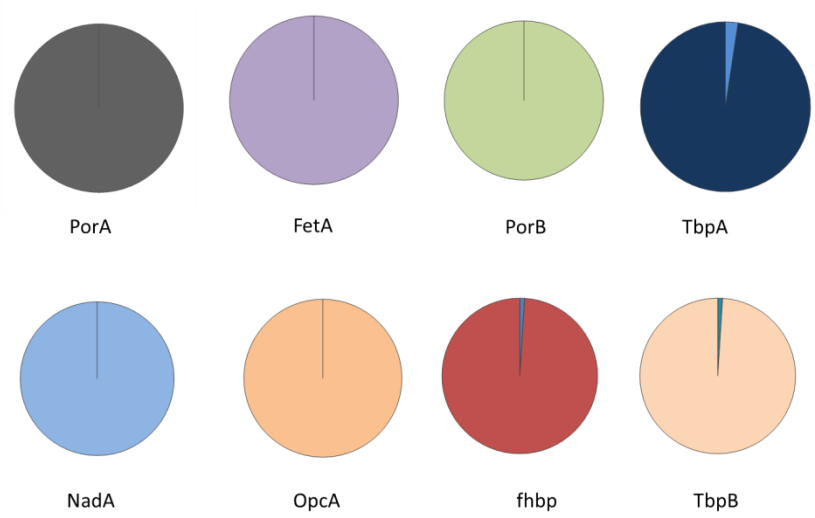


Figure 7.4. Allelic diversity of eight major sub-capsular antigens across 153 isolates of meningococci belonging to the ST-5 complex.

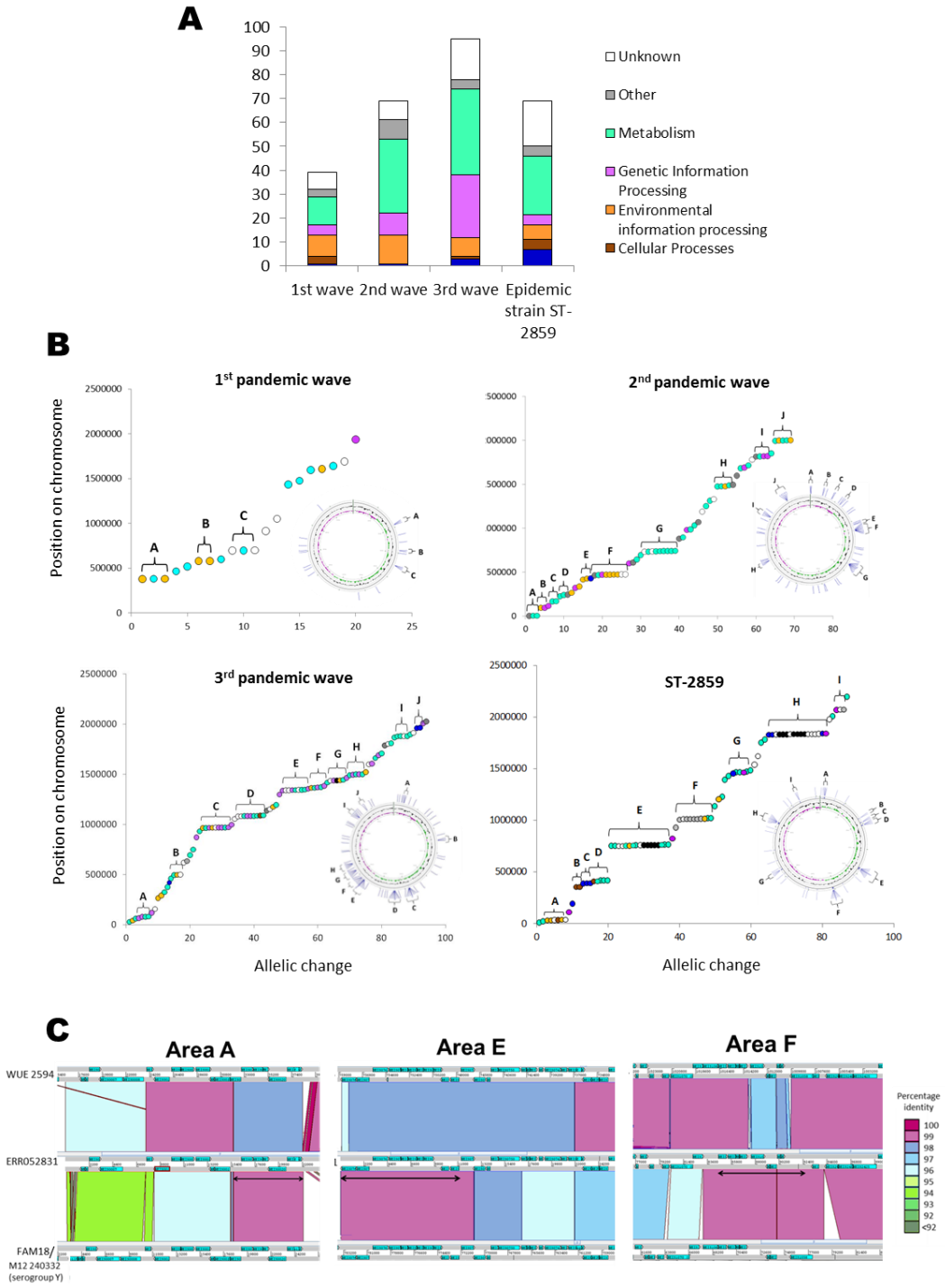


Figure 7.5. Functional characterisation and location of alleles specific to each epidemic wave. *(A) Functional characterisations of alleles specific to each epidemic wave. (B) Plots of successive allelic changes against their position in the reference genome for each epidemic wave, with an accompanying plot of these changes annotated on the circular chromosome of Z2491 (or WUE2594 for ST-2859). Letters indicate areas of allelic changes which are adjacent on the chromosome. (C) Genomic areas of putative recombination between ST-2859 and ST11/ST167 strains. Comparison plots show hypothetical donor strains on the bottom level (either reference strain FAM18, an isolate of serogroup C ST11, or M12 240332, a serogroup Y ST-167 complex strain), ST-2859 on the central level (isolate ERR052831), and reference*

strain WUE 2594 (serogroup A, ST-5 complex) on the top level (representative of the recipient strain). The arrows signify putative areas of recombination, and correspond to higher sequence identity shared between FAM18/ M12 240332 and ST-2859, than the ancestral WUE 2594 strain and ST2859. Diagrams are modified outputs of the Artemis Comparison Tool.

A number of the contiguous groups of alleles within the isolates from the epidemic waves had exact sequence matches to other hyperinvasive meningococcal clonal complexes (Table 7.1). Two contiguous genomic areas observed among isolates from the second pandemic wave had exact matches to large numbers of globally distributed isolates belonging to the highly virulent ST-11 complex, which has given rise to large numbers of epidemics in both the developing and developed world (Krizova & Musilek 1995; Whalen *et al*. 1995; Koumare *et al*. 2007). Alleles from area H of the third pandemic wave also had exact matches to large numbers of ST-11 complex meningococci. The earliest matching ST-11 isolates date from 1964 which is several years in advance of the onset of the second pandemic wave, suggesting that these large contiguous areas may be imports from an ST-11 strain. It is noteworthy that many of the introgressed alleles were present in the majority of ST-11 complex genomes present on the PubMLST database (>800). Further, eight areas of contiguous allelic changes among ST-2859 strains contained exact matches to isolates from hypervirulent clonal complexes. The alleles from seven such areas were found in a variety of globally-distributed ST-11 complex strains, but were universally present in a small group of serogroup W ST-11 strains circulating in Burkina Faso and Niger in 2001 and 2002, thus representing potential donors for the emergence of the ST-2859 clone. Area F also matched isolates from several serogroup Y isolates of the hyperinvasive ST167 complex, the earliest allele of which was identified in an isolate dating from 1940. It is possible, therefore, that multiple introgression events with several hyperinvasive meningococci gave rise to the epidemic ST-2859 clone. Alleles within the majority of introgressed areas were found in relatively early reported isolates in the PubMLST

database (Table 7.1) – with the earliest dating from 1937 - thus suggesting that these potentially virulent loci are long-lived and have been circulating for several decades.

The majority of introgressed genes with characterised functions pertained to metabolism (62.5%, 25 out of 40 loci; 50% of all 50 loci including those with unknown functions) (Table 7.1; Figure A5), and 86.3% of alleles translated into amino acid changes (Table A14). Five genes within area G among isolates from the second pandemic wave encode proteins which form the majority of the ubiquinone complex, and are approximately predicted to interact within the same functional metabolic network (along with gene NMA0745 in area G) (Figure 7.6). This suggests that the import of metabolic genes may have contributed to the emergence of the epidemic strains, and possibly that the genetic re-assortment of functional metabolic units among hyperinvasive lineages may also confer virulence or transmission-related traits.

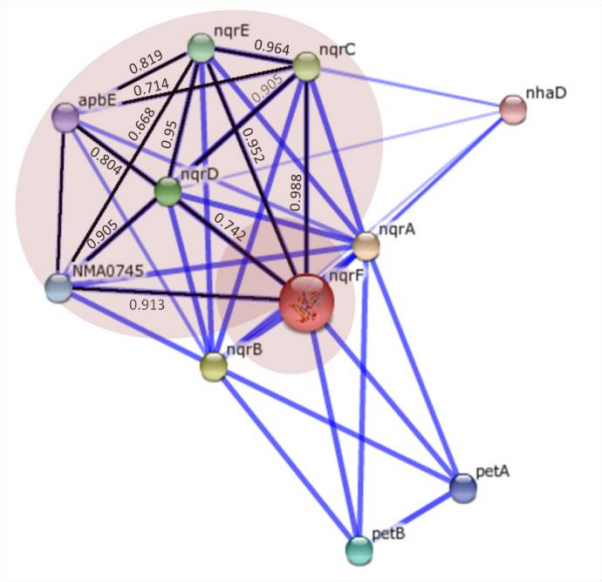


Figure 7.6. Predicted interactions between proteins of the ubiquinone complex (connected by black lines), encoded within introgressed area G among isolates from the second pandemic wave (shaded area). *The numbers indicate the predicted confidence of functional interaction between two protein partners (ranging from 0-1) according to the STRING database. The images were modified from the STRING database (string-db.org).*

	Genome position	Locus in reference genome	Matches on PubMLST database (no. isolates)	Nucleotide differences to match	Clonal complexes of isolates with allelic matches in PubMLST database (no. isolates)	Oldest isolate with allele on PubMLST database	Product	Functional Characterisation
Isolates belonging to ST-2859								
A	32717	NMAA_0029	513	Exact match	11 (457), 162 (38), 269 (13), 865 (8)	USA, 1964	Thiamine transport system substrate-binding protein	Environmental information processing
A	33812	NMAA_0030	521	Exact match	11 (430), 162 (42), 269 (12), 865 (8), 41/44 (5)	USA, 1964	Mechanosensitive ion channel	Environmental information processing
A	34686	NMAA_0031	926	Exact match	11 (681), 41/44 (43), 162 (41), 269 (27), 35 (18), 167 (15), 60 (9), 461 (9), 282 (8), 865 (8), 213 (6)	USA, 1964	Competence-damaged protein (CinA family)	Unknown
A	35273	NMAA_0032	438	Exact match	11 (418)	USA, 1964	Peptide methionine sulfoxide reductase (putative pilin biogenesis)	Cellular processes
A	36986	NMAA_0033	604	Exact match	11 (544), 269 (32), 865 (6)	USA, 1964	Probable signal recognition particle protein	Environmental information processing
C	389418	NMAA_0322	15	Exact match	11 (15)	UK, 1998	Pilin glycosylation protein pglC	Antigenic
D	409695	NMAA_0337	1046	Exact match	11 (739), 22 (88), 41/44 (56), 167 (20), 8 (19), 18 (16)	The Netherlands, 1963	GTP binding protein engB	Cellular processes
D	410530	NMAA_0338	2585	Exact match	11 (746), 41/44 (521), 269 (374), 23 (262), 32 (135), 22 (88), 162 (47), 35 (41) , 174 (36), 167 (31), 8 (21), 18 (18), 198 (12), 282 (11), 60 (10), 213 (5)	Denmark, 1940	Cytochrome C	Metabolism
D	413375	NMAA_0340	839	Exact match	11 (794), 8 (19)	USA, 1964	Cytochrome C biogenesis protein (CcsA)	Metabolism
D	414679	NMAA_0341	1	3 nucleotide differences	11 (1)	UK, 2013	tRNA N6-adenosine threonylcarba	Metabolism

							moyltransferase.(EC 2.3.1.-	
E	749433	NMAA_0635	153	5 nucleotide differences	41/44 (69), 5 (21), 11 (14), 4 (13)	Burkina Faso, 1963	Ribosomal large subunit pseudouridine synthase F (rluF)	Metabolism
E	750216	NMAA_0636	1068	Exact match	11 (802), 213 (157), 41/44 (20), 254 (10)	Denmark, 1940	ATP-NAD kinase	Metabolism
E	751126	NMAA_0637	1159	Exact match	11 (794), 23 (258), 162 (47), 41/44 (5)	USA, 1964	None	Unknown
E	751743	NMAA_0638	863	Exact match	11 (796), 162 (45)	USA, 1964	NADH(P)-binding	Unknown
E	753236	NMAA_0640	1385	Exact match	11 (812), 41/44 (407), 60 (54), 4 (14), 865 (11), 254 (9), 213 (9)	USA, 1937	UDP-N-acetylenolpyruvoylglucosamine reductase (EC 1.3.1.98)	Metabolism
E	754454	NMAA_0641	842	Exact match	11 (795), 1 (11)	Niger, 1963	Multidrug efflux protein	Environmental information processing
E	756168	NMAA_0642	844	Exact match	11 (602), 41/44 (92)	The Netherlands, 1963	ATP phosphoribosyltransferase regulatory subunit	Metabolism
E	757422	NMAA_0643	362	6 nucleotide differences	11 (820)	USA, 1964	Adenylosuccinate synthetase	Metabolism
E	762625	NMAA_0651	984	Exact match	11 (838), 1 (47), 8 (20), 4 (14), 35 (7), 4821 (5)	USA, 1937	Adenylate kinase	Metabolism
E	764593	NMAA_0653	605	Exact match	11 (586)	USA, 1964	pfkB family carbohydrate kinase (BIGS: D-beta-D-heptose-7-phosphate kinase)	Metabolism
E	765601	NMAA_0654	1343	Exact match	11 (832), 41/44 (434), 8 (20), 22 (5), 269 (5)	USA, 1964	Cytosine-specific methyltransferase	Metabolism
F	1007824	NMAA_0886	10	6 nucleotide differences	167 (7)	UK, 2010	Putative phage tail fiber protein	Other
F	1010855	NMAA_0888	1398	Exact match	41/44 (446), 269 (388), 213 (153), 23 (139), 1, (41), 461 (37), 174 (36), 167 (30), 4(13)	Burkina Faso, Niger, 1963	Protein of unknown function (DUF497)	Unknown

F	1011222	NMAA_0889	42	Exact match	167 (30). 41/44 (6)	The Netherlands, 1986	Caudovirales tail fibre assembly protein	Other
F	1011806	NMAA_0890	41	Exact match	167 (27), 41/44 (7)	The Netherlands, 1986	None	Unknown
F	1013947	NMAA_0892	1523	Exact match	11 (577), 41/44 (443), 213 (148), 23 (137), 461 (34), 167 (23), 103 (19), 8 (14), 865 (12), 269 (10), 1 (8), 32 (7)	UK, 1941	Putative bacterial lipoprotein (DUF799)	Unknown
F	1014591	NMAA_0893	1091	3 differences	11 (569), 23, (137), 22 (82), 41/44(43), 1 (42), 53 (35), 35 (30), 167 (30), 8(8)	Denmark, 1940	None	Unknown
F	1014969	NMAA_0894	146	Exact match	53 (32), 167 (31), 198 (12), 35 (8), 11 (7), 32 (6), 41/44 (5), 269 (5), 1136 (5)	Denmark, 1962	Curli production assembly/transport component CsgG	Environmental information processing
F	1015805	NMAA_0895	203	Exact match	32 (154), 167 (22)	Denmark, 1962	Short chain dehydrogenase	Metabolism
G	1459282	NMAA_1235	796	Exact match	11 (764), 8 (6)	USA, 1964	CTP synthase	Metabolism
G	1461028	NMAA_1236	793	Exact match	11 (768)	USA, 1964	Long-chain-fatty-acid--CoA-ligase	Metabolism
G	1462769	NMAA_1237	691	Exact match	11 (644), 8 (7)	USA, 1964	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	Genetic information processing
G	1464586	NMAA_1239	2508	Exact match	11 (811), 41/44 (508), 23 (231), 269 (347), 213 (157), 60 (59), 461 (37), 35 (37), 174 (34), 167 (30), 8 (19), 198 (12), 32 (9), 1157 (7), 1136 (5)	Denmark, 1962	Diacylglycerol kinase	Metabolism
H	1829030	NMAA_1551	29	Exact match	11 (28)	Canada, 1970	Adhesin MafB2	Antigenic
H	1832666	NMAA_1557	830	Exact match	11 (807)	Norway, 1969	mafB-CTo1MGI-1 alternative toxic C-	Unknown

							terminal extremity	
H	1837213	NMAA_1566	622	Exact match	11 (601)	USA, 1964	None	Unknown
I	2069801	NMAA_1788	603	Exact match	11 (577), 865 (7)	USA, 1964	Ribosomal RNA small subunit methyltransferase G	Genetic information processing
I	2071324	NMAA_1790	543	Exact match	11 (509), 41/44 (7)	USA, The Netherlands, 1964	Fusaric acid resistance protein family	Metabolism
Isolates belonging to second pandemic wave								
G	734253	NMA0741	1288	Exact match	11 (821), 41/44 (313), 22 (85)	USA, 1964	Ubiquinone biosynthesis protein UbiB	Metabolism
G	737825	NMA0745	1053	Exact match	11 (837), 22 (88), 103 (17), 41/44 (10), 5 (5), 4821 (5)	USA, 1964	Putative periplasmic protein	Unknown
G	738066	NMA0746	505	Exact match	11 (346), 22 (84), 32 (5), 4821 (5), 5 (5)	USA, 1964	Thiamine biosynthesis protein	Metabolism
G	739274	NMA0747	959	Exact match	11 (800), 22 (86), 41/44 (5), 32 (5), 4821 (5), 5 (5)	USA, 1964	Na(+)-translocating NADH-quinone reductase subunit F	Metabolism
G	740505	NMA0748	1203	Exact match	11 (805), 22 (76), 41/44 (44), 162 (40), 461 (38), 174 (36), 32 (11), 254 (10), 269 (10) , 4821 (5), 5 (5)	USA, 1964	Na(+)-translocating NADH-quinone reductase subunit E	Metabolism
G	741102	NMA0749	1343	Exact match	11 (833), 32 (158), 60 (68), 174 (36), 53 (35), 162 (40), 167 (20), 41/44 (12), 22 (11), 254 (10), 4821 (5), 269 (5)	USA, 1964	Na(+)-translocating NADH-quinone reductase subunit D	Metabolism
G	741728	NMA0750	885	Exact match	11 (798), 41/44 (30), 22 (10)	USA, 1964	Na(+)-translocating NADH-quinone reductase subunit C	Metabolism

H	1476300	NMA1572	46	Exact match	167 (19), 865 (11)	The Netherlands, 1986	Pyridoxamine-5'-phosphate oxidase	Metabolism
H	1477432	NMA1573	42	Exact match	167 (15), 5 (5)	The Netherlands, 1986	Pseudouridine synthase	Metabolism
H	1478271	NMA1574	22	Exact match	167 (14)	The Netherlands, 1986	Integral membrane transporter	Environmental information processing
Isolates belonging to third pandemic wave								
H	1496471	NMA1591	634	Exact match	11 (580), 8 (20), 213 (12)	USA, 1964	type III restriction/modification system enzyme	Genetic Information Processing
H	1499437	NMA1592	765	Exact match	11 (601), 41/44 (83), 8 (19), 5 (10)	Denmark, 1962	L-lactate dehydrogenase	Metabolism
H	1501374	NMA1594	861	Exact match	11 (829), 5 (10)	Denmark, 1962	NifS-like aminotranfserase	Metabolism; Genetic Information Processing

Table 7.1. Loci of putative recombined areas with sequence matches to other hyperinvasive clonal complexes. *The letters refer to the area of contiguous allelic changes observed on the chromosome.*

7.6 Discussion

The prevailing hypothesis for the emergence of the three pandemic waves and ST-2859 in sub-Saharan Africa invokes immune escape at the sub-capsular antigens, on the basis of observations of a number of antigenic variants between ST-5 and ST-7 clones, in addition to between ST-7 and ST-2859. Differences in transferrin-binding protein B, IgA1 protease, OpaB, OpaD, and the *lgt* gene (involved in lipooligopolysaccharide synthesis) were observed between isolates of the second and third pandemics (Olyhoek *et al*. 1987; Achtman *et al*. 1992; Zhu *et*

al. 2001), and variants of the Maf adhesins, the pilin glycosylation locus and the pilU genomic region were recently found among ST-7 and ST-2859 strains by Lamelas *et al.* (2014). In addition to these documented antigenic changes, we note the loss of the antigenic lactoferrin-binding protein from ST-2859, as well as different transferrin-binding protein A alleles between the second and third pandemic waves (Table A8.10, A8.12).

The naturally-acquired immune response to meningococci induced by carriage is not well understood, and on the basis of the results of several studies, it seems unlikely that herd immunity against the sub-capsular antigens is able to develop in a population in the absence of effective vaccines. Further, meningococci would need to vary the amino acid sequences of a large number of antigens in order to evade immune responses at as many possible surface-exposed structures. However, the majority of the sub-capsular antigens are highly stable across these clones.

Differences in metabolic genes may contribute to the emergence of pandemic strains in a number of ways. There is increasing support that metabolic genes play important roles in pathogenesis in meningococci (reviewed by Schoen *et al*. (2014)), as well as in other bacterial pathogens (Shelburne *et al*. 2008) (see also Sections 1.4 & 1.5). Metabolic adaptation allows meningococci to exploit alternative host resources in invasive disseminated infections, for example, and differences have been shown in the expression of metabolic genes between meningococci adhering to lung epithelial cells and invasive meningococci growing in blood (Pagliarulo *et al*. 2004; Echenique-Rivera *et al*. 2011; Hedman *et al*. 2012; Hey 2013). It is also possible that differences in metabolic efficiency may lead to differences in transmissibility between strains, allowing different strains to assimilate resources at different rates (Schoen *et al*. 2014). Indeed, a recent study by Schoen and colleagues (2014) showed a significantly higher *in vitro* growth rate among meningococcal strains from hyperinvasive lineages compared to those from carried lineages. A breakdown of the metabolic functions of putatively introgressed

genes from hyperinvasive strains here indicated that the largest category was energy metabolism (30%), suggesting that processes contributing to the direct release of energy may contribute to the emergence of virulent strains (Figure A6). Notably, introgressed area G from isolates of the second pandemic wave encodes a number of interacting subunits from the Na(+)-translocating NADH-quinone reductase complex in the same functional network, which carries out key redox reactions of the electron transport pathway (Figure 7.6). It is possible that differences in energy metabolism between strains allow some to replicate and transmit at a more rapid rate than others. Several of the allelic differences observed also pertain to genetic information processing (16.2%), which may also lead to differences in transmissibility through differences in replication rates.

It could be argued that the results observed here could be attributed to neutral processes: that the strains from the epidemic waves differ at the loci observed solely as a result of chance and drift. Indeed, the majority of loci across the meningococcal core genome encode metabolic functions, thus it could be expected that most of the alleles which are observed to differ between the epidemic waves are at metabolic loci. The possibility that the allelic changes observed played little or no role in the emergence of the epidemic strains cannot be excluded. However, the observation that several of these variants were co-inherited from other hyperinvasive lineages suggests that there may be some link to virulence.

Our findings are consistent with the mathematical models of immunological and ecological competition of Buckee *et al*. (2008), and of those presented in this thesis, which assume that metabolic differences between strains can lead to small differences in transmission fitness. Simulations showed that only strains with the highest metabolic fitness can afford the costs of virulence, and successive replacement over time by strains with increasing metabolic fitness but similar antigenic properties. Similarly, two waves of infection of the HIN1 influenza virus in the UK between 2009 and 2011 showed no significant evidence of antigenic evolution.

Dorigatti and colleagues used a mathematical model to demonstrate that the third wave of infection could be attributed to increased transmissibility of the virus, as opposed to waning of prior immunity (Dorigatti *et al*. 2013).

The alleles with no other sequence matches to other meningococci outside the ST-5 complex may have originated from mutation. The observation that many allelic changes occurred in groups of contiguous alleles may be explained by epistatic interactions with knock-on effects. If, for example, a mutation within a protein in one part of a complex affects the functioning of an interacting protein, the latter may develop changes – resulting from mutation - to compensate for these changes and restore functioning of the complex. It is still possible, however, that such large areas of contiguous changes may have arisen through recombination, but that potential donor strains have not been sampled.

Invasive meningococcal disease is caused by only a minority of hyperinvasive lineages, which show strong temporal and geographic stability over many decades and global spread (Caugant *et al*. 1986). Here, we show that multiple recombination events between such long-lived hyperinvasive lineages may lead to the emergence of new highly virulent strains. These observations suggest that the introgressed genes from the hyperinvasive lineages may confer increased virulence and/or transmission fitness. The introgressed genes, as well as the majority of differences between epidemic waves, encoded proteins of metabolic functions, several of which were predicted to interact in the same functional network and were present in early isolates dating back several decades. This suggests that the circulation and re-assortment of co-adapted and long-lived metabolic proteins between virulent lineages may be important in the emergence of new epidemic strains. These observations are unusual given that antigenic genes are believed to vary over time as a consequence of diversifying selection, while core metabolic genes are expected to exhibit stabilising selection for conservation of function. Given than recombination between meningococci is frequent, and that a minimum of eleven independent

recombination events were found in the isolates analysed here, it is particularly intriguing that the meningococci indeed did not acquire more antigenic variants than those observed. If effective herd immunity were generated against the sub-capsular antigens, one would expect at least some changes to occur in these genes, given the high fitness advantage conferred.

There are a number of limitations which can be identified with our approach. Firstly, the isolates used do not form a coherent and representative dataset. The isolates of ST-7 and ST-2859, and in Ghana and Burkina Faso, for example, were over-represented relative to those from the first and second pandemic waves and other countries. Although the dataset is not a true epidemiological sample, it is still possible to make a number of biologically-relevant observations, such as genetic differences between the waves. Secondly, there may be additional genes present in the 153 genomes analysed which are not present in the reference strains, which may have been missed in our reference-based analyses. Gene-prediction programmes could be used to identify these additional genes in future studies (Hyatt *et al*. 2010). Thirdly, there were several areas of contiguous allelic changes which may have been acquired through lateral gene transfer, but for which a putative donor was not found. As with many tools for identifying putative donors of introgressed regions, the capacity to recognise such donors depends on their presence in the sample analysed. It is possible that these large contiguous areas may have been acquired from old or rare meningococci which have not been genetically characterised, or are no longer in circulation.

Further support that the identified metabolic genes do play roles in transmissibility and/or virulence could be obtained from experimental studies. As many of the genes identified are part of the core metabolism of meningococci, it is possible that similar genes may play roles in transmission/virulence in other bacterial pathogens.

7.7 Summary

The emergence of three pandemic waves of the ST-5 complex, in addition to ST-2859 in sub-Saharan Africa, is not well understood given their high antigenic and genetic uniformity. Here, we investigated the genetic differences between 153 meningococci of the ST-5 complex using a gene-by-gene analysis approach. The emergence of virulent strains in meningococci and other pathogenic bacteria was often demonstrated to result from the import of, or mutation within, antibiotic resistance genes, antigens or virulence factors (e.g. Croucher *et al*. 2011), but here we found very few allelic changes between the waves among the antigenic genes. While it is possible that antigenic escape at a number of loci may have played a role, our findings suggest that the emergence of the epidemic clones of the ST-5 complex has been associated with changes in the metabolic genes, including a number of large recombination events with other hyperinvasive lineages. These results are consistent with the growing body of support that metabolic genes play important roles in virulence and pathogenesis. They are also consistent with the model assumptions originally used by Buckee *et al*. (2008), and emphasised throughout the theoretical frameworks in this thesis, that the metabolic genes play important roles in transmission fitness and virulence.

Chapter 8.

Discussion

8.1 Summary of Work in Thesis

With the advent of next-generation sequencing techniques and high-throughput functional assays, metabolic genes have been demonstrated to play important roles in virulence and pathogenesis in recent years. Many theoretical frameworks do not distinguish explicitly between different types of genes, however, and few have considered the metabolic genes. Further, although insights from several studies suggest that different selection pressures operate in different areas of the genome (Chen *et al*. 2006; Lefébure & Stanhope 2007), few theoretical frameworks examine the effects of distinct competitive processes acting on different loci in the genome. Buckee *et al*. (2008) investigated the combined effects of metabolic and immunological competition on strain diversity using a stochastic model. A key prediction was that strains should show stable, non-overlapping associations between alleles of antigenic and metabolic genes. In support of their model, Buckee *et al*. (2008) found non-overlapping associations between the variable regions of the antigen PorA and groups of housekeeping alleles across a sample of *N. meningitidis* spanning over 20 years, suggesting that the combination of metabolic and antigenic competition may play a role in the maintenance of meningococcal population structure. In this work, I build upon the hypotheses and principles of the framework by Buckee *et al*., and explore further the potential role of immunological competition at antigenic loci and, subsequently ecological competition at metabolic loci in *N. meningitidis* and *S. pneumoniae,* using mathematical models and whole genome data.

In Chapter 2, the conflicting effects of within- and between- host immunological competition was explored using a deterministic model, and applied to understanding the diversity of

meningococcal and gonococcal antigenic Opa repertoires. The framework expands beyond previous theoretical frameworks that have investigated either the importance of maximising within-host duration of infection or maximising between-host transmission, providing a single mathematical framework accommodating both components of pathogen fitness. The model results showed that for pathogen systems experiencing high cross-immunity, shorter duration of infection and higher R_0 values, such as *N. meningitidis,* strains generally exhibited minimal within-strain antigenic diversity, but high levels of population-level diversity with non-overlapping antigenic repertoires; whereas for pathogens with lower levels of cross-immunity, lower R_0 values and longer durations of infection, strains with diverse antigenic repertoires and long durations of infection are more likely to arise, such as in *N. gonorrhoea*. The results were consistent with published genotypic Opa data, and may help explain the contrasting diversity and patterning among gonococcal and meningococcal Opa diversity.

In Chapter 3, metabolic loci were incorporated into the model framework, and a number of deterministic models combining both immunological competition at antigenic loci (in the form of strain-specific immunity) and ecological competition at metabolic loci were investigated. The strengths of ecological and immunological competition were relaxed over a wide range of values, and the conditions under which non-overlapping associations between antigenic type (AT) and metabolic type (MT) arose were explored. We observed a variety of population structures including stable non-overlapping associations between AT and MT over a large range of parameter space, suggesting that even weak levels of competition may be important in structuring meningococcal and pneumococcal populations.

Subsequently, the predictions of these deterministic frameworks, and the stochastic model of Buckee *et al*. (2008), were investigated in detail using 3460 isolates of *N. meningitidis* from the public PubMLST database in Chapter 4. We found that a subset of associations between PorA, FetA and clonal complex persisted for decades and proliferated globally, suggesting that

particular combinations of AT and MT are highly stable, in line with the model predictions. Consistent with the results of Buckee *et al*. (at intermediate differences in transmissibility between STs), fluctuating associations between PorA:FetA type and clonal complex were also observed. As both PorA and FetA have been used as components of protein-based vaccines, the observations of such long-lived PorA:FetA:clonal complex associations suggest that such vaccines targeted against specific hyperinvasive clonal complexes may be effective for several decades.

If MLST-type and MLEE-type are used as surrogates of MT, the predictions of the frameworks incorporating both ecological and immunological competition may be upheld across other bacterial pathogens (Table 1.4). However, this genotypic support encompasses diversity at only a small minority of genes. Using 616 published genomes of *S. pneumoniae*, we show that hypothesis of associations between serotype and MT is consistent at the whole genome level, and that MLST-defined sequence type (ST) and whole genome-defined metabolic profile are also generally correlated. The observation that variants of metabolic genes showed a significantly higher level of association (using the linkage disequilibrium and mutual information metrics) than variants of genes not belonging to metabolic processes, lends support to the hypothesis that the metabolic profile of pneumococci encodes a number of tightly linked and interacting proteins.

Bacterial genomes comprise a large variety of genes, the functions of which can be characterised into more than just antigenic and metabolic categories. In the subsequent chapter, virulence-associated genes were incorporated into the model framework comprising antigenic and metabolic genes. Simulations over a range of parameter space showed that a range of population structures are able to arise, which may correspond to a number of real pathogen examples, from avirulent commensals to obligately virulent life history strategies (Figure 6.2; Table 6.1). The effects of strain-targeted vaccinations were then investigated using the model

in the context of polysaccharide conjugate vaccines of *S. pneumoniae*. The results showed that the metabolic and virulence type originally linked with the vaccine AT before vaccination were observed in association with non-vaccine ATs after vaccination, perhaps resulting in increased virulence of the non-vaccine strains. The results of the model were supported by data from 616 whole genomes published by Croucher *et al*. (2013).

In Chapter 7, I demonstrate how insights from the multilocus strain models can be applied to inform our interpretation of genomic data, using a case study in which a succession of epidemic meningococcal strains is analysed using the gene-by-gene approach. The majority of the allelic differences between three pandemic strains of the ST-5 complex and the recently emerged ST-2859 strain had functions pertaining to metabolic processes, with very few changes in antigenic genes. This finding seemed counterintuitive given that strains of new outbreaks are often associated with novel antigenic variants/antibiotic resistance genes, however consistent with the hypothesis that metabolism may play important roles in transmissibility and virulence.

8.2 Alternative Explanations

Although the observed associations between antigenic variants and MLST/MLEE type across bacterial pathogens (Table 1.4; Chapter 4), and the associations between serotype and metabolic alleles at the whole genome level in *S. pneumoniae* (Chapter 5), are consistent with the immunological and metabolic competitive processes posited in the theoretical frameworks presented in this thesis, there are a number of other explanations which could account for such observations.

The propagation of persistent multilocus associations over time and space is a hallmark of clonal populations, in which genetic exchange is too rare to break the prevalent pattern of clonal population structure. Clonal descent will undoubtedly have contributed to the observations of

associations between ATs and MTs in the pathogens listed in Table 1.4. Such an explanation is especially pertinent to those pathogens which do not experience frequent recombination. For example, the STs and OspC antigen of *Borrelia lusitaniae* show non-overlapping associations, but recombination across the chromosome is relatively rare in this species (Victorino *et al*. 2008).

However, the bacterial species presented in Table 1.4 have been previously shown to manifest frequent rates of recombination; the thousands of AT:ST combinations found among *S. pneumonia* and *N. meningitidis* (Figure 8.1) reflects the large amounts of diversity generated by recombination in these populations. For example, among the *S. pneumoniae* dataset extracted from the public MLST database, there were over 9500 variants exhibiting different associations between serotype and ST. Out of the large pool of genetic variants of associations between metabolic and antigenic alleles observed (Figure 8.1), only a subset of successful genotypes are able to propagate to high frequencies, which manifest a non-overlapping pattern between these alleles. (Figure 8.2, Tables 8.1-2). The model of Buckee *et al*. (2008), and those presented in this thesis, assume that all strain combinations are able to circulate in the population, but the most prevalent strains are observed to manifest a roughly non-overlapping in order to minimise competition for hosts and metabolic resources, allowing each to exploit a unique antigenic and metabolic niche. It is this fitness advantage which may permit them to persist for many years and sometimes disseminate on a global scale (Watkins & Maiden 2012). The low frequency variants generated by recombination which show overlapping metabolic and antigenic variants are therefore at a fitness disadvantage with the dominant non-overlapping variants.

Although pneumococci, meningococci and the other pathogens listed in Table 1.4 demonstrate frequent recombination, and although the aforementioned observation that only the most

frequent, successful strains exhibit non-overlapping associations between antigenic and metabolic alleles is compelling, it should be noted that clonal descent cannot definitively be excluded as the most important factor maintaining the associations observed between antigenic and metabolic alleles (both at the MLST- and whole genome- levels), and indeed for maintaining the population structure of bacterial pathogens generally. The processes of clonal descent and interstrain competition are not mutually exclusive, however. It may be that ecological and immunological competition played an important role in the diversification of bacterial strains earlier on in the evolution of these pathogens (in the establishment of non-overlapping niches to avoid resource competition, for example), and that clonal descent has since maintained the allelic associations. Understanding the contribution of selective processes (such as competition) relative to clonal descent in the maintenance of bacterial population structure is a challenging but important question to resolve.

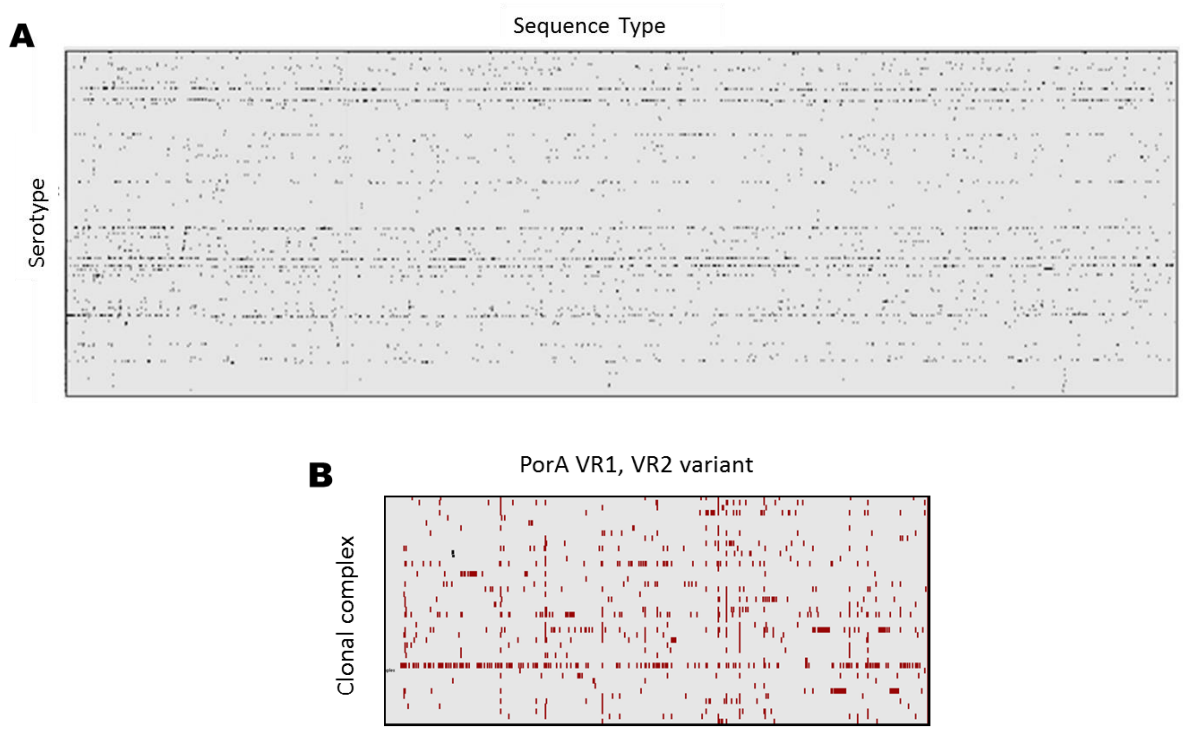


Figure 8.1. Diversity of associations between MLST-defined metabolic alleles and antigenic alleles for* S. pneumoniae *and* N. meningitidis*, as recorded from public databases. *(A) Plot of 9,559 allelic combinations found between serotype and sequence type among 15,156 isolates of* Streptococcus pneumoniae*. Data were extracted from the MLST website on 28/04/2013 (http://spneumoniae.mlst.net/). Of the isolates which could be delineated into carriage and invasive infection, 44% were obtained from carriage, and 56% from invasive disease. (B) Plot of 786 allelic combinations found between PorA and clonal complex among 4325 isolates of Neisseria meningitidis. Data were extracted from the PubMLST database on 20/04/2013 (http://pubmlst.org/neisseria/. Of the isolates which could be delineated into carriage and invasive infection, 31% were obtained from carriage, and 69% from invasive disease. (*

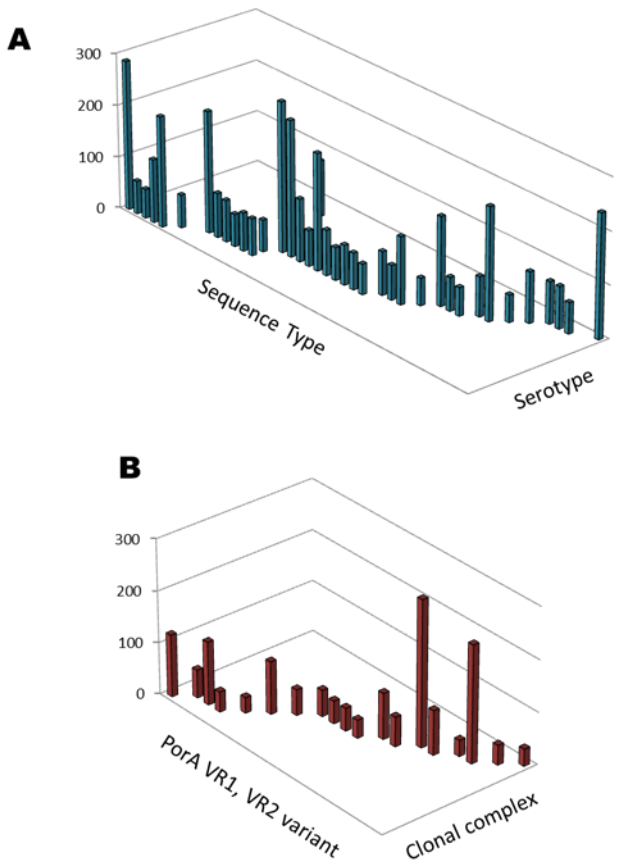


Figure 8.2. The most frequent strains of* S. pneumoniae *and* N. meningitidis *exhibit non-overlapping associations between MLST-defined metabolic alleles and antigenic alleles. *(A) The most frequent allelic combinations of serotype and sequence type (with a frequency > 50) among* S. pneumoniae. *(B) The most frequent allelic combinations of PorA and clonal complex (with a frequency > 20 isolates per combination) among* N. meningitidis. *All isolate information was extracted from the PubMLST and MLST websites as described Figure 8.1.*

	Serotype											
ST	1	11A	14	15A	19A	19F	22F	23F	3	35B	9V	6B	7F
217	287												
612	63												
303	56												
618	122												
306	213												
62		63											
9			234										
124			85										
15			80										
156			62								108		
143			73										
63			71	61									
199					288								
320					261								
1201					120								
193					69								
276					224								
847					87								
172					63								
2062					76								
202					70								
416					58								
4414						83							
423						66							
236						129							
433							52						
81								170					
802								64					
242								54					
458									76				
180									216				
558										52			
162											97		
90												80	
315												80	
176												59	
191													236

Table 8.1. Combinations of serotype and sequence type for* S. pneumoniae *for which more than 50 isolates have been reported.

				Clonal	Complex					
PorA: FetA variant	ST-1	ST-11	ST-213	ST-22	ST-23	ST-269	ST-32	ST-41/44	ST-549	ST-8
P1.5-2,10: F3-5	120									
P1.5,2: F1-1		54								
P1.5,2: F3-6		122								
P1.5-1,10-8: F3-6		39								
P1.18,25-15: F5-5										
P1.22,14: F5-5			102							
P1.18-1,3: F4-1				50						
P1.5-2,10-1: F4-1					51					
P1.5-1,2-2: F5-8					44					
P1.5-1,2-2: F5-8					44					
P1.5-1,10-1: F4-1					35					
P1.22,9: F5-12						89				
P1.19-1,15-11: F5-						57				
P1.19,15: F5-1							283			
P1.7,16: F3-3							86			
P1.19,15: F1-7								31		
P1.7-2,4: F1-5								227		
P1.5-1,2-2: F4-3									38	
P1.5-1,2-2: F3-6										33

Table 8.2. Combinations of PorA, FetA and clonal complex for* N. meningitidis *for which more than 20 isolates have been reported.

Other models of bacterial population structure include the epidemic model (Smith *et al*. 1993) and neutral models, such as the neutral microepidemic model (Fraser *et al*. 2005). As discussed in Section 1.3, we believe it is likely that both neutral processes and the transmission success of epidemic strains play important roles in the maintenance of pathogen population structures. However, the results of previous studies which aimed to reproduce the population structure of various pathogens using neutral/microepidemic processes significantly underestimated the diversity/longevity of STs observed (Jolley *et al*. 2005; Buckee *et al*. 2008), and the epidemic model is also unable to account for the longevity of lineages observed. The structuring of bacterial populations as a result of the distribution of different restriction-modification systems across clades has also been proposed (Budroni *et al*. 2011). However, as discussed in Section

1.3, the association between clades and restriction-modification systems could also be interpreted as a consequence of the diversification processes through other mechanisms (such as variation in transmissibility and immune selection).

Linkage disequilibrium in recombining populations can also result superficially from biased sampling of hyperinvasive clones, in samples which are not realistic representations of the natural population diversity as a whole. Although the PubMLST database does not represent a balanced population sample, the isolates shown in these "collapsed" populations in Figure 8.2 do comprise isolates from both carriage and invasive disease – thus the linkage disequilibrium observed in Figure 8.2 is unlikely to result from biased sampling of hyperinvasive isolates. Further, non-overlapping associations are evident in a number of population samples obtained solely from carriage, such as the associations between serotype and metabolic profile of 616 genomes from carriage in Chapter 5. The studies listed in Table 1.4 comprise bacterial isolates from a variety of sources, including both invasive and carriage collections. The possibility that the non-overlapping patterns observed among the invasive isolates in these studies may be attributed to biased sampling cannot be ruled out, however.

It is also possible that some of the specific non-overlapping associations observed in Table 1.4 and Chapter 5 have arisen as a result of epistatic interactions between antigenic and metabolic areas of the genome, resulting in heightened fitness. It may be that specialised metabolic machinery is required for the uptake and synthesis of particular polysaccharide sugars, for example, which then shapes the remainder of the metabolome. A recent study by Croucher *et al*. (2015) used an experimental approach with knockouts of the *cps* locus to test if particular pneumococcal genotypes were more suited to certain serogroups, and found that reduced growth rates were found among both the within-serogroup and between-serogroup switched variants – thus suggesting that serogroup-specific adaptations may not be wholly responsible for the frequent rates of within-serogroup changes observed. Croucher *et al*. also examined the

distribution of the carbohydrate uptake genes functionally characterised by Bidossi *et al*. (2012) across 616 genomes of *S. pneumoniae* to test the hypothesis that limitations to acquiring the necessary carbohydrates to synthesise a given serogroup may account for the patterns observed, but the resulting distribution did not generally support this hypothesis. It is possible that epistasis at other carbohydrate metabolism genes may play a role; but it is also difficult to postulate whether the specialisation of metabolic genes responsible for antigen production would result in the modification of the remainder of the metabolic genes required for core metabolic processes. Perhaps insights from functional genomic studies will shed light on the dominance of antigen production (and any specialised downstream processes) in the metabolome. The theoretical frameworks predict that associations between AT and MT are essentially arbitrary, suggesting that an AT may associate primarily with one MT in one location/time period, and with another in a different location/period. This is consistent with the observation that one AT associates with multiple MTs (and vice versa) across different time periods and geographical regions in *N. meningitidis* (Buckee *et al*. 2008). Associations between serotype and ST, however, are often consistent in *S. pneumoniae*. There are nevertheless a few exceptions, as different STs have been observed with the same serotype in distinct geographical locations (Enright & Spratt 1998). For example, pneumococci of serotype 19A were primarily associated with ST482 in a study in Finland, but with a variety of STs in a study in the UK (Brueggemann *et al*. 2003; Hanage *et al*. 2005). The occurrence of epistasis and ecological and immunological competition in pathogen populations is not mutually exclusive – it may be that both processes act together to shape the diversity of pathogen populations.

8.3 Parallels with Species Co-existence in Community Ecology Theory

In addition to enhancing our understanding of the mechanisms of competition which promote the maintenance of strain diversity in pathogen populations, the model frameworks of ecological and immunological competition presented in this thesis are also pertinent to the question of species co-existence in macro-ecological communities, which remains a central question in community ecology.

The roots of species co-existence research can be traced to one of the earliest axioms of ecology, Gause's Principle, which dictates that no two species can occupy the same niche as the superior competitor species would exclude the other (Gause 1934). The niche model of species co-existence thus invokes "stabilising" processes to reduce niche overlap between species, thereby minimising the impact of fitness differences on competitive interactions (Chesson 2000). The classic Lotka-Volterra models invoke such stabilising processes for example: so long as intraspecific exceeds interspecific competition, species should limit their own populations more than they limit others (Lotka 1925; Voltera, 1926; Macarthur & Levins 1967). Alternatively, the "neutral" theory, emphasises the importance of similarity among individuals, rather than differences, in promoting species co-existence (Hubbell 2001). Neutral models assume equivalent fitness among individuals, with population dynamics driven by random variation in births, deaths, and dispersal. Processes which reduce fitness differences between species – known as "equalising" – are therefore key to neutral dynamics (Chesson 2000).

Although neutral frameworks have successfully predicted a number of empirical ecological patterns (Bell 2001), the key assumption of species equivalence generally has little empirical support (Chave 2004). In contrast, despite substantial empirical evidence for niche differences

between co-occurring species (McKane *et al*. 2002), the collective importance of these differences for maintaining species diversity has been difficult to ascertain. The continuing debates over the relative importance of neutral and selective processes in the creation and maintenance of species diversity have resulted in a surge of theoretical frameworks over the past 50 years (Kirkpatrick and Ravigne 2002). In a review from the last 40 years, Sergey Gavrilets emphasised the need for simple models comparable to other areas of ecology, with "a shift in focus to identifying more general rules and patterns" (Gavrilets 2003).

Our general deterministic framework, based on competition linked to different traits, may have broad application to the questions of both strain co-existence in infectious disease epidemiology and species co-existence in community ecology. The model framework invokes the interaction of competition for different types of resources among individuals, some of which are depleted more rapidly than others (determined by μ and σ). As colonisation in the host is often transient, ecological competition at metabolic loci can be thought of as a relatively short-term process, whereas the infection process of many pathogens confers immunity which is longer lasting, and acts for a period longer than that of colonisation/infection alone. Thus we can infer that immunological competition in pathogen systems is analogous to long-term competition for more permanent occupation of physical sites in macro-ecological systems (such as nesting sites), or for resources with slow replenishment/recovery rates; and ecological competition for metabolic resources in pathogen systems parallels short-term competition for the temporary occupying of sites (such as space in a feeding site), or for resources with a rapid replenishment rate (Table 8.3). The interactions of varying short-term and long-term competition, to our knowledge, has not yet been explored in models of species co-existence in macro-ecology. The models suggest that competition for short- and long- term resources can lead to segregation of unique resource requirements among genotypes, thereby enabling stable species co-existence.

	Short-term competition	Long-term competition
Ecological	Temporary physical occupation of a given site Depletion of resources with rapid recovery/ replenishment rates	Permanent physical occupation of a given site Depletion of resources with low rats of recovery/replenishment
Immunological	Specific immune response activated only in the presence of the pathogen (e.g. innate responses to PAMPs; Pathogen Associated Molecular Patterns)	Specific immune response with immunological memory (e.g. adaptive responses)

Table 8.3. Interpretations of short- and long-term competitive processes in ecological and immunological contexts.

Let us consider a hypothetical system in which various animal species are in direct competition for three plant resources, 1, 2 and 3, the sites of which they visit several times each day, as well as space for permanent nesting sites in three micro-habitats A, B and C. The temporary occupation of a feeding site by one individual prevents others from feeding on that resource, and the nesting sites are permanently unavailable for others to colonise once claimed. Our

models suggest that the interplay of strong short-term competition for space at the foraging site and strong long-term competition for the nesting site will eventually lead to non-overlapping associations between nesting sites A, B and C and resources 1, 2, and 3 – thus leading to long-term resource-partitioning on both resource axes and therefore species co-existence.

An example of largely non-overlapping trait combinations across distinct species is provided by the *Anolis* lizards, a highly speciose group of insectivorous lizards found across the Caribbean and in certain parts of the Americas. Anoles are classified as belonging to one of six ecomorph classes, which define a lizard's food resource and microhabitat. Anoles also manifest a large range of dewlap phenotypes, used particularly by males to mark territories and attract mates. The model would predict that the particular associations between the traits of interest, although non-overlapping, are random and arbitrary. Non-overlapping associations between dewlap phenotype and ecomorph can be observed across several islands of the Antilles, which seem to differ across islands (Figure 8.3). An empirical characterisation of dewlap diversity was attempted for the first time by Nicholson *et al*. (2007). Dewlap colours and patterns were divided into discrete categories and systematically recorded for 140 species across four islands. They found that species of the same ecomorph were no more similar in dewlap configuration than expected by chance. However, other factors may influence the evolution of dewlap phenotypes, such as light conditions (Leal and Fleishman, 2004).

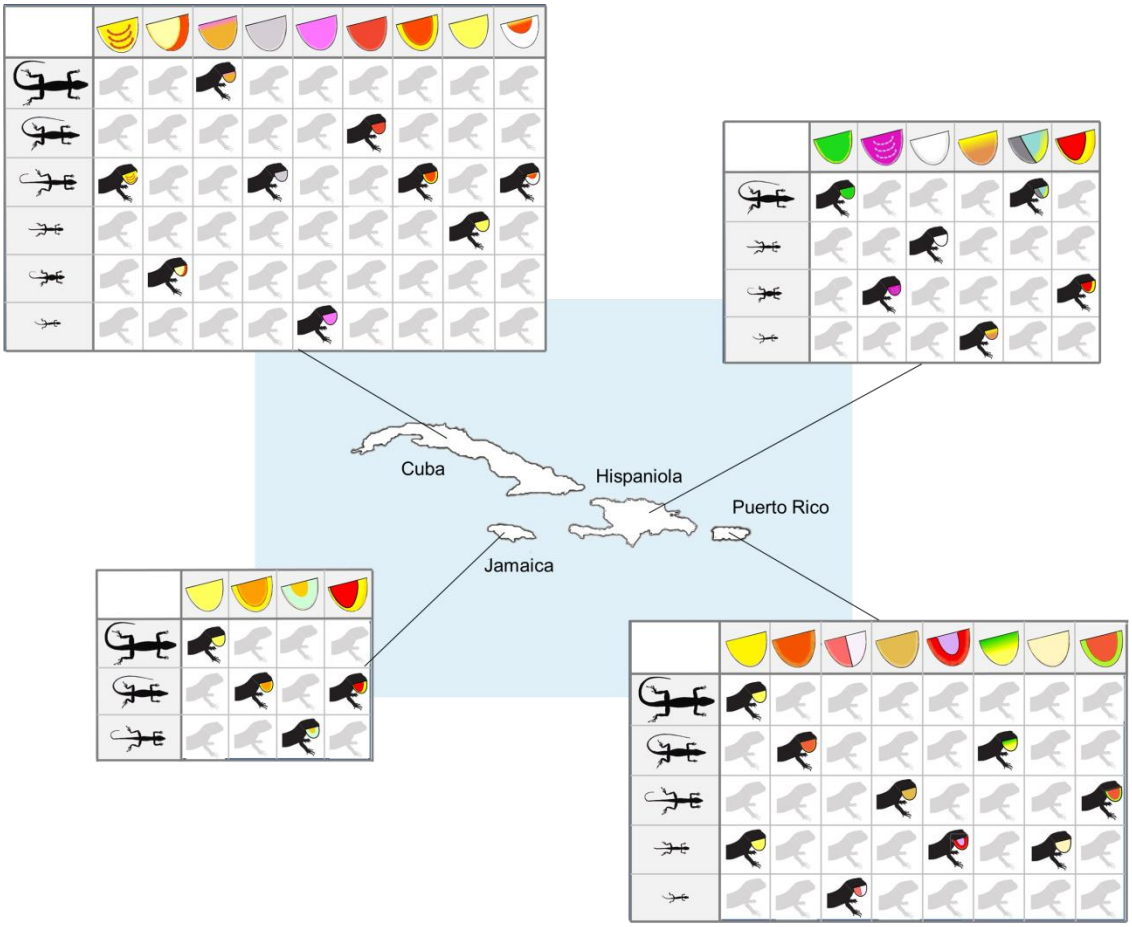


Figure 8.3. Anolis lizard trait distribution in sympatric communities across four Caribbean islands. *Each grid represents the trait space of possible associations between dewlap morphologies (columns) and ecomorph traits (rows) on each island. The number of possible trait combinations (grey outlines) is large compared to the smaller assemblage of species which are actually observed (black outlines). The list of sympatric species from each island and dewlap characteristics for each species were taken from Nicholson et al. (2007). Two species did not possess a dewlap and were not included, thus providing 28 species in total. Each dewlap showing a different pattern and/or colour was counted as a unique dewlap configuration, and ecotype information for each species was taken from Losos (2009).*

8.3.1 Importance of Stabilising Effects

The interplay between ecological and immunological competition in our model supports the importance of niche differentiation in co-existence theory. Niche partitioning increases competition with conspecific individuals, and species consequently limit their own growth to a greater extent than others. This provides a stabilising mechanism to explain the co-existence of distinct species despite fitness differences. Therefore, if strains are able to inhibit their own growth to a greater extent than the growth of other strains, this provides a stabilising mechanism to explain the co-existence of distinct strains. The models described here invoke such processes, as strains with the same MT are prohibited – to some extent – from co-infecting the same host but can freely co-infect hosts infected with different MTs. Equally, if strains elicit specific immune responses which act to repress subsequent infection against the same strain to a greater extent than infection with other strains, the strength of intraspecific competition would be greater than interspecific competition. Thus, the results of our model support the importance of niche partitioning on multiple, independent resource axes, as the population is ultimately segregated into distinct strains exploiting distinct antigenic and metabolic niches.

The results of a series of models by Colijn *et al*. investigating the mechanisms of co-existence of drug-resistant and -susceptible strains of *S. pneumoniae* also showed that strain-specific immunity increased the likelihood of stable strain co-existence (Colijn *et al*. 2010). A stochastic model by Cobey and Lipsitch (2012), in contrast, emphasised the importance of both stabilising and equalising mechanisms. In line with known data on pneumococcal cross-immunity, the model assumed a non-specific immune response which had the effect of decreasing the duration of colonisation by a given serotype, upon repeated exposure, over the course of a host's lifetime. Such an effect acted to decrease fitness differences between serotypes, as the most

successful genotypes become less prevalent over the course of a host's lifetime. Serotype-specific immunity acted as a stabilising mechanism. The combination of intermediate anticapsular immunity and a non-linear increase in acquired non-specific immunity was able to reproduce the observed diversity in serotypes in addition to a number of other features of pneumococcal epidemiology.

The assumptions of strain-specific immunity and ecological interference in the models presented here contravene the structural neutrality of strain-coexistence models; that is, they contain implicit mechanisms which lead to the stable co-existence of strains "for free" (Lipsitch *et al*. 2009). They break the ecological criterion of neutral models, which stipulates that the ecological dynamics of the model should not depend on the properties of the strains themselves. Thus, perhaps it would be fruitful to examine a model system where the immune states depend on the number of times a host has been infected as demonstrated by Lipsitch *et al*. (2009), or to examine strain-specific immunity in a stochastic individual-based framework (Cobey & Lipsitch, 2012) to decrease any bias towards stable co-existence.

8.4 Future Studies

The theoretical frameworks and genomic analyses presented in this thesis support the hypothesis that ecological competition at metabolic loci and immunological competition at antigenic loci may contribute to the maintenance of the structure of meningococcal and pneumococcal populations (at least to the structuring observed at the pneumococcal capsule and meningococcal PorA and FetA antigens). To ascertain whether the model predictions are applicable to the other pathogens listed in Table 1.4, antigenic-specific immunity which protects against subsequent infection (to some extent) needs to be demonstrated. Although strong antigenic-specific immune responses have been recorded for many of these pathogens,

the strength of protective immunity against subsequent infection is not yet clear. Furthermore, as mentioned in Chapter 3, although some protective immunity has demonstrated for some pneumococcal capsule types, the strengths seem to vary considerably across serotypes, with some providing negligible measurable protective effects. Furthermore, there are a number of pneumococcal protein antigens which may stimulate strong immune responses, but do not seem to protect against subsequent infection (e.g. Goldblatt *et al*., 2005); the results of the models in this thesis are not applicable to these antigens.

The genetic data which were used in combination with the theoretical frameworks comprised only nucleotide sequence data; an area of future study may involve the testing of such frameworks using functional genomics or methylome approaches. Using sequence data to investigate the hypothesis of ecological competition between strains with similar metabolic profiles operating within the host is also relatively limited. There are a number of experimental approaches which may ultimately allow for the *in vivo* and *in vitro* testing of this hypothesis, such as continuous-culture chemostat systems and flowcells. Throughout this work, I have reviewed a number of competing hypotheses which account for the structure of highly recombining bacterial pathogen populations. Although each one is able to explain a number of observations from bacterial pathogens, the processes and selection pressures are not mutually exclusive, and it is likely that a combination of neutral, selective and mechanistic factors play important roles.

A growing number of studies are focussing on the importance of the metabolic genes in transmission and virulence, and it is becoming increasingly apparent that conceptual frameworks which account explicitly for the metabolic genes are required. The models presented in this thesis provide examples of such analytical platforms on which to base and interpret the accumulating data acquired from whole genome studies in the coming years.

References

Abu Kwaik, Y. & Bumann, D., 2013. Microbial quest for food in vivo: "nutritional virulence" as an emerging paradigm. *Cell Microbiol*, 15, pp.882–890.

Achtman, M. et al., 1992. A comparison of the variable antigens expressed by clone IV-1 and subgroup III of Neisseria meningitidis serogroup A. *J Infect Dis*, 165, pp.53–68.

Achtman, M., 2008. Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. *Annual Review of Microbiology*, 62, pp.53–70.

Achtman, M. et al., 2012. Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella enterica. *Plos Pathogens*, 8.

Achtman, M. et al., 1999. Recombination and clonal groupings within Helicobacter pylori from different geographical regions. *Molecular Microbiology*, 32, pp.459–470.

Alber, D. et al., 2001. Genetic diversity of Neisseria lactamica strains from epidemiologically defined carriers. *Journal of Clinical Microbiology*, 39, pp.1710–1715.

Amir, J., Louie, L. & Granoff, D.M., 2005. Naturally-acquired immunity to Neisseria meningitidis group A. *Vaccine*, 23, pp.977–983.

Anderson, R.M. & May, R.M., 1991. Infectious Diseases Of Humans Dynamics And Control. *Oxford University Press: Oxford, UK; New York, USA.*

Anderton, J.M. et al., 2007. E-cadherin is a receptor for the common protein pneumococcal surface adhesin A (PsaA) of Streptococcus pneumoniae. *Microbial Pathogenesis*, 42, pp.225–236.

Antia, R., Nowak, M.A. & Anderson, R.M., 1996. Antigenic variation and the within-host dynamics of parasites. *Proceedings of the National Academy of Sciences of the United States of America*, 93, pp.985–989.

Auranen, K. et al., 2010. Between-strain competition in acquisition and clearance of pneumococcal carriage-epidemiologic evidence from a longitudinal study of day-care children. *Am J Epidemiol*, 171, pp.169–176.

Baart, G.J.E. et al., 2008. Modeling Neisseria meningitidis B metabolism at different specific growth rates. *Biotechnology and Bioengineering*, 101, pp.1022–1035.

Baart, G.J.E. et al., 2007. Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes. *Genome biology*, 8, p.R136.

Baek, J.Y. et al., 2011. Comparison of genotypes of Streptococcus pneumoniae serotypes 6A and 6B before and after the introduction of PCV7 vaccination in Korea. *Diagnostic Microbiology and Infectious Disease*, 69, pp.370–375.

Bagnoli, F. et al., 2008. A second pilus type in Streptococcus pneumoniae is prevalent in emerging serotypes and mediates adhesion to host cells. *Journal of Bacteriology*, 190, pp.5480–5492.

Bambini, S. et al., 2013. An Analysis of the Sequence Variability of Meningococcal fHbp, NadA and NHBA over a 50-Year Period in the Netherlands. *Plos One*, 8.

Barocchi, M.A. et al., 2006. A pneumococcal pilus influences virulence and host inflammatory responses. *Proc Natl Acad Sci U S A*, 103, pp.2857–2862.

Bart, A. et al., 2001. The population structure of Neisseria meningitidis serogroup A fits the predictions for clonality. *Infection Genetics and Evolution*, 1, pp.117–122.

Basset, A. et al., 2007. Association of the pneumococcal pilus with certain capsular serotypes but not with increased virulence. *Journal of Clinical Microbiology*, 45, pp.1684–1689.

Baumler, D.J. et al., 2011. The evolution of metabolic networks of E. coli. *Bmc Systems Biology*, 5.

Beall, B. et al., 2006. Pre- and postvaccination clonal compositions of invasive pneumococcal serotypes for isolates collected in the United States in 1999, 2001, and 2002. *Journal of Clinical Microbiology*, 44, pp.999–1017.

Beall, B.W. et al., 2011. Shifting Genetic Structure of Invasive Serotype 19A Pneumococci in the United States. *Journal of Infectious Diseases*, 203, pp.1360–1368.

Van Belkum, A., 2007. Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *Fems Immunology and Medical Microbiology*, 49, pp.22–27.

Bell, G., 2001. Neutral macroecology. *Science (New York, N.Y.)*, 293, pp.2413–2418.

Bennett, J.S. et al., 2009. A common gene pool for the Neisseria FetA antigen. *International Journal of Medical Microbiology*, 299, pp.133–139.

Bennett, J.S. et al., 2005. Genetic diversity and carriage dynamics of Neisseria lactamica in infants. *Infection and Immunity*, 73, pp.2424–2432.

Bennett, J.S. et al., 2010. Independent evolution of the core and accessory gene sets in the genus Neisseria: insights gained from the genome of Neisseria lactamica isolate 020-06. *Bmc Genomics*, 11.

Bennett, J.S. et al., 2007. Species status of Neisseria gonorrhoeae: evolutionary and epidemiological inferences from multilocus sequence typing. *Bmc Biology*, 5.

Berry, A.M. et al., 1994. Cloning And Nucleotide-Sequence Of The Streptococcus-Pneumoniae Hyaluronidase Gene And Purification Of The Enzyme From Recombinant Escherichia-Coli. *Infection and Immunity*, 62, pp.1101–1108.

Berry, A.M. et al., 1989. Contribution Of Autolysin To Virulence Of Streptococcus-Pneumoniae. *Infection and Immunity*, 57, pp.2324–2330.

Bessen, D.E., 2009. Population biology of the human restricted pathogen, Streptococcus pyogenes. *Infection Genetics and Evolution*, 9, pp.581–593.

Bethe, G. et al., 2001. The cell wall-associated serine protease PrtA: a highly conserved virulence factor of Streptococcus pneumoniae. *Fems Microbiology Letters*, 205, pp.99–104.

Bhat, K.S. et al., 1991. The Opacity Proteins Of Neisseria-Gonorrhoeae Strain Ms11 Are Encoded By A Family Of 11 Complete Genes. *Molecular Microbiology*, 5, pp.1889–1901.

Bidossi, A. et al., 2012. A Functional Genomics Approach to Establish the Complement of Carbohydrate Transporters in Streptococcus pneumoniae. *Plos One*, 7.

Bilek, N., Ison, C.A. & Spratt, B.G., 2009. Relative Contributions of Recombination and Mutation to the Diversification of the opa Gene Repertoire of Neisseria gonorrhoeae. *Journal of Bacteriology*, 191, pp.1878–1890.

Bjune, G. et al., 1991. *Effect of outer membrane vesicle vaccine against group B meningococcal disease in Norway.*,

Blakebrough, I.S. et al., 1982. The epidemiology of infections due to *Neisseria meningitidis* and *Neisseria lactamica* in a northern Nigerian community. *The Journal of Infectious Diseases*, 146, pp.626–637.

Bogaert, D., de Groot, R. & Hermans, P.W.M., 2004. Streptococcus pneumoniae colonisation: the key to pneumococcal disease. *Lancet Infectious Diseases*, 4, pp.144–154.

Boisier, P. et al., 2007. Meningococcal meningitis: unprecedented incidence of serogroup X-related cases in 2006 in Niger. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 44, pp.657–663.

Bos, M.P., Hogan, D. & Belland, R.J., 1997. Selection of Opa(+) Neisseria gonorrhoeae by limited availability of normal human serum. *Infection and Immunity*, 65, pp.645–650.

Bottomley, C. et al., 2013. A mathematical model of serotype replacement in pneumococcal carriage following vaccination. *J R Soc Interface*, 10, p.20130786.

Boulton, I.C. & Gray-Owen, S.D., 2002. Neisserial binding to CEACAM1 arrests the activation and proliferation of CD4(+) T lymphocytes. *Nature Immunology*, 3, pp.229–236.

Bratcher, H.B. et al., 2014. A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. *BMC Genomics*, 15, p.1138.

Brehony, C., Wilson, D.J. & Maiden, M.C.J., 2009. Variation of the factor H-binding protein of Neisseria meningitidis. *Microbiology-Sgm*, 155, pp.4155–4169.

Bremermann, H.J. & Thieme, H.R., 1989. A Competitive-Exclusion Principle For Pathogen Virulence. *Journal of Mathematical Biology*, 27, pp.179–190.

Briles, D.E. et al., 1992. Strong association between capsular type and virulence for mice among human isolates of Streptococcus pneumoniae. *Infect Immun*, 60, pp.111–116.

Brito, D.A., Ramirez, M. & De Lencastre, H., 2003. Serotyping Streptococcus pneumoniae by multiplex PCR. *Journal of Clinical Microbiology*, 41(6), pp.2378–2384.

Broeker, M. et al., 2012. Increase of meningococcal serogroup Y cases in Europe A reason for concern? *Human Vaccines & Immunotherapeutics*, 8, pp.685–688.

Brooks, G.F. & Ingwer, I., 1978. Studies On Relationships Between Serum Bactericidal Activity And Uncomplicated Genital Infections Due To Neisseria-Gonorrhoeae. *Journal of Infectious Diseases*, 138, pp.333–339.

Brueggemann, A.B. et al., 2003. Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. *Journal of Infectious Diseases*, 187, pp.1424–1432.

Brueggemann, A.B. et al., 2004. Temporal and geographic stability of the serogroup-specific invasive disease potential of Streptococcus pneumoniae in children. *Journal of Infectious Diseases*, 190, pp.1203–1211.

Brueggemann, A.B. et al., 2007. Vaccine escape recombinants emerge after pneumococcal vaccination in the united states. *Plos Pathogens*, 3, pp.1628–1636.

Bryant, J.M. et al., 2013. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. *Lancet*, 381, pp.1551–1560.

Buchan, H. et al., 1993. Morbidity Following Pelvic Inflammatory Disease. *British Journal of Obstetrics and Gynaecology*, 100, pp.558–562.

Buchanan, T.M. et al., 1980. Gonococcal Salpingitis Is Less Likely To Recur With Neisseria-Gonorrhoeae Of The Same Principal Outer-Membrane Protein Antigenic Type. *American Journal of Obstetrics and Gynecology*, 138, pp.978–980.

Buckee, C.O. et al., 2010. Long-term evolution of antigen repertoires among carried meningococci. *Proceedings of the Royal Society B-Biological Sciences*, 277, pp.1635–1641.

Buckee, C.O. et al., 2008. Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America*, 105, pp.15082–15087.

Buckee, C.O. et al., 2011. Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108, pp.15504–15509.

Buckwalter, C.M. & King, S.J., 2012. Pneumococcal carbohydrate transport: food for thought. *Trends in Microbiology*, 20, pp.517–522.

Budroni, S. et al., 2011. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 108, pp.4494–4499.

Callaghan, M.J. et al., 2011. Potential of Recombinant Opa Proteins as Vaccine Candidates against Hyperinvasive Meningococci. *Infection and Immunity*, 79, pp.2810–2818.

Callaghan, M.J. et al., 2008. The effect of immune selection on the structure of the meningococcal Opa protein repertoire. *Plos Pathogens*, 4.

Callaghan, M.J., Jolley, K.A. & Maiden, M.C.J., 2006. Opacity-associated adhesin repertoire in hyperinvasive Neisseria meningitidis. *Infection and Immunity*, 74, pp.5085–5094.

Caron, F. et al., 2011. From tailor-made to ready-to-wear meningococcal B vaccines: Longitudinal study of a clonal meningococcal B outbreak. *The Lancet Infectious Diseases*, 11, pp.455–463.

Cartwright, K.A., Stuart, J.M. & Noah, N.D., 1986. An outbreak of meningococcal disease in Gloucestershire. *Lancet*, 2, pp.558–561.

Cartwright, K.A. V et al., 1987. The Stonehouse Survey - Nasopharyngeal Carriage Of Meningococci And Neisseria-Lactamica. *Epidemiology and Infection*, 99, pp.591–601.

Carvalho, M.D.G. et al., 2010. Revisiting pneumococcal carriage by use of broth enrichment and PCR techniques for enhanced detection of carriage and serotypes. *Journal of Clinical Microbiology*, 48, pp.1611–1618.

Carver, T.J. et al., 2005. ACT: The Artemis comparison tool. *Bioinformatics*, 21(16), pp.3422–3423.

Castillochavez, C. et al., 1989. Epidemiological Models With Age Structure, Proportionate Mixing, And Cross-Immunity. *Journal of Mathematical Biology*, 27, pp.233–258.

Caugant, D.A. et al., 1988. Clonal diversity of Neisseria meningitidis from a population of asymptomatic carriers. *Infection and Immunity*, 56, pp.2060–2068.

Caugant, D.A. et al., 1987. Genetic-Structure Of Neisseria-Meningitidis Populations In Relation To Serogroup, Serotype, And Outer-Membrane Protein Pattern. *Journal of Bacteriology*, 169, pp.2781–2792.

Caugant, D.A. et al., 1986. Intercontinental spread of a genetically distinctive complex of clones of Neisseria meningitidis causing epidemic disease. *Proc Natl Acad Sci U S A*, 83, pp.4927–4931.

Caugant, D.A. et al., 2012. Molecular characterization of invasive meningococcal isolates from countries in the African meningitis belt before introduction of a serogroup A conjugate vaccine. *PLoS One*, 7, p.e46019.

Caugant, D.A. & Maiden, M.C.J., 2009. Meningococcal carriage and disease-Population biology and evolution. *Vaccine*, 27, pp.B64–B70.

Caugant, D.A., Tzanakaki, G. & Kriz, P., 2007a. Lessons from meningococcal carriage studies. *FEMS Microbiology Reviews*, 31(1), pp.52–63.

Caugant, D.A., Tzanakaki, G. & Kriz, P., 2007b. Lessons from meningococcal carriage studies. *FEMS Microbiology Reviews*, 31(1), pp.52–63.

CE, F., MP, P. & FM, L., 2012. Development of a group A meningococcal conjugate vaccine, MenAfriVac. *Hum Vaccin Immunother*, 8, pp.715–724.

Chave, J., 2004. Neutral theory and community ecology. *Ecology Letters*, 7, pp.241–253.

Chen, S.L. et al., 2006. Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proceedings of the National Academy of Sciences of the United States of America*, 103, pp.5977–5982.

Chesson, P., 2000. Mechanisms Of Maintenance Of Species Diversity. *Annu. Rev. Ecol. Syst*, 31, pp.343–366.

Chewapreecha, C. et al., 2014. Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics*, 46, p.305–+.

Chin, C.-S. et al., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10, pp.563–9.

Clarke, S.C. et al., 2006. Pneumococci causing invasive disease in children prior to the introduction of pneumococcal conjugate vaccine in Scotland. *Journal of Medical Microbiology*, 55, pp.1079–1084.

Claus, H. et al., 2005. Genetic analysis of meningococci carried by children and young adults. *J Infect Dis*, 191, pp.1263–1271.

Coates, A.R.M. et al., 1981. Antigenic diversity of mycobacterium tuberculosis and mycobacterium bovis detected by means of monoclonal antibodies. *Lancet*, 2, pp.167–169.

Cobey, S. & Lipsitch, M., 2012. Niche and Neutral Effects of Acquired Immunity Permit Coexistence of Pneumococcal Serotypes. *Science*, 335, pp.1376–1380.

Coffey, T.J. et al., 1995. Genetic analysis of clinical isolates of Streptococcus pneumoniae with high-level resistance to expanded-spectrum cephalosporins. *Antimicrobial Agents and Chemotherapy*, 39, pp.1306–1313.

Cohen, T., Colijn, C. & Murray, M., 2008. Modeling the effects of strain diversity and mechanisms of strain competition on the potential performance of new tuberculosis vaccines. *Proceedings of the National Academy of Sciences of the United States of America*, 105, pp.16302–16307.

Cohn, A.C. et al., 2010. Changes in Neisseria meningitidis Disease Epidemiology in the United States, 1998-2007: Implications for Prevention of Meningococcal Disease. *Clinical Infectious Diseases*, 50, pp.184–191.

Cole, J.G. & Jerse, A.E., 2009. Functional Characterization of Antibodies against Neisseria gonorrhoeae Opacity Protein Loops. *Plos One*, 4.

Colijn, C. et al., 2010. What is the mechanism for persistent coexistence of drug-susceptible and drug-resistant strains of Streptococcus pneumoniae? *Journal of the Royal Society, Interface / the Royal Society*, 7, pp.905–919.

Criss, A.K., Kline, K.A. & Seifert, H.S., 2005. The frequency and rate of pilin antigenic variation in Neisseria gonorrhoeae. *Molecular Microbiology*, 58, pp.510–519.

Croucher, N. et al., 2015. Selective and Genetic Constraints on Pneumococcal Serotype Switching. *Plos Genetics*, 11(3): e1005095.

Croucher, N.J. et al., 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature Genetics*, 45, p.656–+.

Croucher, N.J. et al., 2011. Rapid Pneumococcal Evolution in Response to Clinical Interventions. *Science*, 331, pp.430–434.

Dagerhamn, J. et al., 2008. Determination of accessory gene patterns predicts the same relatedness among strains of Streptococcus pneumoniae as sequencing of housekeeping genes does and represents a novel approach in molecular epidemiology. *Journal of Clinical Microbiology*, 46, pp.863–868.

Daugla, D.M. et al., 2014. Effect of a serogroup A meningococcal conjugate vaccine (PsA-TT) on serogroup A meningococcal meningitis and carriage in Chad: A community study. *The Lancet*, 383, pp.40–47.

Davies, R.L., Arkinsaw, S. & Selander, R.K., 1997. Genetic relationships among Pasteurella trehalosi isolates based on multilocus enzyme electrophoresis. *Microbiology-Uk*, 143, pp.2841–2849.

Davies, R.L., Campbell, S. & Whittam, T.S., 2002. Mosaic structure and molecular evolution of the leukotoxin operon (lktCABD) in mannheimia (Pasteurella) haemolytica, Mannheimia glucosida, and Pasteurella trehalosi. *Journal of Bacteriology*, 184, pp.266–277.

Dehio, C., Gray-Owen, S.D. & Meyer, T.F., 1998. The role of neisserial Opa proteins in interactions with host cells. *Trends in Microbiology*, 6, pp.489–495.

Delorme, C. et al., 2007. Extent of horizontal gene transfer in evolution of streptococci of the salivarius group. In *Journal of Bacteriology*. pp. 1330–1341.

Didelot, X. et al., 2011. Recombination and Population Structure in Salmonella enterica. *Plos Genetics*, 7.

Diermayer, M. et al., 1999. Epidemic serogroup B meningococcal disease in Oregon: the evolving epidemiology of the ET-5 strain. *JAMA : the journal of the American Medical Association*, 281, pp.1493–1497.

Donati, C. et al., 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome biology*, 11, p.R107.

Dorigatti, I., Cauchemez, S. & Ferguson, N.M., 2013. Increased transmissibility explains the third wave of infection by the 2009 H1N1 pandemic virus in England. *Proc Natl Acad Sci U S A*, 110, pp.13422–13427.

Echenique-Rivera, H. et al., 2011. Transcriptome analysis of Neisseria meningitidis in human whole blood and mutagenesis studies identify virulence factors involved in blood survival. *PLoS Pathog*, 7, p.e1002027.

Elias, J. et al., 2010. Vaccine preventability of meningococcal clone, Greater Aachen Region, Germany. *Emerging Infectious Diseases*, 16, pp.464–472.

Engvall, E. & Perlmann, P., 1971. Enzyme-Linked Immunosorbent Assay (Elisa) Quantitative Assay Of Immunoglobulin-G. *Immunochemistry*, 8, p.871–&.

Enright, M.C. et al., 2000. Molecular typing of bacteria directly from cerebrospinal fluid. *Eur J Clin Microbiol Infect Dis*, 19, pp.627–630.

Enright, M.C. et al., 2001. Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm type and clone. *Infection and Immunity*, 69, pp.2416–2427.

Enright, M.C. & Spratt, B.G., 1998. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology-Uk*, 144, pp.3049–3060.

Everitt, R.G. et al., 2014. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nature Communications*, 5.

Exley, R.M. et al., 2009. Identification of Meningococcal Genes Necessary for Colonization of Human Upper Airway Tissue. *Infection and Immunity*, 77, pp.45–51.

Facklam, R., 2002. What happened to the streptococci: Overview of taxonomic and nomenclature changes. *Clinical Microbiology Reviews*, 15, p.613–+.

Feavers, I.M. et al., 1996. Antigenic diversity of meningococcal outer membrane protein PorA has implications for epidemiological analysis and vaccine design. *Clinical and diagnostic laboratory immunology*, 3, pp.444–450.

Feikin, D.R. & Klugman, K.P., 2002. Historical changes in pneumococcal serogroup distribution: Implications for the era of pneumococcal conjugate vaccines. *Clinical Infectious Diseases*, 35, pp.547–555.

Feil, E.J. et al., 2004. eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, 186, pp.1518–1530.

Feil, E.J. et al., 2000. Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data. *Genetics*, 154, pp.1439–1450.

Feil, E.J. et al., 2001. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 98, pp.182–187.

Feil, E.J. et al., 1999. The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Molecular biology and evolution*, 16(11), pp.1496–1502.

Ferguson, N.M., Galvani, A.P. & Bush, R.M., 2003. Ecological and immunological determinants of influenza evolution. *Nature*, 422, pp.428–433.

Findlow, H. et al., 2007. Three cases of invasive meningococcal disease caused by a capsule null locus strain circulating among healthy carriers in Burkina Faso. *Journal of Infectious Diseases*, 195, pp.1071–1077.

Flasche, S. et al., 2013. The impact of specific and non-specific immunity on the ecology of Streptococcus pneumoniae and the implications for vaccination. *Proc Biol Sci*, 280, p.20131939.

Fraser, C., Hanage, W.P. & Spratt, B.G., 2005. Neutral microepidemic evolution of bacterial pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 102, pp.1968–1973.

Gandon, S. et al., 2001. Imperfect vaccines and the evolution of pathogen virulence. *Nature*, 414, pp.751–756.

Gause, G.F., 1934. Experimental Analysis Of Vito Volterra's Mathematical Theory Of The Struggle For Existence. *Science (New York, N.Y.)*, 79, pp.16–17.

Gavrilets, S., 2003. Perspective: models of speciation: what have we learned in 40 years? *Evolution; international journal of organic evolution*, 57, pp.2197–2215.

Gherardi, G. et al., 2012. Serotype and Clonal Evolution of Penicillin-Nonsusceptible Invasive Streptococcus pneumoniae in the 7-Valent Pneumococcal Conjugate Vaccine Era in Italy. *Antimicrobial Agents and Chemotherapy*, 56, pp.4965–4968.

Gilchrist, M.A. & Sasaki, A., 2002. Modeling host-parasite coevolution: A nested approach based on mechanistic models. *Journal of Theoretical Biology*, 218, pp.289–308.

Ginsburg, A.S. et al., 2012. Issues and challenges in the development of pneumococcal protein vaccines. *Expert Review of Vaccines*, 11, pp.279–285.

Gog, J.R. & Grenfell, B.T., 2002. Dynamics and selection of many-strain pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 99, pp.17209–17214.

Goldblatt, D. et al., 2005. Antibody responses to nasopharyngeal carriage of Streptococcus pneumoniae in adults: A longitudinal household study. *Journal of Infectious Diseases*, 192, pp.387–393.

Golubchik, T. et al., 2012. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nature Genetics*, 44, pp.352–355.

Gomes, M.G.M., Medley, G.F. & Nokes, D.J., 2002. On the determinants of population structure in antigenically diverse pathogens. *Proceedings of the Royal Society B-Biological Sciences*, 269, pp.227–233.

Gordhan, B.G. et al., 2002. Construction and phenotypic characterization of an auxotrophic mutant of Mycobacterium tuberculosis defective in L-arginine biosynthesis. *Infection and Immunity*, 70(6), pp.3080–3084.

Gossger, N. et al., 2012. Immunogenicity and Tolerability of Recombinant Serogroup B Meningococcal Vaccine Administered With or Without Routine Infant Vaccinations According to Different Immunization Schedules A Randomized Controlled Trial. *Jama-Journal of the American Medical Association*, 307, pp.573–582.

Grad, Y.H. et al., 2012. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences*, 109, pp.3065–3070.

Greenwood, B., 1999. Meningococcal meningitis in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93, pp.341–353.

Gupta, S. et al., 1996. The maintenance of strain structure in populations of recombining infectious agents. *Nature Medicine*, 2, pp.437–442.

Gupta, S., Ferguson, N. & Anderson, R., 1998. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science*, 280, pp.912–915.

Gupta, S., Ferguson, N.M. & Anderson, R.M., 1997. Vaccination and the population structure of antigenically diverse pathogens that exchange genetic material. *Proc Biol Sci*, 264, pp.1435–1443.

Gupta, S. & Galvani, A., 1999. The effects of host heterogeneity on pathogen population structure. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 354, pp.711–719.

Gupta, S., Swinton, J. & Anderson, R.M., 1994a. Theoretical studies of the effects of heterogeneity in the parasite population on the transmission dynamics of malaria. *Proceedings. Biological sciences / The Royal Society*, 256, pp.231–238.

Gupta, S., Swinton, J. & Anderson, R.M., 1994b. THeoretical-Studies Of The Effects Of Heterogeneity In The Parasite Population On The Transmission Dynamics Of Malaria. *Proceedings of the Royal Society B-Biological Sciences*, 256, pp.231–238.

H, S., 1903. Die Artunterscheidung der fur den menschen Pathogen Streptokokken durch Blutagar. *Munch. Med. Wochenschr*, 50, pp.849–853.

Hagblom, P. et al., 1985. Intragenic Recombination Leads To Pilus Antigenic Variation In Neisseria-Gonorrhoeae. *Nature*, 315, pp.156–158.

Halperin, S.A. et al., 2012. The changing and dynamic epidemiology of meningococcal disease. *Vaccine*, 30, pp.B26–B36.

Hanage, W.P. et al., 2004. Ability of pneumococcal serotypes and clones to cause acute otitis media: Implications for the prevention of otitis media by conjugate vaccines. *Infection and Immunity*, 72, pp.76–81.

Hanage, W.P. et al., 2011. Clonal replacement among 19A Streptococcus pneumoniae in Massachusetts, prior to 13 valent conjugate vaccination. *Vaccine*, 29, pp.8877–8881.

Hanage, W.P. et al., 2010. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics*, 2, pp.80–84.

Hanage, W.P. et al., 2005. Invasiveness of serotypes and clones of Streptococcus pneumoniae among children in Finland. *Infection and Immunity*, 73, pp.431–435.

Harrison, L.H. et al., 2006. Antigenic shift and increased incidence of meningococcal disease. *The Journal of infectious diseases*, 193, pp.1266–1274.

Harrison, O.B. et al., 2013. Description and nomenclature of Neisseria meningitidis capsule locus. *Emerging Infectious Diseases*, 19, pp.566–573.

Harrison, O.B. et al., 2009. Epidemiological Evidence for the Role of the Hemoglobin Receptor, HmbR, in Meningococcal Virulence. *Journal of Infectious Diseases*, 200, pp.94–98.

Hausdorff, W.P. et al., 2000. Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I. *Clin Infect Dis*, 30, pp.100–121.

Hausdorff, W.P., Hoet, B. & Schuerman, L., 2010. Do pneumococcal conjugate vaccines provide any cross-protection against serotype 19A? *BMC pediatrics*, 10, p.4.

Hava, D. & Camilli, A., 2002. Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors. *Molecular Microbiology*, 45, pp.1389–1405.

Hedges, S.R. et al., 1999. Limited local and systemic antibody responses to Neisseria gonorrhoeae during uncomplicated genital infections. *Infection and Immunity*, 67, pp.3937–3946.

Hedman, Å.K. et al., 2012. Transcriptional profiling of serogroup B Neisseria meningitidis growing in human blood: an approach to vaccine antigen discovery. *PLoS One*, 7, p.e39718.

Helm, R.A. & Seifert, H.S., 2010. Frequency and Rate of Pilin Antigenic Variation of Neisseria meningitidis. *Journal of Bacteriology*, 192, pp.3822–3823.

Henrichsen, J., 1995. Six newly recognized types of Streptococcus pneumoniae. *J Clin Microbiol*, 33, pp.2759–2762.

Hethcote, H.W., Yorke, J.A. & Nold, A., 1982. Gonorrhea Modeling - A Comparison Of Control Methods. *Mathematical Biosciences*, 58, pp.93–109.

Hey, A., 2013. Transcriptional profiling of Neisseria meningitidis interacting with human epithelial cells in a long-term in vitro colonization model . *Infection & Immunity*, 81, pp.4149–4159.

Higgins, M.A. et al., 2009. Differential Recognition and Hydrolysis of Host Carbohydrate Antigens by Streptococcus pneumoniae Family 98 Glycoside Hydrolases. *Journal of Biological Chemistry*, 284, pp.26161–26173.

Hill, P.C. et al., 2008. Nasopharyngeal carriage of Streptococcus pneumoniae in Gambian infants: A longitudinal study. *Clinical Infectious Diseases*, 46, pp.807–814.

Hobbs, M.M. et al., 1994. Microevolution Within A Clonal Population Of Pathogenic Bacteria - Recombination, Gene Duplication And Horizontal Genetic Exchange In The Opa Gene Family Of Neisseria-Meningitidis. *Molecular Microbiology*, 12, pp.171–180.

Hogberg, L. et al., 2007. Age- and serogroup-related differences in observed durations of nasopharyngeal carriage of penicillin-resistant pneumococci. *Journal of Clinical Microbiology*, 45, pp.948–952.

Holmes, A.R. et al., 2001. The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence. *Molecular Microbiology*, 41, pp.1395–1408.

Holmes, E.C., Urwin, R. & Maiden, M.C.J., 1999. The influence of recombination on the population structure and evolution of the human pathogen Neisseria meningitidis. *Molecular Biology and Evolution*, 16, pp.741–749.

Holst, J. et al., 2003. Serum bactericidal activity correlates with the vaccine efficacy of outer membrane vesicle vaccines against Neisseria meningitidis serogroup B disease. *Vaccine*, 21, pp.734–737.

Hsieh, Y.C. et al., 2009. Characterization of invasive isolates of Streptococcus pneumoniae among Taiwanese children. *Clinical Microbiology and Infection*, 15, pp.991–996.

Hu, F.Z. et al., 2012. In vivo capsular switch in Streptococcus pneumoniae--analysis by whole genome sequencing. *PLoS One*, 7, p.e47983.

Hubbell, S.P., 2001. *The Unified Neutral Theory of Biodiversity and Biogeography*,

Huber, C.A. et al., 2012. Lack of antigenic diversification of major outer membrane proteins during clonal waves of Neisseria meningitidis serogroup A colonization and disease. *FEMS Immunol Med Microbiol*.

Hyams, C. et al., 2011. Effects of Streptococcus pneumoniae Strain Background on Complement Resistance. *Plos One*, 6.

Hyams, C. et al., 2013. Streptococcus pneumoniae Capsular Serotype Invasiveness Correlates with the Degree of Factor H Binding and Opsonization with C3b/iC3b. *Infection and Immunity*, 81, pp.354–363.

Hyams, C. et al., 2010. The Streptococcus pneumoniae Capsule Inhibits Complement Activity and Neutrophil Phagocytosis by Multiple Mechanisms. *Infection and Immunity*, 78, pp.704–715.

Hyatt, D. et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, p.119.

Ihekweazu, C.A. et al., 2008. Trends in incidence of pneumococcal disease before introduction of conjugate vaccine: South West England, 1996-2005. *Epidemiol Infect*, 136, pp.1096–1102.

Iyer, R., Baliga, N.S. & Camilli, A., 2005. Catabolite control protein A (CcpA) contributes to virulence and regulation of sugar metabolism in Streptococcus pneumoniae. *Journal of Bacteriology*, 187, pp.8340–8349.

James, J.F. & Swanson, J., 1978. Studies On Gonococcus Infection .13. Occurrence Of Color-Opacity Colonial Variants In Clinical Cultures. *Infection and Immunity*, 19, pp.332–340.

Jefferies, J.M. et al., 2010. Temporal analysis of invasive pneumococcal clones from Scotland illustrates fluctuations in diversity of serotype and genotype in the absence of pneumococcal conjugate vaccine. *J Clin Microbiol*, 48, pp.87–96.

Jerse, A.E. et al., 1994. Multiple Gonococcal Opacity Proteins Are Expressed During Experimental Urethral Infection In The Male. *Journal of Experimental Medicine*, 179, pp.911–920.

Jodar, L. et al., 2002. Development of vaccines against meningococcal disease. *Lancet*, 359, pp.1499–1508.

Johnson, D.R. et al., 2006. Characterization of group A streptococci (Streptococcus pyogenes): correlation of M-protein and emm-gene type with T-protein agglutination pattern and serum opacity factor. *Journal of Medical Microbiology*, 55, pp.157–164.

Johnson, H.L. et al., 2010. Systematic Evaluation of Serotypes Causing Invasive Pneumococcal Disease among Children Under Five: The Pneumococcal Global Serotype Project. *Plos Medicine*, 7.

Johswich, K.O. et al., 2013. In Vivo Adaptation and Persistence of Neisseria meningitidis within the Nasopharyngeal Mucosa. *Plos Pathogens*, 9.

Jolley, K.A. et al., 2000. Carried meningococci in the Czech Republic: a diverse recombining population. *J Clin Microbiol*, 38, pp.4492–4498.

Jolley, K.A. et al., 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology-Sgm*, 158, pp.1005–1015.

Jolley, K.A. et al., 2005. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Molecular Biology and Evolution*, 22, pp.562–569.

Jolley, K.A., Brehony, C. & Maiden, M.C.J., 2007. Molecular typing of meningococci: recommendations for target choice and nomenclature. *Fems Microbiology Reviews*, 31, pp.89–96.

Jolley, K.A., Chan, M.-S. & Maiden, M.C.J., 2004. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC bioinformatics*, 5, p.86.

Jolley, K.A. & Maiden, M.C.J., 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *Bmc Bioinformatics*, 11.

Jones, N. et al., 2003. Multilocus sequence typing system for group B streptococcus. *Journal of Clinical Microbiology*, 41(6), pp.2530–2536.

De Jonge, M.I. et al., 2003. Functional activity of antibodies against the recombinant OpaJ protein from Neisseria meningitidis. *Infection and Immunity*, 71, pp.2331–2340.

Joseph, B. et al., 2011. Virulence Evolution of the Human Pathogen Neisseria meningitidis by Recombination in the Core and Accessory Genome. *Plos One*, 6.

Kahler, C.M. & Stephens, D.S., 1998. Genetic basis for biosynthesis, structure, and function of meningococcal lipooligosaccharide (Endotoxin). *Critical Reviews in Microbiology*, 24, pp.281–334.

Kawamura, Y. et al., 1995. Determination Of 16s Ribosomal-Rna Sequences Of Streptococcus-Mitis And Streptococcus-Gordonii And Phylogenetic-Relationships Among Members Of The Genus Streptococcus. *International Journal of Systematic Bacteriology*, 45, pp.406–408.

Kelly, D.F. et al., 2006. CRM197-conjugated serogroup C meningococcal capsular polysaccharide, but not the native polysaccharide, induces persistent antigen-specific memory B cells. *Blood*, 108, pp.2642–2647.

Kelly, T., Dillard, J.P. & Yother, J., 1994. Effect Of Genetic Switching Of Capsular Type On Virulence Of Streptococcus-Pneumoniae. *Infection and Immunity*, 62, pp.1813–1819.

Kermack, W.O. & McKendrick, A.G., 1991. Contributions To The Mathematical-Theory Of Epidemics .1. (Reprinted From Proceedings Of The Royal Society, VOL 115A, PG 700-721, 1927). *Bulletin of Mathematical Biology*, 53, pp.33–55.

King, S.J., Hippe, K.R. & Weiser, J.N., 2006. Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by Streptococcus pneumoniae. *Molecular Microbiology*, 59, pp.961–974.

Kong, F. & Gilbert, G.L., 2003. Using cpsA-cpsB sequence polymorphisms and serotype-/group-specific PCR to predict 51 Streptococcus pneumoniae capsular serotypes. *Journal of Medical Microbiology*, 52(12), pp.1047–1058.

Koonin, E. V & Wolf, Y.I., 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36, pp.6688–6719.

Koumare, B. et al., 2007. The first large epidemic of meningococcal disease caused by serogroup W135, Burkina Faso, 2002. *Vaccine*, 25, pp.A37–A41.

Kristiansen, P.A. et al., 2014. Persistent low carriage of serogroup A Neisseria meningitidis two years after mass vaccination with the meningococcal conjugate vaccine, MenAfriVac. *BMC Infect Dis*, 14, p.663.

Krizova, P. & Musilek, M., 1995. Changing epidemiology of meningococcal invasive disease in the Czech republic caused by new clone Neisseria meningitidis C:2a:P1.2(P1.5), ET-15/37. *Cent Eur J Public Health*, 3, pp.189–194.

Lalitha, M.K. et al., 1996. Serotyping of Streptococcus pneumoniae by agglutination assays: A cost-effective technique for developing countries. *Bulletin of the World Health Organization*, 74(4), pp.387–390.

Lamelas, A. et al., 2014. Emergence of a new epidemic Neisseria meningitidis serogroup A Clone in the African meningitis belt: high-resolution picture of genomic changes that mediate immune evasion. *MBio*, 5, pp.e01974–14.

Lange, A. & Ferguson, N.M., 2009. Antigenic Diversity, Transmission Mechanisms, and the Evolution of Pathogens. *Plos Computational Biology*, 5.

Lau, G.W. et al., 2001. A functional genomic analysis of type 3 Streptococcus pneumoniae virulence. *Molecular Microbiology*, 40, pp.555–571.

Lee, C.A., 1996. Pathogenicity islands and the evolution of bacterial pathogens. *Infectious agents and disease*, 5, pp.1–7.

Lefébure, T. & Stanhope, M.J., 2007. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome biology*, 8, p.R71.

Lehmann, D. et al., 2010. The Changing Epidemiology of Invasive Pneumococcal Disease in Aboriginal and Non-Aboriginal Western Australians from 1997 through 2007 and Emergence of Nonvaccine Serotypes. *Clinical Infectious Diseases*, 50, pp.1477–1486.

Leimkugel, J. et al., 2007. Clonal waves of Neisseria colonisation and disease in the African meningitis belt: Eight-year longitudinal study in northern Ghana. *PLoS Medicine*, 4, pp.535–544.

Levin, B.R., 1981. Periodic Selection, Infectious Gene Exchange And The Genetic-Structure Of Escherichia-Coli Populations. *Genetics*, 99, pp.1–23.

Levin, S. & Pimentel, D., 1981. Selection Of Intermediate Rates Of Increase In Parasite-Host Systems. *American Naturalist*, 117, pp.308–315.

Van der Ley, P. et al., 1991. Topology of outer membrane porins in pathogenic Neisseria spp. *Infection and Immunity*, 59(9), pp.2963–2971.

Linke, C.M. et al., 2013. The ABC Transporter Encoded at the Pneumococcal Fructooligosaccharide Utilization Locus Determines the Ability To Utilize Long- and Short-Chain Fructooligosaccharides. *Journal of Bacteriology*, 195, pp.1031–1041.

Lipsitch, M. et al., 2000. Competition among Streptococcus pneumoniae for intranasal colonization in a mouse model. *Vaccine*, 18, pp.2895–2901.

Lipsitch, M. et al., 2012. Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model. *Epidemiology*, 23, pp.510–519.

Lipsitch, M., 2001. Interpreting results from trials of pneumococcal conjugate vaccines: A statistical test for detecting vaccine-induced increases in carriage of nonvaccine serotypes. *American Journal of Epidemiology*, 154, pp.85–92.

Lipsitch, M. et al., 2009. No coexistence for free: Neutral null models for multistrain pathogens. *Epidemics*, 1, pp.2–13.

Lipsitch, M., 1997. Vaccination against colonizing bacteria with multiple serotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 94, pp.6571–6576.

Lipsitch, M. & O'Hagan, J.J., 2007. Patterns of antigenic diversity and the mechanisms that maintain them. *Journal of the Royal Society Interface*, 4, pp.787–802.

Lock, R.A. et al., 1996. Sequence variation in the Streptococcus pneumoniae pneumolysin gene affecting haemolytic activity and electrophoretic mobility of the toxin. *Microbial Pathogenesis*, 21, pp.71–83.

Loman, N.J. et al., 2013. Clonal Expansion within Pneumococcal Serotype 6C after Use of Seven-Valent Vaccine. *PLoS ONE*, 8.

Lotka, A.J. 1925. *Elements of Physical Biology*, Baltimore: Williams and Wilkins.

Macarthur, R. & Levins, R., 1967. The Limiting Similarity, Convergence, and Divergence of Coexisting Species. *The American Naturalist*, 101, p.377.

Maiden, M.C. et al., 2008. Impact of meningococcal serogroup C conjugate vaccines on carriage and herd immunity. *J Infect Dis*, 197, pp.737–743.

Maiden, M.C., 2008. Population genomics: diversity and virulence in the Neisseria. *Curr Opin Microbiol*, 11, pp.467–471.

Maiden, M.C.J. et al., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology*, 11, pp.728–736.

Maiden, M.C.J., 2006. Multilocus sequence typing of bacteria. *Annual review of microbiology*, 60, pp.561–588.

Maiden, M.C.J. et al., 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95, pp.3140–3145.

Malorny, B. et al., 1998. Sequence diversity, predicted two-dimensional protein structure, and epitope mapping of neisserial Opa proteins. *Journal of Bacteriology*, 180, pp.1323–1330.

Manco, S. et al., 2006. Pneumococcal neuraminidases A and B both have essential roles during infection of the respiratory tract and sepsis. *Infection and Immunity*, 74, pp.4014–4020.

Marsh, J.W. et al., 2007. Deletion of fetA gene sequences in serogroup B and C Neisseria meningitidis isolates. *Journal of Clinical Microbiology*, 45, pp.1333–1335.

Marzoa, J. et al., 2012. Induction of immune responses by purified outer membrane protein complexes from Neisseria meningitidis. *Vaccine*, 30, pp.2387–2395.

Maslow, J.N. et al., 1995. Clonal Relationships Among Blood-Stream Isolates Of Escherichia-Coli. *Infection and Immunity*, 63, pp.2409–2417.

McGregor, K.F. et al., 2004. Multilocus sequence typing of Streptococcus pyogenes representing most known emm types and distinctions among subpopulation genetic structures. *Journal of Bacteriology*, 186, pp.4285–4294.

McIntosh, E.D.G. & Reinert, R.R., 2011. Global prevailing and emerging pediatric pneumococcal serotypes. *Expert Review of Vaccines*, 10, pp.109–129.

McKane, R.B. et al., 2002. Resource-based niches provide a basis for plant species diversity and dominance in arctic tundra. *Nature*, 415, pp.68–71.

McLean, A.R., 1995. Vaccination, evolution and changes in the efficacy of vaccines: a theoretical framework. *Proceedings. Biological sciences / The Royal Society*, 261, pp.389–393.

McMichael, A.J. & Phillips, R.E., 1997. Escape of human immunodeficiency virus from immune control. *Annual Review of Immunology*, 15, pp.271–296.

Meats, E. et al., 2003. Characterization of encapsulated and noncapsulated Haemophilus influenzae and determination of phylogenetic relationships by multilocus sequence typing. *Journal of Clinical Microbiology*, 41, pp.1623–1636.

Mell, J.C. et al., 2011. Transformation of Natural Genetic Variation into Haemophilus Influenzae Genomes. *Plos Pathogens*, 7.

Mitchell, A.M. & Mitchell, T.J., 2010. Streptococcus pneumoniae: Virulence factors and variation. *Clinical Microbiology and Infection*, 16(5), pp.411–418.

Mitchinson, W., 2013. *Body Failure: Medical Views of Women*, University of Toronto Press.

Morelli, G. et al., 1997. Clonal descent and microevolution of Neisseria meningitidis during 30 years of epidemic spread. *Molecular Microbiology*, 25, pp.1047–1064.

Mueller, J.E. et al., 2007. Molecular characteristics and epidemiology of meningococcal carriage, Burkina Faso, 2003. *Emerging Infectious Diseases*, 13, pp.847–854.

Mueller, J.E. et al., 2006. Neisseria meningitidis serogroups A and W-135: carriage and immunity in Burkina Faso, 2003. *The Journal of infectious diseases*, 193, pp.812–820.

Muenzner, P. et al., 2005. CEACAM engagement by human pathogens enhances cell adhesion and counteracts bacteria-induced detachment of epithelial cells. *Journal of Cell Biology*, 170, pp.825–836.

Munoz, R. et al., 1992. Geographic-Distribution Of Penicillin-Resistant Clones Of Streptococcus-Pneumoniae - Characterization By Penicillin-Binding Protein Profile, Surface Protein-A Typing, And Multilocus Enzyme Analysis. *Clinical Infectious Diseases*, 15, pp.112–118.

Munoz-Almagro, C. et al., 2008. Emergence of invasive pneumococcal disease caused by nonvaccine serotypes in the era of 7-valent conjugate vaccine. *Clinical Infectious Diseases*, 46, pp.174–182.

Murphy, E. et al., 2011. Challenges for the evaluation of Staphylococcus aureus protein based vaccines Monitoring antigenic diversity. *Human Vaccines*, 7, pp.51–59.

Musser, J.M. et al., 1988. Evolutionary Genetics Of The Encapsulated Strains Of Hemophilus-Influenzae. *Proceedings of the National Academy of Sciences of the United States of America*, 85, pp.7758–7762.

Musser, J.M. et al., 1989. Identification of a high-virulence clone of type III Streptococcus agalactiae (group B Streptococcus) causing invasive neonatal disease. *Proceedings of the National Academy of Sciences of the United States of America*, 86, pp.4731–4735.

Nebenzahl, Y.M. et al., 2004. Virulence of Streptococcus pneumoniae may be determined independently of capsular polysaccharide. *Fems Microbiology Letters*, 233, pp.147–152.

Neufeld, F. & Handel, L., 1909. Über die Herstellung von Pneumokokkenserum und über die Aussichten einer spezifischen Behandlung der Pneumonie. *Z Immunitätsforsch*, III, pp.159–171.

Nicolas, P. et al., 2005. Molecular epidemiology of neisseria meningitidis isolated in the African Meningitis Belt between 1988 and 2003 shows dominance of sequence type 5 (ST-5) and ST-11 complexes. *J Clin Microbiol*, 43, pp.5129–5135.

Nieto, P.A. et al., 2013. Gene Elements that Regulate Streptococcus pneumoniae Virulence and Immunity Evasion. *Current Gene Therapy*, 13, pp.51–64.

Noble, R.C. et al., 1977. Recidivism Among Patients With Gonococcal Infection Presenting To A Venereal-Disease Clinic. *Sexually Transmitted Diseases*, 4, pp.39–43.

Noirot, P. & Noirot-Gros, M.-F., 2004. Protein interaction networks in bacteria. *Current opinion in microbiology*, 7, pp.505–512.

Norheim, G. et al., 2006. Characterization of Neisseria meningitidis isolates from recent outbreaks in Ethiopia and comparison with those recovered during the epidemic of 1988 to 1989. *J Clin Microbiol*, 44, pp.861–871.

Norris, C., 1913. *Gonorrhea in women : its pathology, symptomatology, diagnosis, and treatment ; together with a review of the rare varieties of the disease which occur in men, women and children*, Philadelphia & London, WB Saunders Company.

Nowak, M.A., May, R.M. & Anderson, R.M., 1990. The Evolutionary Dynamics Of Hiv-1 Quasispecies And The Development Of Immunodeficiency Disease. *Aids*, 4, pp.1095–1103.

Ochman, H. & Selander, R.K., 1984. Evidence For Clonal Population-Structure In Escherichia-Coli. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 81, pp.198–201.

Olyhoek, T., Crowe, B.A. & Achtman, M., 1987. Clonal Population-Structure Of Neisseria-Meningitidis Serogroup-A Isolated From Epidemics And Pandemics Between 1915 And 1983. *Reviews of Infectious Diseases*, 9, pp.665–692.

Organisation, W.H., 2012. Global incidence and prevalence of selected curable sexually transmitted infections – 2008.

Orihuela, C.J. et al., 2004. Microarray analysis of pneumococcal gene expression during invasive disease. *Infection and Immunity*, 72, pp.5582–5596.

Orourke, M. & Stevens, E., 1993. Genetic-Structure Of Neisseria-Gonorrhoeae Populations - A Non-Clonal Pathogen. *Journal of General Microbiology*, 139, pp.2603–2611.

Oster, P. et al., 2005. MeNZB: A safe and highly immunogenic tailor-made vaccine against the New Zealand Neisseria meningitidis serogroup B disease epidemic strain. In *Vaccine*. pp. 2191–2196.

Ota, M.O. et al., 2011. The immunogenicity and impact on nasopharyngeal carriage of fewer doses of conjugate pneumococcal vaccine immunization schedule. *Vaccine*, 29, pp.2999–3007.

Pagliarulo, C. et al., 2004. Regulation and differential expression of gdhA encoding NADP-specific glutamate dehydrogenase in Neisseria meningitidis clinical isolates. *Mol Microbiol*, 51, pp.1757–1772.

Pai, R. et al., 2005. Postvaccine genetic structure of Streptococcus pneumoniae serotype 19A from children in the United States. *The Journal of infectious diseases*, 192, pp.1988–1995.

Pantelic, M. et al., 2005. Neisseria gonorrhoeae kills carcinoembryonic antigen-related cellular adhesion molecule 1 (CD66a)-expressing human B cells and inhibits antibody production. *Infection and Immunity*, 73, pp.4171–4179.

Pareek, C.S., Smoczynski, R. & Tretyn, A., 2011. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52, pp.413–435.

Parra, E.L. et al., 2014. Serotype and Genotype Distribution among Invasive Streptococcus pneumoniae Isolates in Colombia, 2005-2010. *Plos One*, 9.

Pettigrew, M.M. et al., 2006. Variation in the presence of neuraminidase genes among Streptococcus pneumoniae isolates with identical sequence types. *Infection and Immunity*, 74, pp.3360–3365.

Philips, B.J. et al., 2003. Factors determining the appearance of glucose in upper and lower respiratory tract secretions. *Intensive Care Medicine*, 29, pp.2204–2210.

Pichon, B. et al., 2013. Changes in Molecular Epidemiology of Streptococcus pneumoniae Causing Meningitis following Introduction of Pneumococcal Conjugate Vaccination in England and Wales. *Journal of Clinical Microbiology*, 51, pp.820–827.

Piet, J.R. et al., 2014. Streptococcus pneumoniae Arginine Synthesis Genes Promote Growth and Virulence in Pneumococcal Meningitis. *Journal of Infectious Diseases*, 209, pp.1781–1791.

Pillai, D.R. et al., 2009. Genome-wide dissection of globally emergent multi-drug resistant serotype 19A Streptococcus pneumoniae. *Bmc Genomics*, 10.

Pizarro-Cerdá, J. & Cossart, P., 2006. Bacterial adhesion and entry into host cells. *Cell*, 124, pp.715–727.

Pizza, M. et al., 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, 287, pp.1816–1820.

Van der Ploeg, L.H. et al., 1984. Antigenic variation in Trypanosoma brucei analyzed by electrophoretic separation of chromosome-sized DNA molecules. *Cell*, 37, pp.77–84.

Plummer, F.A. et al., 1994. Antibodies To Opacity Proteins (Opa) Correlate With A Reduced Risk Of Gonococcal Salpingitis. *Journal of Clinical Investigation*, 93, pp.1748–1755.

Polissi, A. et al., 1998. Large-scale identification of virulence genes from Streptococcus pneumoniae. *Infection and Immunity*, 66, pp.5620–5629.

Pollard, A.J., Perrett, K.P. & Beverley, P.C., 2009. Maintaining protection against invasive bacteria with protein-polysaccharide conjugate vaccines. *Nature reviews. Immunology*, 9, pp.213–220.

Porco, T.C. & Blower, S.M., 1998. Designing HIV vaccination policies: Subtypes and cross-immunity. *Interfaces*, 28, pp.167–190.

Read, R.C. et al., 2014. Effect of a quadrivalent meningococcal ACWY glycoconjugate or a serogroup B meningococcal vaccine on meningococcal carriage: an observer-blind, phase 3 randomised clinical trial. *Lancet*, 384, pp.2123–2131.

Recker, M. et al., 2004. Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature*, 429, pp.555–558.

Regev-Yochay, G. et al., 2010. Re-emergence of the type 1 pilus among Streptococcus pneumoniae isolates in Massachusetts, USA. *Vaccine*, 28, pp.4842–4846.

Regev-Yochay, G. et al., 2009. The pneumococcal pilus predicts the absence of Staphylococcus aureus co-colonization in pneumococcal carriers. *Clin Infect Dis*, 48, pp.760–763.

Ren, B. et al., 2003. Both family 1 and family 2 PspA proteins can inhibit complement deposition and confer virulence to a capsular serotype 3 strain of Streptococcus pneumoniae. *Infection and Immunity*, 71, pp.75–85.

Restif, O. & Grenfell, B.T., 2007. Vaccination and the dynamics of immune evasion. *Journal of the Royal Society, Interface / the Royal Society*, 4, pp.143–153.

Ribeiro, G.S. et al., 2003. Prevention of Haemophilus influenzae type b (Hib) meningitis and emergence of serotype replacement with type a strains after introduction of Hib immunization in Brazil. *The Journal of infectious diseases*, 187, pp.109–116.

Rinta-Kokko, H. et al., 2009. Estimation of vaccine efficacy against acquisition of pneumococcal carriage. *Vaccine*, 27, pp.3831–3837.

Roberts, L., 2008. An ill wind, bringing meningitis. *Science*, 320, pp.1710–1715.

Robertson, J.N., Vincent, P. & Ward, M.E., 1977. Preparation And Properties Of Gonococcal Pili. *Journal of General Microbiology*, 102, pp.169–177.

Robinson, D.A. et al., 2001. Clones of Streptococcus pneumoniae isolated from nasopharyngeal carriage and invasive disease in young children in central Tennessee. *Journal of Infectious Diseases*, 183, pp.1501–1507.

Ronald, A., 1996. Molecular and cell biology of sexually transmitted diseases - Wright,D, Archard,L. *Social Science & Medicine*, 42, p.784.

Rosenqvist, E. et al., 1995. Human antibody responses to meningococcal outer membrane antigens after three doses of the Norwegian group B meningococcal vaccine. *Infection and Immunity*, 63, pp.4642–4652.

Ross, J.D.C., 1996. Systemic gonococcal infection. *Genitourinary Medicine*, 72, pp.404–407.

Rouaud, P. et al., 2006. Prolonged outbreak of B meningococcal disease in the Seine-Maritime department, France, January 2003 to June 2005. *Euro surveillance: European communicable disease bulletin*, 11, pp.178–181.

Russell, J.E. et al., 2008. Molecular epidemiology of meningococcal disease in England and Wales 1975-1995, before the introduction of serogroup C conjugate vaccines. *Microbiology*, 154, pp.1170–1177.

Russell, J.E. et al., 2004. PorA Variable Regions of Neisseria meningitidis. *Emerging Infectious Diseases*, 10(4), pp.674–678.

Saleem, M. et al., 2013. Use of a Molecular Decoy to Segregate Transport from Antigenicity in the FrpB Iron Transporter from Neisseria meningitidis. *PLoS ONE*, 8.

Sandgren, A. et al., 2004. Effect of clonal and serotype-specific properties on the invasive capacity of Streptococcus pneumoniae. *Journal of Infectious Diseases*, 189, pp.785–796.

Satzke, C. et al., 2013. Standard method for detecting upper respiratory carriage of Streptococcus pneumoniae: Updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. *Vaccine*, 32, pp.165–179.

Schoen, C. et al., 2014. Metabolism and virulence in Neisseria meningitidis. *Frontiers in Cellular and Infection Microbiology*, 4.

Schwarz-Linek, U., Höök, M. & Potts, J.R., 2004. The molecular basis of fibronectin-mediated bacterial adherence to host cells. *Molecular Microbiology*, 52, pp.631–641.

Seifert, H.S. et al., 1994. Multiple Gonococcal Pilin Antigenic Variants Are Produced During Experimental Human Infections. *Journal of Clinical Investigation*, 93, pp.2744–2749.

Selander, R.K. & Levin, B.R., 1980. Genetic Diversity And Structure In Escherichia-Coli Populations. *Science*, 210, pp.545–547.

Shannon, C.E., 1948. A Mathematical Theory Of Communication. *Bell System Technical Journal*, 27, pp.379–423.

Sharma, D. et al., 2013. Pneumococcal Carriage and Invasive Disease in Children Before Introduction of the 13-valent Conjugate Vaccine: Comparison With the Era Before 7-valent Conjugate Vaccine. *Pediatric Infectious Disease Journal*, 32, pp.E45–E53.

Shelburne, S.A. et al., 2008. The role of complex carbohydrate catabolism in the pathogenesis of invasive streptococci. *Trends in Microbiology*, 16, pp.318–325.

Shendure, J. & Ji, H.L., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26, pp.1135–1145.

Sié, A. et al., 2008. ST2859 serogroup A meningococcal meningitis outbreak in Nouna Health District, Burkina Faso: a prospective study. *Trop Med Int Health*, 13, pp.861–868.

Sierra, G. V et al., 1991. *Vaccine against group B Neisseria meningitidis: protection trial and mass vaccination results in Cuba.*,

Simões, A.S. et al., 2011. Clonal evolution leading to maintenance of antibiotic resistance rates among colonizing pneumococci in the PCV7 era in Portugal. *Journal of Clinical Microbiology*, 49, pp.2810–2817.

Sleeman, K. et al., 2001. Invasive pneumococcal disease in England and Wales: Vaccination implications. *Journal of Infectious Diseases*, 183, pp.239–246.

Sleeman, K.L. et al., 2006. Capsular serotype-specific attack rates and duration of carriage of Streptococcus pneumoniae in a population of children. *Journal of Infectious Diseases*, 194, pp.682–688.

Smith, J.M. et al., 1993. How Clonal Are Bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 90, pp.4384–4388.

Smith, J.M., Dowson, C.G. & Spratt, B.G., 1991. Localized Sex In Bacteria. *Nature*, 349, pp.29–31.

Snape, M.D. et al., 2013. Persistence of bactericidal antibodies following early infant vaccination with a serogroup B meningococcal vaccine and immunogenicity of a preschool booster dose. *CMAJ*, 185.

Sreevatsan, S. et al., 1997. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, 94, pp.9869–9874.

Stern, A. et al., 1986. Opacity Genes In Neisseria-Gonorrhoeae - Control Of Phase And Antigenic Variation. *Cell*, 47, pp.61–71.

Suerbaum, S. et al., 1998. Free recombination within Helicobacter pylori. *Proceedings of the National Academy of Sciences of the United States of America*, 95, pp.12619–12624.

Sun, Y.H. et al., 2000. Functional genomics of Neisseria meningitidis pathogenesis. *Nature Medicine*, 6, pp.1269–1273.

Swanson, J. et al., 1988. Expression Of Outer-Membrane Protein-Ii By Gonococci In Experimental Gonorrhea. *Journal of Experimental Medicine*, 168, pp.2121–2129.

Swartz, M.N., 2004. Bacterial meningitis - A view of the past 90 years. *New England Journal of Medicine*, 351, pp.1826–1828.

Tanaka, J. et al., 2012. Incidence of childhood pneumonia and serotype and sequence-type distribution in Streptococcus pneumoniae isolates in Japan. *Epidemiology and Infection*, 140, pp.1111–1121.

Tavares, A. et al., 2014. Population structure of methicillin-susceptible Staphylococcus aureus (MSSA) in Portugal over a 19-year period (1992-2011). *European Journal of Clinical Microbiology & Infectious Diseases*, 33, pp.423–432.

Tettelin, H. et al., 2001. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. *Science*, 293, pp.498–506.

Tettelin, H. et al., 2000. Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. *Science*, 287, pp.1809–1815.

Thirumoorthi, M.C. & Dajani, A.S., 1979. Comparison Of Staphylococcal Coagglutination, Latex Agglutination, And Counterimmunoelectrophoresis For Bacterial-Antigen Detection. *Journal of Clinical Microbiology*, 9, pp.28–32.

Thompson, E.A.L., Feavers, I.M. & Maiden, M.C.J., 2003. Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component. *Microbiology-Sgm*, 149, pp.1849–1858.

Thompson, M.J. et al., 2006. Clinical recognition of meningococcal disease in children and adolescents. *Lancet*, 367, pp.397–403.

Tikhomirov, E., Santamaria, M. & Esteves, K., 1997. Meningococcal disease: Public health burden and control. *World Health Statistics Quarterly*, 50, pp.170–177.

Tocheva, A.S. et al., 2013. Distribution of carried pneumococcal clones in UK children following the introduction of the 7-valent pneumococcal conjugate vaccine: A 3-year cross-sectional population based analysis. *Vaccine*, 31, pp.3187–3190.

Tondella, M.L.C. et al., 2000. Distribution of Neisseria meningitidis serogroup B serosubtypes and serotypes circulating in the United States. *Journal of Clinical Microbiology*, 38, pp.3323–3328.

Trotter, C.L., Gay, N.J. & Edmunds, W.J., 2005. Dynamic models of meningococcal carriage, disease, and the impact of serogroup C conjugate vaccination. *American Journal of Epidemiology*, 162, pp.89–100.

Trotter, C.L. & Greenwood, B.M., 2007. Meningococcal carriage in the African meningitis belt. *Lancet Infectious Diseases*, 7, pp.797–803.

Turner, P. et al., 2011. Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray. *Journal of Clinical Microbiology*, 49, pp.1784–1789.

Urwin, R. et al., 2004. Distribution of surface protein variants among hyperinvasive meningococci: Implications for vaccine design. *Infection and Immunity*, 72, pp.5955–5962.

V., V., 1926. Variazoni e fluttazioni del numero d'individui in specie animali conviventi. *Mém. Acad. Lincei Roma*, 2, pp.31–113.

Vermont, C.L. et al., 2005. Porin A-specific antibody avidity in patients who are convalescing from meningococcal B disease. *Pediatric Research*, 58(1), pp.149–152.

Vestrheim, D.F. et al., 2010. Impact of a pneumococcal conjugate vaccination program on carriage among children in Norway. *Clinical and Vaccine Immunology*, 17, pp.325–334.

Victorino, L.R. et al., 2008. Fine-scale phylogeographic structure of Borrelia lusitaniae revealed by multilocus sequence typing. *PLoS ONE*, 3(12).

Virji, M. et al., 1999. Critical determinants of host receptor targeting by Neisseria meningitidis and Neisseria gonorrhoeae: identification of Opa adhesiotopes on the N-domain of CD66 molecules. *Molecular Microbiology*, 34, pp.538–551.

Vyse, A. et al., 2011. Meningococcal disease in Asia: an under-recognized public health burden. *Epidemiology and Infection*, 139, pp.967–985.

De Wals, P. & Bouckaert, A., 1985. Methods for estimating the duration of bacterial carriage. *International journal of epidemiology*, 14(4), pp.628–634.

Wartha, F. et al., 2007. Capsule and D-alanylated lipoteichoic acids protect Streptococcus pneumoniae against neutrophil extracellular traps. *Cellular Microbiology*, 9, pp.1162–1171.

Watkins, E.R. & Maiden, M.C.J., 2012. Persistence of Hyperinvasive Meningococcal Strain Types during Global Spread as Recorded in the PubMLST Database. *Plos One*, 7.

Wedege, E. et al., 2003. Antibody specificities and effect of meningococcal carriage in Icelandic teenagers receiving the Norwegian serogroup B outer membrane vesicle vaccine. *Infection and Immunity*, 71, pp.3775–3781.

Weinberger, D.M. et al., 2010. Association of serotype with risk of death due to pneumococcal pneumonia: a meta-analysis. *Clin Infect Dis*, 51, pp.692–699.

Weinberger, D.M. et al., 2008. Epidemiologic evidence for serotype-specific acquired immunity to pneumococcal carriage. *J Infect Dis*, 197, pp.1511–1518.

Weinberger, D.M. et al., 2009. Pneumococcal Capsular Polysaccharide Structure Predicts Serotype Prevalence. *Plos Pathogens*, 5.

Weinberger, D.M., Malley, R. & Lipsitch, M., 2011. Serotype replacement in disease after pneumococcal vaccination. *Lancet*, 378, pp.1962–1973.

Wertheim, H.F.L. et al., 2005. Associations between Staphylococcus aureus Genotype, Infection, and In-Hospital Mortality: A Nested Case-Control Study. *The Journal of infectious diseases*, 192, pp.1196–1200.

Whalen, C.M. et al., 1995. The changing epidemiology of invasive meningococcal disease in Canada, 1985 through 1992. Emergence of a virulent clone of Neisseria meningitidis. *JAMA*, 273, pp.390–394.

White, L.J., Cox, M.J. & Medley, G.F., 1998. Cross immunity and vaccination against multiple microparasite strains. *IMA journal of mathematics applied in medicine and biology*, 15, pp.211–233.

Whittam, T.S., Ochman, H. & Selander, R.K., 1983. Multilocus Genetic-Structure In Natural-Populations Of Escherichia-Coli. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 80, pp.1751–1755.

Wiesner, P.J. & Thompson 3rd, S.E., 1980. Gonococcal diseases. *Disease-a-month : DM*, 26, pp.1–44.

Williams, J.M. et al., 1998. Using the yeast two-hybrid system to identify human epithelial cell proteins that bind gonococcal Opa proteins: intracellular gonococci bind pyruvate kinase via their Opa proteins and require host pyruvate for growth. *Molecular Microbiology*, 27, pp.171–186.

Wilson, D.J., Falush, D. & McVean, G., 2005. Germs, genomes and genealogies. *Trends in Ecology & Evolution*, 20, pp.39–45.

Wirth, T. et al., 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. *Molecular Microbiology*, 60, pp.1136–1151.

Wisniewski-Dye, F. & Vial, L., 2008. Phase and antigenic variation mediated by genome modifications. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology*, 94, pp.493–515.

Wyres, K.L. et al., 2013. Pneumococcal Capsular Switching: A Historical Perspective. *Journal of Infectious Diseases*, 207, pp.439–449.

Xie, O. et al., 2013. Emergence of serogroup X meningococcal disease in Africa: Need for a vaccine. *Vaccine*, 31, pp.2852–2861.

Yazdankhah, S.P. et al., 2004. Distribution of serogroups and genotypes among disease-associated and carried isolates of Neisseria meningitidis from the Czech Republic, Greece, and Norway. *Journal of Clinical Microbiology*, 42, pp.5146–5153.

Yorke, J.A., Hethcote, H.W. & Nold, A., 1978. Dynamics And Control Of Transmission Of Gonorrhea. *Sexually Transmitted Diseases*, 5, pp.51–56.

Yother, J., 2011. Capsules of Streptococcus pneumoniae and Other Bacteria: Paradigms for Polysaccharide Biosynthesis and Regulation. *Annual Review of Microbiology, Vol 65*, 65, pp.563–581.

Zahner, D., Gudlavalleti, A. & Stephens, D.S., 2010. Increase in Pilus Islet 2-encoded Pili among Streptococcus pneumoniae Isolates, Atlanta, Georgia, USA. *Emerging Infectious Diseases*, 16, pp.955–962.

Zak, K. et al., 1984. Antigenic Variation During Infection With Neisseria-Gonorrhoeae - Detection Of Antibodies To Surface-Proteins In Sera Of Patients With Gonorrhea. *Journal of Infectious Diseases*, 149, pp.166–174.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18, pp.821–829.

Zhang, J.R. et al., 2000. The polymeric immunoglobulin receptor translocates pneumococci across human nasopharyngeal epithelial cells. *Cell*, 102, pp.827–837.

Zhang, X. et al., 2008. Genetic characteristics of serogroup A meningococci circulating in China, 1956-2005. *Clinical Microbiology and Infection*, 14, pp.555–561.

Zhang, Y., Auranen, K. & Eichner, M., 2004. The influence of competition and vaccination on the coexistence of two pneumococcal serotypes. *Epidemiology and infection*, 132(6), pp.1073–1081.

Zhao, J.H., 2004. 2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis. *Bioinformatics*, 20, pp.1325–1326.

Zhu, P. et al., 2001. Fit genotypes and escape variants of subgroup III Neisseria meningitidis during three pandemics of epidemic meningitis. *Proc Natl Acad Sci U S A*, 98, pp.5234–5239.

Appendix I.

Additional Results

Comparing the Simulation Results of the Full SIR and Simplified Determinsitic Frameworks Presented in Chapter 3

Two-allele system

Qualitative patterns of behaviour of the full SIR system were identical to those of the simplified system: with non-overlapping associations between AT and MT at low differences in R0 between MTs (Figures A1 & A3), and competitive exclusion beyond a certain threshold (Figures A3 & A4). As well as consistency between overall strain patterns, the thresholds at which competitive exclusion of the least transmissible MT occurred were similar within both frameworks (Table A1). The total proportion of the population in the infected and immune categories at equilibrium was also highly similar across both models (Table A2, Table A3).

The dynamics were highly similar in both frameworks, with damped oscillations in Za and Zb settling down to equilibrium at 115 and 120 years in the full SIR and simplified frameworks respectively where non-overlapping associations were observed (Figures A1 & A2), and at 100 years where competitive exclusion was observed (Figures A3 & A4). The change in infected individuals over time was also highly comparable in both models.

R0	Exclusion Threshold for Full SIR model (%)	Exclusion Threshold for Simplified model (%)
1.5	0.01	0.07
3	0.13	0.17
5	0.5	0.3
10	1.3	1

Table A1. The exclusion threshold indicates the percentage difference of R0 between MTs at which competitive exclusion occurs. ($\beta_3 > \beta_2 > \beta_1$; $\beta_3 = 30$; σ =13.3, 10, 6, 3; $\mu = 0.02$).

R0	Full SIR Model	Simplified Model
1.5	0.00066	0.00067
3	0.0027	0.0027
5	0.0053	0.0053
10	0.012	0.012

Table A2. Total proportions of infected individuals for the two-allele frameworks. ($\beta_3 = \beta_2 = \beta_1$; $\beta = 30$; $\sigma = 13.3, 10, 6, 3$; $\mu = 0.02$)

R0	Full SIR Model	Simplified Model
1.5	0.46	0.50
3	0.80	0.80
5	0.88	0.89
10	0.94	0.95

Table A3. Total proportions of immune individuals for the two-allele models ($\beta_3 = \beta_2 = \beta_1$; $\beta = 30$; $\sigma = 13.3, 10, 6, 3$; $\mu = 0.02$).

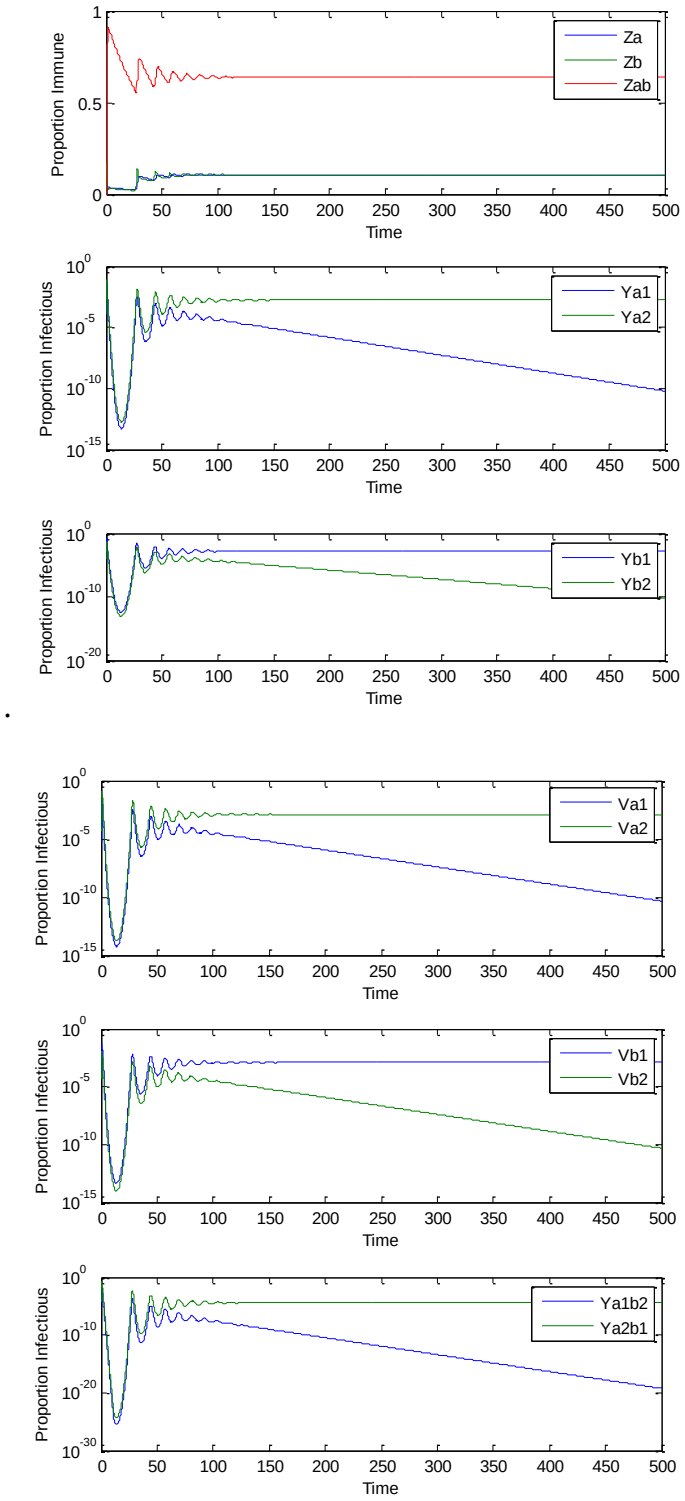


Figure A1. The proportion immune (Z) and infected (Y and V) as simulated by the full SIR two-allele model, for strains with equal R0 values, over 500 years *(β1=β2=20; σ = 4; μ = 0.02). The discordant strains a2 and b1 are dominant.*

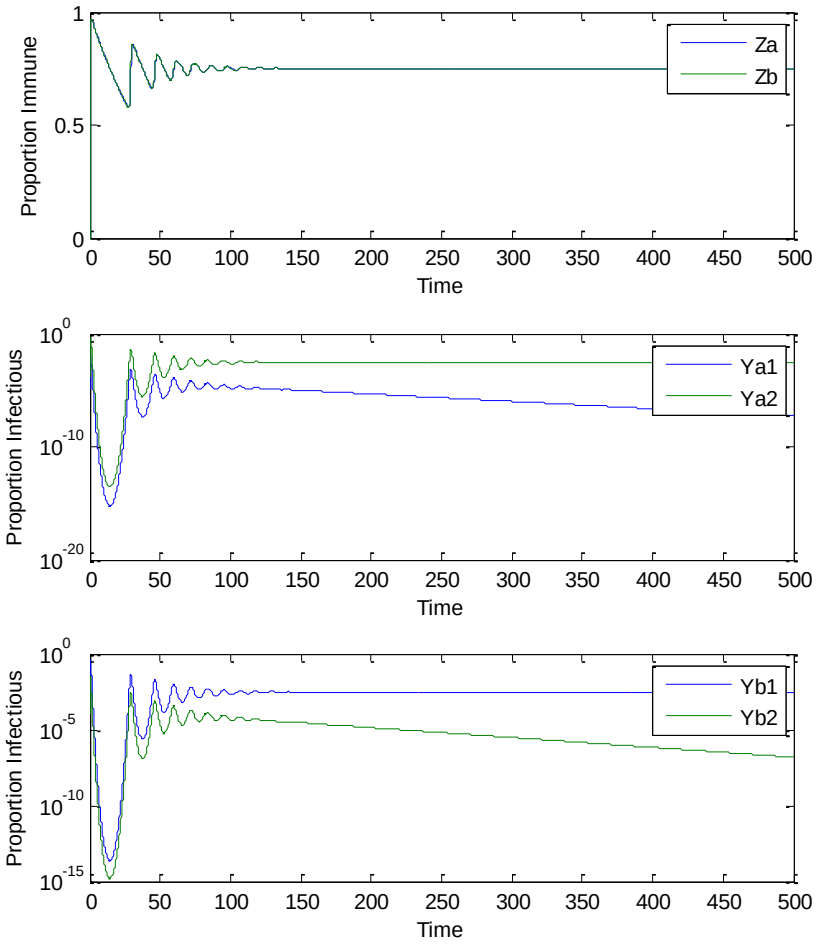


Figure A2. The proportion immune (Z) and infected (Y) as simulated by simplified two-allele model, for 500 years (β1=β2=20; σ = 4; μ = 0.02). The discordant strains a2 and b1 are dominant.

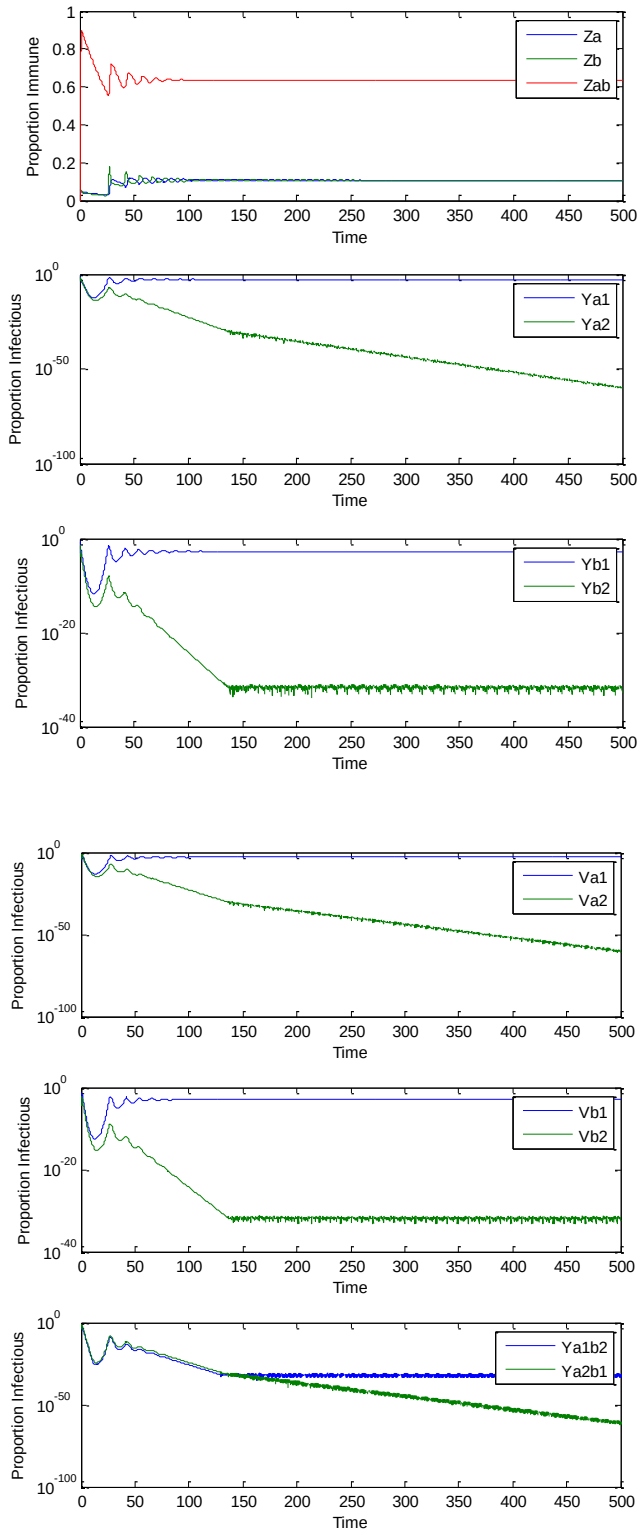


Figure A3. The proportion immune (Z) and infected (Y and V) as simulated by the non-overlapping single-locus two-allele model, with strong interstrain competition, over 500 years *($\beta1=20$; $\beta2=18$; $\sigma = 4$; $\mu = 0.02$). ($\Delta R_0^{MT}=10\%$). MT2 has been competitively excluded, leaving MT1 to dominate both ATs.*

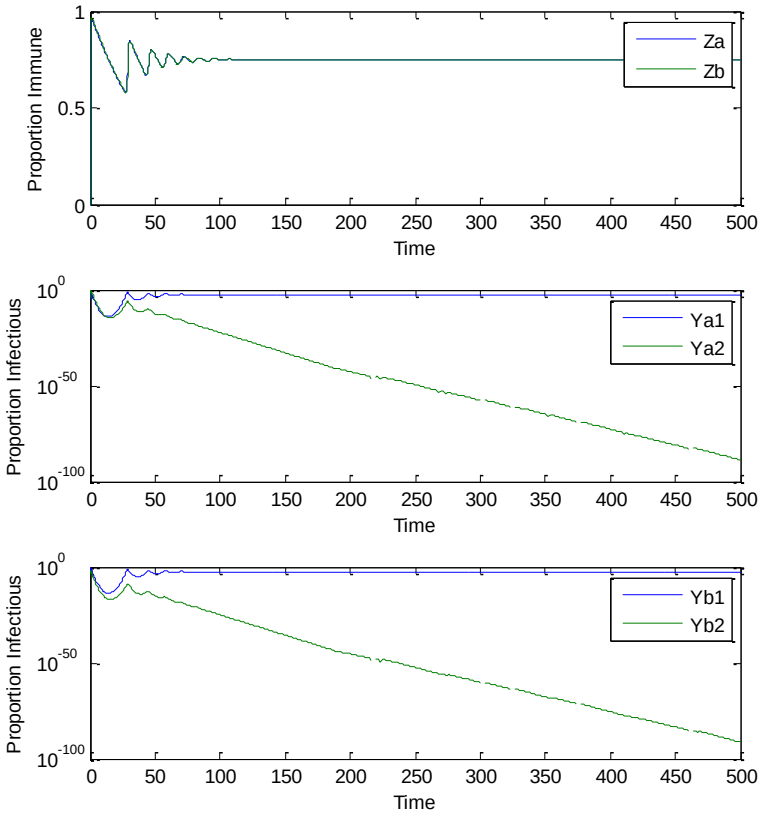


Figure A4. The proportion immune (Z) and infected (Y) as simulated by the simplified single-locus two-allele model, with strong interstrain competition, over 500 years *(β1=20; β2=18; σ = 4; μ = 0.02). (ΔR0=10%). MT2 has been competitively excluded, leaving MT1 to dominate both ATs.*

Three-allele system

The behaviours exhibited by both models were again highly similar both qualitatively and quantitatively. At absent to low levels of differences in R_0 between MTs, all six MTs and ATs were permitted to co-exist, but emerged as three strains characterised by non-overlapping associations of AT and MT. After Δβ passed a primary threshold, the least transmissible MT decayed in prevalence to zero. As the differences in R_0 increased further, the MT with the highest fitness displaced both other MTs. The thresholds in ΔR_0^{MT} at which competitive

exclusion occurred in both frameworks at different R0s are shown in Table A4. The proportions of infected and immune individuals in both models are almost identical, shown in Tables A5 and A6 respectively.

R0	Exclusion Threshold of MT3 for Full SIR model (%)	Exclusion Threshold of MT3 for Simplified model (%)	Exclusion Threshold of MT2 and MT3 for Full SIR Model (%)	Exclusion Threshold of MT2 and MT3 for Simplified model (%)
1.5	0.01	0.07	0.30	0.13
3	0.13	0.17	0.93	0.60
5	0.5	0.3	2.0	1.3
10	1.3	1	4.0	2.6

Table A4. The exclusion threshold indicates the percentage difference of R0 between MTs at which competitive exclusion occurs. ($\beta_3 > \beta_2 > \beta_1$; $\beta_3 = 30$; $\sigma = 13.3, 10, 6, 3$; $\mu = 0.02$).

R0	Full SIR Model	Simplified Model
1.5	0.0020	0.0020
3	0.0040	0.0040
5	0.0080	0.0080
10	0.017	0.018

Table A5. Total proportions of infected individuals for the three-allele frameworks. ($\beta_3 = \beta_2 = \beta_1$; $\beta = 30$; $\sigma = 13.3, 10, 6, 3$; $\mu = 0.02$).

R0	Full SIR Model	Simplified Model
1.5	0.70	0.73
3	0.85	0.86
5	0.91	0.92
10	0.95	0.96

Table A6. Total proportions of immune individuals for the three-allele frameworks. *($\beta_3 = \beta_2 = \beta_1$; $\beta = 30$; $\sigma = 13.3, 10, 6, 3$; $\mu = 0.02$).*

Supplementary Information for Chapter 7

When all 153 genomes were analysed against reference strain Z2491, identical alleles across all isolates were found at 352 loci including the reference genome, and 103 loci excluding the reference genome loci (out of 1993 annotated loci present in the reference genome). Variable alleles were found at 999 loci, truncated sequences at one or more isolates were found at 528 loci, and 56 loci were paralogous and excluded from further analysis.

When isolates from the third pandemic wave and ST-2859 were analysed against reference strain WUE 2594, identical alleles across all isolates were found at 948 loci including the reference genome, and 109 loci excluding the reference genome loci (out of 2070 annotated loci present in the reference genome). Variable alleles were found at 560 loci, truncated sequences at one or more isolates were found at 392 loci, and 37 loci were paralogous and excluded from further analysis.

Isolate	Aliases/Accession Number	Country	Year	Strain Designation	ST (MLST)	clonal_complex (MLST)
A4/M1027	NIBSC:2803; Z1001; mlst082	USA	1937	A: P1.5-2,10: F1-5: ST-4 (cc4)	4	ST-4 complex
120M	NIBSC:2822; Z1035; mlst101	Pakistan	1967	A: P1.5-2,10: F5-1: ST-1 (cc1)	1	ST-1 complex
7891	NIBSC:2760; Z1054; mlst041	Finland	1975	A: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
6748	NIBSC:2784; Z1073; mlst065	Canada	1971	A: P1.18-1,3: F5-1: ST-1 (cc1)	1	ST-1 complex
129E	NIBSC:2828; Z1092; mlst107	West Germany	1964	A: P1.5-2,10: F3-6: ST-1 (cc1)	1	ST-1 complex
139M	NIBSC:2795; Z1099; mlst074	Philippines	1968	A: P1.5-2,10: F5-1: ST-1 (cc1)	1	ST-1 complex
S3131	NIBSC:2813; Z1213; mlst092	Ghana	1973	A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
S4355	NIBSC:2806; Z1227; mlst085	Denmark	1974	A: P1.5-1,9: F3-1: ST-5 (cc5)	5	ST-5 complex
10	NIBSC:2825; Z1269; mlst104	Burkina Faso	1963	A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
20	NIBSC:2767; Z1275; mlst048	Niger	1963	A: P1.5-2,10: F1-7: ST-1 (cc1)	1	ST-1 complex
26	NIBSC:2764; Z1278; mlst045	Niger	1963	A: P1.7,13: F1-5: ST-4 (cc4)	4	ST-4 complex
255	NIBSC:2811; Z1318; mlst090	Burkina Faso	1966	A: P1.7-2,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
243	NIBSC:2779; Z1362; mlst060	Cameroon	1966	A: P1.7,13: F1-5: ST-4 (cc4)	4	ST-4 complex
393	NIBSC:2823; Z1392; mlst102	Greece	1968	A: P1.5-2,10: F5-1: ST-1 (cc1)	1	ST-1 complex
254	NIBSC:2812; Z5010; mlst091	Djibouti	1966	A: P1.5-2,10: F1-7: ST-1 (cc1)	1	ST-1 complex
S5611	NIBSC:2765; Z1466; mlst046	Australia	1977	A: P1.5-2,10: F5-1: ST-1 (cc1)	1	ST-1 complex
11-004	NIBSC:2826; Z1503; mlst105	China	1984	A: P1.20,9: F3-8: ST-5 (cc5)	5	ST-5 complex
IAL2229	NIBSC:2816; Z1506; mlst095	Brazil	1976	A: P1.20,9: F2-1: ST-5 (cc5)	5	ST-5 complex
F4698	NIBSC:2731; Z3515; mlst005	Saudi	1987	A: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
F6124	NIBSC:2730; Z3524; mlst002	Chad	1988	A: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
1014	NIBSC:2821; Z3667; mlst100	Sudan	1985	A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
H1964	NIBSC:2796; Z3771; mlst075	UK	1987	A: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
NMBI	Z5463BC	France		A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
Z814	Z5463	Gambia		A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
153	NIBSC:2733; Z3905; mlst001	China	1966	A: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
154	NIBSC:2766; Z3906; mlst047	China	1966	A: P1.20,9: F3-1: ST-6 (cc5)	6	ST-5 complex
80049	NIBSC:2793; Z4099; mlst072	China	1963	A: P1.5-2,10: F1-5: ST-5 (cc5)	5	ST-5 complex
243	Z5463PI	Gambia		A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
D8	NIBSC:2762; Z4186; mlst043	Mali	1990	A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
2059001	NIBSC:2745; Z4421; mlst026	Mali	1990	A: P1.7,13: F1-5: ST-4 (cc4)	4	ST-4 complex
BZ 133	NIBSC:2824; Z4665; mlst103	Netherlands	1977	B: P1.7,16: F5-1: ST-10300 (cc1)	10300	ST-1 complex
14/1455	NIBSC:2732; Z4717; mlst004	USSR	1970	A: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
371	NIBSC:2769; Z4756; mlst050	India	1980	A: P1.5-2,10: F5-1: ST-1 (cc1)	1	ST-1 complex
690	NIBSC:2805; Z4757; mlst084	India	1980	A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
106	NIBSC:2808; Z5005; mlst087	Morocco	1967	A: P1.5-2,10: F5-1: ST-1 (cc1)	1	ST-1 complex
79128	NIBSC:2800; Z5035; mlst079	China	1979	A: P1.7-1,10: F5-5: ST-3 (cc1)	3	ST-1 complex
322/85	NIBSC:2810; Z5037; mlst089	East Germany	1985	A: P1.5-2,10: F5-2: ST-2 (cc1)	2	ST-1 complex
79126	NIBSC:2777; Z5043; mlst058	China	1979	A: P1.7-3,10-5: F5-5: ST-3 (cc1)	3	ST-1 complex
92001	NIBSC:2783; Z5826; mlst064	China	1992	A: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
Z2491	C751; NIBSC:2763; Z6244; mlst044	Gambia	1983	A: P1.7,13-1: F1-5: ST-4 (cc4)	4	ST-4 complex
M13220		Philippines	2005	A: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
M18575		Burkina Faso	2003	A: P1.ND,ND: F-ND: ST-2859 (cc5)	2859	ST-5 complex
WUE 2594		Germany	1991	A: P1.20,9: F1-21: ST-5 (cc5)	5	ST-5 complex
ERR051676				ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex

ERR051677	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR051678	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR051693	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
ERR051687	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
ERR052799	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052812	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052817	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
SRR513938	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
ERR052816	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052813	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052829	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
SRR513846	ND: P1.20,9: F3-1: ST-580 (cc5)	580	ST-5 complex
ERR052831	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052822	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052826	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
SRR513840	ND: P1.20,9: F3-1: ST-4789 (cc5)	4789	ST-5 complex
ERR052811	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052800	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052821	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052795	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052823	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052787	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052828	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052792	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
SRR514831	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR513836	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052767	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052775	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052754	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052768	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052746	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052814	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052819	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
SRR513853	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052827	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
SRR513841	ND: P1.20,9: F-ND: ST-8428 (cc5)	8428	ST-5 complex
ERR052825	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052796	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052820	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052737	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052779	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052793	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052824	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052747	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052807	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052806	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex

ERR052789	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052739	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052763	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052832	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052790	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052764	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052781	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052749	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052818	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052750	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052766	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052815	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052751	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052808	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052738	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052758	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052743	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052757	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR513837	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052745	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052774	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052791	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052810	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052778	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052804	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052771	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052765	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052741	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052786	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052752	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052805	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052780	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052782	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052744	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052809	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052802	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052769	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052759	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052773	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052740	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052756	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052830	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052794	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052755	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052784	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052761	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex

ERR052753	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052801	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052783	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052798	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052770	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052788	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052803	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052797	ND: P1.20,9: F3-1: ST-2859 (cc5)	2859	ST-5 complex
ERR052760	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052777	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052785	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052776	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052762	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052742	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052772	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
ERR052748	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606726	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606727	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606739	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606725	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606730	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606738	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606721	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606732	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606729	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606723	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606720	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606737	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606722	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606736	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606724	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606731	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606696	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606697	ND: P1.20,9: F2-1: ST-5 (cc5)	5	ST-5 complex
SRR606719	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606733	ND: P1.20-1,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606699	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606734	ND: P1.20-1,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606695	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606698	ND: P1.20,9: F2-1: ST-5 (cc5)	5	ST-5 complex
SRR606700	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606718	ND: P1.20,9: F3-1: ST-5 (cc5)	5	ST-5 complex
SRR606703	ND: P1.20,9: F3-1: ST-580 (cc5)	580	ST-5 complex
SRR606691	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex
SRR606704	ND: P1.20,9: F3-1: ST-580 (cc5)	580	ST-5 complex
SRR606692	ND: P1.20,9: F3-1: ST-7 (cc5)	7	ST-5 complex

M11 240262	UK	2011	ND: P1.20,9: F3-1: ST-4789 (cc5)	4789	ST-5 complex	

Table A7. List of isolates used to plot Figure 7.1 and for analysis of the ST-5 complex.

Locus in reference genome	Product	Genome Position	Functional characterisation
NMA0408	putative integral membrane protein	377328	Metabolism
NMA0410	3-demethylubiquinone-9 3-methyltransferase	380455	Metabolism
NMA0411	putative homoserine kinase	381201	Metabolism
NMA0478	putative outer membrane peptidase	463713	Environmental Information Processing
NMA0537	putative integral membrane protein	517957	Unknown
NMA0606	hypothetical protein NMA0606	580081	Unknown
NMA0608	carbamoyl phosphate synthase small subunit	580546	Metabolism
NMA0621	putative riboflavin kinase/FMN adenylyltransferase	599508	Metabolism
NMA0708	putative hexosaminidase	696057	Metabolism
NMA0709	putative integral membrane protein	697199	Environmental Information Processing
NMA0710	putative periplasmic serine protease	698921	Metabolism
NMA0946	putative regulatory protein	914617	Environmental Information Processing
NMA1104	putative superoxide dismutase	1052308	Cellular processes
NMA1539	putative cation-transporting ATPase	1432649	Environmental Information Processing
NMA1571	iron/sulphur-binding oxidoreductase	1475160	Metabolism
NMA1673	putative integral membrane transporter	1595427	Environmental Information Processing
NMA1683	ClpB protein	1605805	Genetic Information Processing
NMA1708	hypothetical protein	1639365	Metabolism
NMA1748	glutaminyl-tRNA synthetase	1691533	Genetic Information Processing
NMA1997	hypothetical protein (pseudogene)	1936559	Other

Table A8. Loci which contained alleles unique to the strains from the first pandemic wave.

Locus in reference genome	Product	Genome Position	Functional characterisation
NMA0004	Fic/DOC family	2952	Other
NMA0005	NADH dehydrogenase I chain K	3561	Metabolism
NMA0006	NADH dehydrogenase I chain J	3863	Metabolism
NMA0093	Zinc transporter ZupT	93576	Environmental information processing
NMA0094	valyl-tRNA synthetase	94482	Genetic Information Processing
NMA0132	50S ribosome-binding GTPase	115873	Genetic Information Processing
NMA0185	capsule polysaccharide modification protein	167551	Metabolism
NMA0186	capsule polysaccharide modification protein	168947	Metabolism
NMA0241	electron transfer flavoprotein alpha-subunit	225154	Metabolism
NMA0258	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	241392	Metabolism
NMA0260	putative integral membrane protein	244168	Unknown
NMA0280	Iron permease FTR1 family	265572	Environmental information processing
NMA0345	GTPase ObgE	320559	Genetic Information Processing
NMA0365	Permease	337526	Environmental information processing
NMA0442	Divalent cation transporter	415729	Environmental information processing
NMA0452	membrane transport solute-binding protein	426802	Environmental information processing
NMA0453	TonB dependent receptor (FetA)	428465	Antigen/virulence; Environmental Information Processing
NMA0478	outer membrane peptidase	463713	Other
NMA0480	aldehyde dehydrogenase A	468118	Metabolism
NMA0483	putative transcriptional regulator	470401	Genetic Information Processing
NMA0485	ABC transporter ATP binding protein	471431	Environmental information processing
NMA0486	ABC transport inner membrane subunit	472339	Environmental information processing
NMA0487	outer membrane transport protein	473166	Environmental information processing
NMA0488	periplasmic transport protein	473697	Environmental information processing
NMA0489	STAS domain	474320	Unknown
NMA0490	putative periplasmic/outer membrane protein	474619	Unknown
NMA0622	Isoleucine--tRNA ligase	600606	Genetic Information Processing
NMA0631	HNH endonuclease	610868	Other
NMA0658	Cytochrome C	644324	Metabolism
NMA0707	7-carboxy-7-deazaguanine synthase	694741	Metabolism
NMA0740	Unknown	733754	Unknown
NMA0741	ubiquinone biosynthesis protein UbiB	734253	Metabolism
NMA0745	putative periplasmic protein	737825	Unknown
NMA0746	thiamine biosynthesis protein	738066	Metabolism
NMA0747	Na(+)-translocating NADH-quinone reductase subunit F	739274	Metabolism
NMA0748	Na(+)-translocating NADH-quinone reductase subunit E	740505	Metabolism

NMA0749	Na(+)-translocating NADH-quinone reductase subunit D	741102	Metabolism
NMA0750	Na(+)-translocating NADH-quinone reductase subunit C	741728	Metabolism
NMA0751	Na(+)-translocating NADH-quinone reductase subunit B	742497	Metabolism
NMA0906	competence protein	881651	Other
NMA0925	Cytochrome C	897012	Metabolism
NMA1020	transcriptional regulator	981761	Genetic Information Processing
NMA1024	Adenylosuccinate synthetase	986609	Metabolism
NMA1084	Putative MetA-pathway of phenol degradation	1032148	Metabolism
NMA1120	Oxidoreductase	1070172	Other
NMA1263	Nudix family	1189567	Unknown
NMA1356	UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase	1253355	Metabolism
NMA1416	riboflavin synthase alpha subunit	1314389	Metabolism
NMA1438	putative integral membrane protein	1331072	Unknown
NMA1572	pyridoxamine-5'-phosphate oxidase	1476300	Metabolism
NMA1573	pseudouridine synthase	1477432	Metabolism
NMA1574	integral membrane transporter	1478271	Environmental information processing
NMA1584	acetylornithine aminotransferase	1489480	Metabolism
NMA1589	putative type III restriction/modification system modification methylase (pseudogene)	1494377	Other
NMA1675	Transposase	1598453	Other
NMA1742	CTP synthase	1684464	Metabolism
NMA1748	glutaminyl-tRNA synthetase	1691533	Genetic Information Processing
NMA1771	histidinol-phosphate aminotransferase	1716562	Metabolism
NMA1838	DNA-binding protein	1781866	Unknown
NMA1891	Nitroreductase	1816250	Other
NMA1894	phosphoserine aminotransferase	1818798	Metabolism
NMA1896	N utilisation substance protein A	1820713	Genetic Information Processing
NMA1897	initiation factor IF2	1822242	Genetic Information Processing
NMA1923	Peptidase	1853690	Metabolism
NMA2052	aconitate hydratase	1996523	Metabolism
NMA2053	Sulfite exporter TauE/SafE	1999207	Environmental information processing
NMA2054	citrate synthase	2000156	Metabolism
NMA2055	Phosphoenolpyruvate phosphomutase	2001396	Metabolism
NMA2056	Permease	2002897	Environmental information processing

Table A9. Loci which contained alleles unique to the strains from the second pandemic wave.

Locus in reference genome	Product	Genome Position	Functional characterisation
NMA0035	Glutamate-ammonia-ligase adenylyltransferase	28626	Metabolism
NMA0049	RNA polymerase sigma factor	41790	Environmental Information Processing
NMA0066	dihydrodipicolinate reductase	60973	Metabolism
NMA0071	ribonuclease E	65414	Genetic Information Processing
NMA0080	ribosome recycling factor	79151	Genetic Information Processing
NMA0081	Isoprenyl transferase	79764	Metabolism
NMA0083	1-deoxy-D-xylulose 5-phosphate reductoisomerase	81366	Metabolism
NMA0135	elongation factor G	118817	Genetic Information Processing
NMA0174	unknown	157181	Unknown
NMA0280	Iron permease FTR1 family	265572	Environmental Information Processing
NMA0299	Transferrin binding protein-like solute binding protein	291661	Environmental Information Processing
NMA0350	Serine hydrolase	326157	Metabolism
NMA0408	Sulfatase	377328	Metabolism
NMA0477	para-aminobenzoate synthase component I	460532	Metabolism
NMA0519	ATP synthase beta chain	496617	Metabolism
NMA0521	glycyl-tRNA synthetase alpha chain	498802	Genetic Information Processing
NMA0522	unknown	499784	Unknown
NMA0634	unknown	618565	Unknown
NMA0650	pilus secretin	635588	Other
NMA0707	7-carboxy-7-deazaguanine synthase (CDG synthase)	694741	Metabolism
NMA0759	glycine cleavage system component H	750467	Metabolism
NMA0902	Ribosomal RNA small subunit methyltransferase A	871027	Genetic Information Processing
NMA0968	Phosphoribosylaminoimidazole-succinocarboxamide synthase	932217	Metabolism
NMA1000	ABC transporter	964511	Environmental Information Processing
NMA1001	phosphoglucomutase	965286	Metabolism
NMA1002	peptidyl-prolyl cis-trans isomerase B	966831	Genetic Information Processing
NMA1003	Transmembrane transporter	967459	Environmental Information Processing
NMA1003A	unknown	969207	Unknown
NMA1004	Peptidyl-tRNA hydrolase	969731	Genetic Information Processing
NMA1005	RnfH family Ubiquitin	970362	Genetic Information Processing
NMA1006	Polyketide cyclase / dehydrase and lipid transport	970633	Metabolism
NMA1007	ATP-dependent zinc metallopeptidase	971286	Genetic Information Processing
NMA1038	type I restriction-modification system protein	998259	Genetic Information Processing
NMA1107	putative membrane protein	1055468	Unknown
NMA1131	unknown	1079957	Unknown
NMA1134	putative lipoprotein	1082178	Unknown

NMA1136	homoserine O-acetyltransferase	1083718	Metabolism
NMA1137	50S ribosomal protein L36	1085479	Genetic Information Processing
NMA1138	additional 50S ribosomal protein L31	1085604	Genetic Information Processing
NMA1139	Methylenetetrahydrofolate reductase	1086078	Metabolism
NMA1140	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase	1087095	Metabolism
NMA1141	Redoxin	1089650	Cellular Processes
NMA1142	dihydrolipoamide dehydrogenase	1090636	Metabolism
NMA1200	Putative surface fibril protein	1139032	Other
NMA1210	unknown	1153976	Unknown
NMA1249	ABC transporter	1175111	Environmental Information Processing
NMA1267	Gamma-glutamyl phosphate reductase	1195261	Metabolism
NMA1400	Methyltransferase small domain	1299257	Genetic Information Processing
NMA1445	GTPase HflX (GTP-binding protein HflX)	1339576	Genetic Information Processing
NMA1446	unknown	1340759	Unknown
NMA1447	unknown	1342192	Unknown
NMA1448	DNA repair protein RadC	1342990	Genetic Information Processing
NMA1449	Glutamate-cysteine ligase	1343790	Metabolism
NMA1451	putative lipoprotein	1346901	Unknown
NMA1452	3-isopropylmalate dehydratase small subunit	1347217	Metabolism
NMA1453	DNA modification methylase	1348033	Metabolism
NMA1465	tRNA 2-thiocytidine biosynthesis protein TtcA	1357496	Genetic Information Processing
NMA1472	protein-tyrosine-phosphatase	1365637	Environmental Information Processing
NMA1473	glycerate kinase	1366217	Metabolism
NMA1476	mercuric ion binding protein	1370357	Genetic Information Processing
NMA1478	polysaccharide modification protein	1371201	Metabolism
NMA1491	transcription-repair coupling factor	1387161	Genetic Information Processing
NMA1524	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	1419151	Metabolism
NMA1541	sel1 repeat	1437010	Unknown
NMA1542	tRNA (guanine-N(7)-)-methyltransferase	1438565	Genetic Information Processing
NMA1544	virus-related protein	1440251	Other
NMA1545	excinuclease ABC subunit B	1441083	Genetic Information Processing
NMA1547	peptidase	1445043	Metabolism
NMA1563	SUN-family protein	1466222	Genetic Information Processing
NMA1588	tRNA pseudouridine synthase B	1493137	Metabolism
NMA1591	type III restriction/modification system enzyme	1496471	Genetic Information Processing
NMA1592	L-lactate dehydrogenase	1499437	Metabolism
NMA1593	Transcriptional regulator	1500899	Genetic Information Processing
NMA1594	NifS-like aminotranfserase	1501374	Metabolism; Genetic Information Processing
NMA1620	cytolysin secretion ABC transporter	1523857	Environmental Information Processing
NMA1677	unknown	1600980	Unknown
NMA1683	ClpB protein	1605805	Genetic Information Processing

NMA1730	succinyl-diaminopimelate desuccinylase	1663067	Metabolism
NMA1748	glutaminyl-tRNA synthetase	1691533	Genetic Information Processing
NMA1766	acetolactate synthase isozyme III large subunit	1709994	Metabolism
NMA1850	Phage Mu protein F like protein	1788811	Other
NMA1882	transposase	1804208	Unknown
NMA1887	nitrite reductase, major outer membrane copper-containing protein	1810404	Metabolism
NMA1936	cytochrome	1868180	Metabolism
NMA1946	DNAase	1879996	Metabolism
NMA1947	L-asparaginase	1880890	Metabolism
NMA1948	DedA-family integral membrane protein	1881946	Unknown
NMA1949	Phosphoglucosamine mutase	1882819	Metabolism
NMA1964	glutamate dehydrogenase	1898767	Metabolism
NMA1975	integral membrane protein	1915001	Unknown
NMA2024	transferrin-binding protein A	1962649	Antigen/virulence
NMA2025	transferrin-binding protein B	1965468	Antigen/virulence
NMA2063	cell division protein	2011353	Genetic Information Processing
NMA2080	Chelatase	2028669	Unknown

Table A10. Loci which contained alleles unique to the strains from the third pandemic wave.

Locus in reference genome	Product	Genome Position	Functional characterisation
murA | NMAA_0011	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	7582	Metabolism
glmS | NMAA_0021	glucosamine--fructose-6-phosphate aminotransferase [isomerizing]	21789	Metabolism
thiB | NMAA_0029	Thiamine transport system substrate-binding protein	32717	Environmental information processing
NMAA_0030	Mechanosensitive ion channel	33812	Environmental information processing
NMAA_0031	Competence-damaged protein (CinA family)	34686	Unknown
msrAB | NMAA_0032	peptide methionine sulfoxide reductase (putative pilin biogenesis)	35273	Cellular processes
ftsY | NMAA_0033	probable signal recognition particle protein	36986	Environmental information processing
NMAA_0034	GIY-YIG catalytic domain	38840	Unknown
NMAA_0100	putative transcriptional activator protein METR	113047	Genetic information processing
fetA | NMAA_0164	TonB dependent receptor (FetA)	193640	Antigenic
NMAA_0299	ATP-binding protein involved in chromosome partitioning	353893	Cellular processes
NMAA_0300	putative membrane-associated thioredoxin	355325	Cellular processes
pglD | NMAA_0321	pilin glycosylation protein pglD	387460	Antigenic
pglC | NMAA_0322	pilin glycosylation protein pglC	389418	Antigenic
pglB1 | NMAA_0323	pilin glycosylation protein pglB1	390723	Antigenic
NMAA_0337	GTP binding protein engB	409695	Cellular processes
cyc | NMAA_0338	Cytochrome C	410530	Metabolism
NMAA_0339	Cytochrome C biogenesis protein	411367	Metabolism
ccsA | NMAA_0340	Cytochrome C biogenesis protein	413375	Metabolism
NMAA_0341	tRNA N6-adenosine threonylcarbamoyltransferase (EC 2.3.1.-)	414679	Metabolism
rluF | NMAA_0635	ribosomal large subunit pseudouridine synthase F	749433	Metabolism
NMAA_0636	ATP-NAD kinase	750216	Metabolism
NMAA_0637	None	751126	Unknown
NMAA_0638	NADH(P)-binding	751743	Unknown
murB | NMAA_0640	UDP-N-acetylenolpyruvoylglucosamine reductase (EC 1.3.1.98)	753236	Metabolism
NMAA_0641	multidrug efflux protein	754454	Environmental information processing
hisZ | NMAA_0642	ATP phosphoribosyltransferase regulatory subunit	756168	Metabolism
purA | NMAA_0643	Adenylosuccinate synthetase	757422	Metabolism
NMAA_0644	None	758839	Unknown
NMAA_0645	None	759132	Unknown
NMAA_0646	None	759583	Unknown
NMAA_0647	None	760110	Unknown
NMAA_0648	None	760563	Unknown
NMAA_0649	None	760708	Unknown
adk | NMAA_0651	adenylate kinase	762625	Metabolism
rfaE1 | NMAA_0653	pfkB family carbohydrate kinase (BIGS: D-beta-D-heptose-7-phosphate kinase)	764593	Metabolism

NMAA_0654	Cytosine-specific methyltransferase	765601	Metabolism
dnaB | NMAA_0710	DnaB-like helicase C terminal domain	823174	Genetic information processing
NMAA_0805	putative phage repressor	930949	Other
NMAA_0886	Putative phage tail fiber protein	1007824	Other
NMAA_0888	None	1010855	Unknown
NMAA_0889	Caudovirales tail fibre assembly protein	1011222	Other
NMAA_0890	None	1011806	Unknown
NMAA_0891	None	1012099	Unknown
NMAA_0892	Putative bacterial lipoprotein (DUF799)	1013947	Unknown
NMAA_0893	None	1014591	Unknown
NMAA_0894	Curli production assembly/transport component CsgG	1014969	Environmental information processing
NMAA_0895	short chain dehydrogenase	1015805	Metabolism
NMAA_0896	Flavin containing amine oxidoreductase	1016929	Metabolism
NMAA_0996	putative type III restriction-modification system enzyme Res	1134598	Cellular processes
NMAA_1049	Sulfite exporter TauE/SafE	1204715	Environmental information processing
NMAA_1067	Fumarylacetoacetate (FAA) hydrolase family	1225027	Metabolism
gdhB | NMAA_1182	Glutamate dehydrogenase	1387610	Metabolism
ackA1 | NMAA_1212	acetate kinase	1424286	Metabolism
lbpB | NMAA_1234	lactoferrin-binding protein	1454533	Antigenic
pyrG | NMAA_1235	CTP synthase	1459282	Metabolism
fadD1 | NMAA_1236	long-chain-fatty-acid--CoA-ligase	1461028	Metabolism
trmU | NMAA_1237	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	1462769	Genetic information processing
dgk | NMAA_1239	diacylglycerol kinase	1464586	Metabolism
acnB | NMAA_1253	aconitate hydratase	1478878	Metabolism
NMAA_1298	None	1538271	Unknown
NMAA_1383	None	1621406	Unknown
NMAA_1495	Nitrilase	1752253	Metabolism
murF | NMAA_1519	UDP-MurNAc-pentapeptide synthetase	1780672	Metabolism
mafA2 | NMAA_1550	adhesin MafA2	1828085	Antigenic
mafB2 | NMAA_1551	adhesin MafB2	1829030	Antigenic
NMAA_1553	None	1830920	Unknown
NMAA_1554	None	1831592	Unknown
NMAA_1555	None	1831846	Unknown
NMAA_1556	Suppressor of fused protein (SUFU)	1831985	Environmental information processing
NMAA_1557	None	1832666	Unknown
NMAA_1558	None	1833347	Unknown
NMAA_1559	Regulator of ribonuclease activity B	1833690	Genetic information processing; metabolism
NMAA_1560	None	1834072	Unknown
NMAA_1561	Suppressor of fused protein (SUFU)	1834426	Environmental information processing
NMAA_1562	None	1835104	Unknown
NMAA_1563	None	1835834	Unknown

NMAA_1564	None	1836359	Unknown
NMAA_1566	None	1837213	Unknown
frpD | NMAA_1567	RTX iron-regulated protein FrpC	1838725	Antigenic
NMAA_1568	GIY-YIG catalytic domain	1839879	Unknown
NMAA_1707	None	1976980	Unknown
nuoB | NMAA_1739	NADH dehydrogenase I chain B	2006433	Metabolism
gidB | NMAA_1788	Ribosomal RNA small subunit methyltransferase G	2069801	Genetic information processing
NMAA_1789	Putative integral membrane protein	2070582	Unknown
NMAA_1790	Fusaric acid resistance protein family	2071324	Other
NMAA_1914	Alanine racemase, N-terminal domain	2195758	Metabolism

Table A11. Loci which contained alleles unique to ST-2859 strains.

	Genome position	Locus in reference genome	Product	Functional characterisation
Genes missing in ST-2859 isolates				
A	38840	NMAA_0034	GIY-YIG catalytic domain	Unknown
E	759132	NMAA_0645	None characterised	Unknown
E	759583	NMAA_0646	None characterised	Unknown
E	760110	NMAA_0647	None characterised	Unknown
E	760563	NMAA_0648	None characterised	Unknown
E	760708	NMAA_0649	None characterised	Unknown
	145453	lbpB		
G	3	NMAA_1234	Lactoferrin-binding protein	Antigenic
	183159			
H	2	NMAA_1554	None characterised	Unknown
	183184			
H	6	NMAA_1555	None characterised	Unknown
	183198			
H	5	NMAA_1556	Suppressor of fused protein	Unknown
	183334			
H	7	NMAA_1558	None characterised	Unknown
	183369			
H	0	NMAA_1559	None characterised	Unknown
	183407			
H	2	NMAA_1560	None characterised	Unknown
	183442			
H	6	NMAA_1561	Suppressor of fused protein	Unknown
Genes missing in third pandemic wave				
	144025			
G	1	NMA1544	Virus-related protein	Other

Table A12. Loci of genes missing in ST-2859 and/or third pandemic wave isolates but present in the ancestral strains.

Isolate	PorA	PorB	FetA	FHBp	NadA	opcA	TbpA	TbpB	NMA1332	GNA33	GNA1220
NIBSC:2760; Z1054; mlst041	P1.20,9	3-27	F3-1	5	7	3	9	9	3	7	45
NIBSC:2806; Z1227; mlst085	P1.5-1,9	3-27	F3-1	5	7	3	9	14	3	7	45
NIBSC:2826; Z1503; mlst105	P1.20,9	3-61	F3-1	5	7	3	9	106	3	17	45
NIBSC:2816; Z1506; mlst095	P1.20,9	3-27	F3-1	5	7	3	9	14	3	7	45
NIBSC:2731; Z3515; mlst005	P1.20,9	3-47	F3-1	5	16	3	9	14	3	19	45
NIBSC:2730; Z3524; mlst002	P1.20,9	3-47	F3-1	5	7		9		3	19	45
NIBSC:2796; Z3771; mlst075	P1.20,9	3-47	F3-1	5	7	3	9	19	3	19	45
NIBSC:2733; Z3905; mlst001	P1.20,9	3-47	F3-1	5	7	3	19	24	3	19	45
NIBSC:2766; Z3906; mlst047	P1.20,9	3-47	F3-1		7	3	9	14	3	19	45
NIBSC:2793; Z4099; mlst072	P1.5-2,10	3-56	F1-5	5	7	3	89	121	3	23	48
NIBSC:2732; Z4717; mlst004	P1.20,9	3-47	F3-1	5	7	3	9	82	3	19	45
NIBSC:2783; Z5826; mlst064	P1.20,9	3-62	F3-1	39	7	3	62	85	3	19	45
FR774048	P1.20,9	3-47	F1-21	5		3	9	14	3	19	2
ERR051676	P1.20,9	3-47	F3-1	39	7	3	1453		1		45
ERR051677	P1.20,9	3-47	F3-1	39	7	3	1453		1	19	45
ERR051678	P1.20,9	3-47	F3-1	39	7	3	1453		1		45
ERR051693	P1.20,9	3-47	F3-1		7	3	9	14	3	19	45
ERR051687	P1.20,9	3-47	F3-1	5	7	3	9		3		45
ERR052799	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052812	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052817	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
SRR513938	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
ERR052816	P1.20,9	3-47	F3-1	39	7	3		85	1	19	45
ERR052813	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052829	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
SRR513846	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
ERR052831	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052822	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052826	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
SRR513840	P1.20,9	3-142	F3-1	39	7	3	1453	85	3	19	45
ERR052811	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052800	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052821	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052795	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052823	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052787	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052828	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052792	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
SRR514831	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
SRR513836	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052767	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052775	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45

ERR052754	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052768	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052746	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052814	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052819	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
SRR513853	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052827	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
SRR513841	P1.20,9	3-142	F-ND	39	7	3	1453		3	19	45
ERR052825	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052796	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052820	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052737	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052779	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052793	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052824	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052747	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052807	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052806	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052789	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052739	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052763	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052832	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052790	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052764	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052781	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052749	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052818	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052750	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052766	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052815	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052751	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052808	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052738	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052758	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052743	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052757	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR513837	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052745	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052774	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052791	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052810	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052778	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052804	P1.20,9	3-47	F3-1	39	7	3	1453		1	19	45
ERR052771	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052765	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052741	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45

ERR052786	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052752	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052805	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052780	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052782	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052744	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052809	P1.20,9	3-47	F3-1	39	7	3	1453		1	19	45
ERR052802	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052769	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052759	P1.20,9	3-47	F3-1	39	7	3	1453		3	19	45
ERR052773	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052740	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052756	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052830	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052794	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052755	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052784	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052761	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052753	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052801	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052783	P1.20,9	3-47	F3-1	39	7		1453	85	3	19	45
ERR052798	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052770	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052788	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052803	P1.20,9	3-47	F3-1	39	7	3	1453	85	1	19	45
ERR052797	P1.20,9	3-47	F3-1	39	7		1453	85	1	19	45
ERR052760	P1.20,9	3-47	F3-1	39	7		1453	85	3	19	45
ERR052777	P1.20,9	3-47	F3-1	39	7		1453	85	3	19	45
ERR052785	P1.20,9	3-47	F3-1	39	7		1453	85	3	19	45
ERR052776	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052762	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052742	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052772	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
ERR052748	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606726	P1.20,9	3-47	F3-1	39	7	3	1453	85	3		45
SRR606727	P1.20,9	3-47	F3-1	39	7	3	1453	85	3		45
SRR606739	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606725	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606730	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606738	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606721	P1.20,9	3-47	F3-1	5	7	3			3	19	45
SRR606732	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606729	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606723	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
SRR606720	P1.20,9	3-47	F3-1	5	7	3			3	19	45
SRR606737	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45

SRR606722	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
SRR606736	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606724	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606731	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606696	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
SRR606697	P1.20,9	3-27	F2-1	5	7	3	9		3	7	45
SRR606719	P1.20,9	3-47	F3-1	5	7	3			3	19	45
SRR606733	P1.20-1,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606699	P1.20,9	3-27	F3-1	5	7	3	9	14	3	7	45
SRR606734	P1.20-1,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606695	P1.20,9	3-47	F3-1		7	3	9	14	3	19	45
SRR606698	P1.20,9	3-27	F3-1	5	7	3	9		3	7	45
SRR606700	P1.20,9	3-27	F3-1	5	7	3	9	14	3	7	45
SRR606718	P1.20,9	3-47	F3-1	5	7	3			3	19	45
SRR606703	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
SRR606691	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
SRR606704	P1.20,9	3-47	F3-1	5	7	3	9	14	3	19	45
SRR606692	P1.20,9	3-47	F3-1	39	7	3	1453	85	3	19	45
M11 240262	P1.20,9	3-47	F3-1	39	7	3	587	654	3	19	45

Table A13. Alleles of major outer membrane antigens of 153 genomes of the ST-5 complex.

	Locus in reference	Sequence length	Genome position	Product	No. nucleotide differences	No. amino acid changes
A	NMAA_0029	1002	32717	Thiamine transport system substrate-binding protein	9	1
A	NMAA_0030	849	33812	Mechanosensitive ion channel	6	2
A	NMAA_0031	486	34686	Competence-damaged protein (CinA family)	10	2
A	NMAA_0032	1569	35273	peptide methionine sulfoxide reductase (putative pilin biogenesis)	12	2
A	NMAA_0033	1266	36986	probable signal recognition particle protein	31	4
B	NMAA_0299	1128	353893	ATP-binding protein involved in chromosome partitioning	51	8
B	NMAA_0300	522	355325	putative membrane-associated thioredoxin	1	0
C	NMAA_0321	1911	387460	pilin glycosylation protein pglD	18	5
C	NMAA_0322	1206	389418	pilin glycosylation protein pglC	85	17
C	NMAA_0323	1242	390723	pilin glycosylation protein pglB1	67	23
D	NMAA_0337	630	409695	GTP binding protein engB	6	0
D	NMAA_0338	624	410530	Cytochrome C	2	0
D	NMAA_0339	2016	411367	Cytochrome C biogenesis protein (resB)	20	5
D	NMAA_0340	1188	413375	Cytochrome C biogenesis protein CcsA	22	5
D	NMAA_0341	1065	414679	tRNA N6-adenosine threonylcarbamoyltransferase (EC 2.3.1.-)	17	2
E	NMAA_0635	759	749433	Ribosomal large subunit pseudouridine synthase F (rluF)	5	3
E	NMAA_0636	891	750216	ATP-NAD kinase	10	0
E	NMAA_0637	567	751126	None	5	2
E	NMAA_0638	777	751743	NADH(P)-binding	4	2
E	NMAA_0640	1041	753236	UDP-N-acetylenolpyruvoylglucosamine reductase (EC 1.3.1.98)	1	1
E	NMAA_0641	1380	754454	multidrug efflux protein	11	4
E	NMAA_0642	1152	756168	ATP phosphoribosyltransferase regulatory subunit	14	6
E	NMAA_0643	1299	757422	Adenylosuccinate synthetase	39	7
E	NMAA_0644	261	758839	None	42	18
E	NMAA_0651	648	762625	Adenylate kinase	1	0
E	NMAA_0653	972	764593	pfkB family carbohydrate kinase (BIGS: D-beta-D-heptose-7-phosphate kinase)	1	0
E	NMAA_0654	1236	765601	Cytosine-specific methyltransferase	1	1
F	NMAA_0886	1974	1007824	Putative phage tail fiber protein	45	20
F	NMAA_0888	288	1010855	Protein of unknown function (DUF497)	1	0
F	NMAA_0889	606	1011222	Caudovirales tail fibre assembly protein	8	3
F	NMAA_0890	297	1011806	None	6	2
F	NMAA_0891	210	1012099	None	1	1
F	NMAA_0892	648	1013947	Putative bacterial lipoprotein (DUF799)	1	1
F	NMAA_0893	369	1014591	None	1	0
F	NMAA_0894	672	1014969	Curli production assembly/transport component CsgG	4	2
F	NMAA_0895	720	1015805	short chain dehydrogenase	15	1
F	NMAA_0896	993	1016929	Flavin containing amine oxidoreductase	5	2
	NMAA_1212	1197	1424286	acetate kinase	3	2

G	NMAA_1235	1635	1459282	CTP synthase	39	1
G	NMAA_1236	1671	1461028	long-chain-fatty-acid--CoA-ligase	87	14
G	NMAA_1237	1104	1462769	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	17	12
	NMAA_1239	384	1464586	diacylglycerol kinase	3	1
	NMAA_1253	2586	1478878	aconitate hydratase	1	0
H	NMAA_1550	942	1828085	adhesin MafA2	6	2
H	NMAA_1551	1497	1829030	adhesin MafB2	26	10
H	NMAA_1553	669	1830920	None	176	75
H	NMAA_1557	678	1832666	None	184	74
H	NMAA_1562	675	1835104	None	145	67
H	NMAA_1563	204	1835834	None	23	8
H	NMAA_1564	345	1836359	None	12	5
H	NMAA_1566	1134	1837213	None	22	12
H	NMAA_1567	705	1838725	RTX iron-regulated protein FrpC	45 (plus 2 insertions)	26
H	NMAA_1568	276	1839879		2	1
I	NMAA_1788	624	2069801	Ribosomal RNA small subunit methyltransferase G	16	5
I	NMAA_1789	372	2070582	putative integral membrane protein	20	8
I	NMAA_1790	1167	2071324	Fusaric acid resistance protein family	21	4
Isolates from the second pandemic wave						
G	NMA0740	465	733754	None	28	11
G	NMA0741	1512	734253	Ubiquinone biosynthesis protein UbiB	20	5
G	NMA0745	216	737825	Putative periplasmic protein	1	1
G	NMA0746	1056	738066	Thiamine biosynthesis protein	7	4
G	NMA0747	1218	739274	Na(+)-translocating NADH-quinone reductase subunit F	17	1
G	NMA0748	594	740505	Na(+)-translocating NADH-quinone reductase subunit E	21	2
G	NMA0749	627	741102	Na(+)-translocating NADH-quinone reductase subunit D	3	2
G	NMA0750	777	741728	Na(+)-translocating NADH-quinone reductase subunit C	23	2
G	NMA0751	1233	742497	Na(+)-translocating NADH-quinone reductase subunit B	15	2
H	NMA1572	633	1476300	pyridoxamine-5'-phosphate oxidase	1	1
H	NMA1573	771	1477432	pseudouridine synthase	28	4
H	NMA1574	1542	1478271	integral membrane transporter	16	3
Isolates from the third pandemic wave						
H	NMA1588	921	1493137	tRNA pseudouridine synthase B	39	9
H	NMA1591	2823	1496471	type III restriction/modification system enzyme	23	11
H	NMA1592	1173	1499437	L-lactate dehydrogenase	62	1
H	NMA1593	447	1500899	Transcriptional regulator	1	0
H	NMA1594	1215	1501374	NifS-like aminotranfserase	17	3

Table A14. Nucleotide differences and amino acid changes between ancestral and introgressed alleles (for the most frequent allele in the sample).

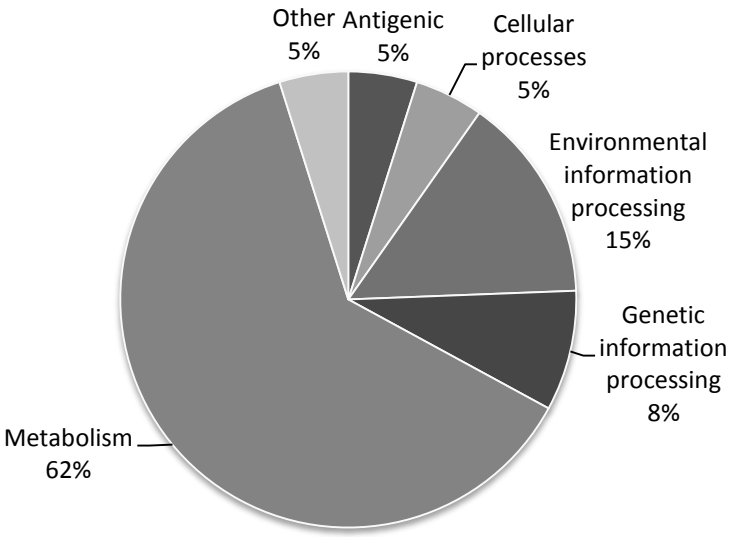


Figure A5. Functional characterisation of the putatively introgressed genes.

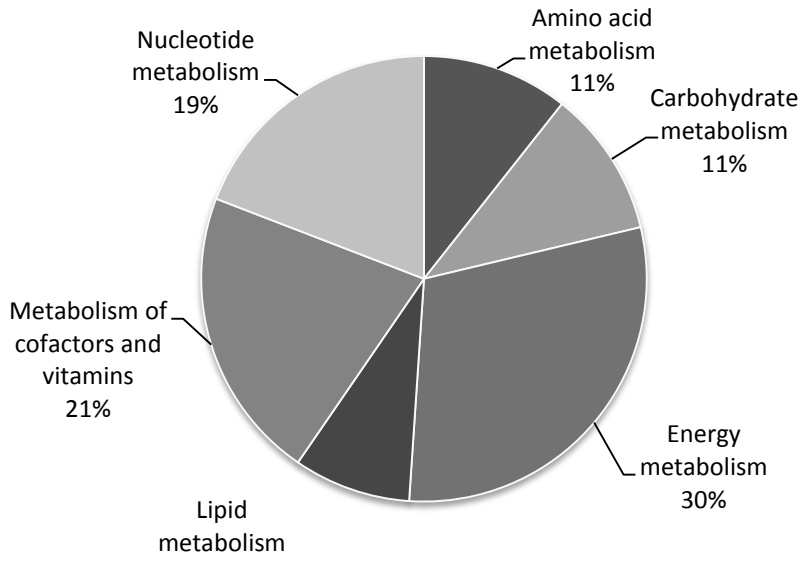


Figure A6. Metabolic functions of the putatively introgressed genes.

Appendix II.

Peer-Reviewed Publications by Author

Watkins, E. R. et al., 2015. Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. *Plos Pathogens*, 11(7): e1005034.

African Meningococcal Carriage Consortium, 2015. The diversity of meningococcal carriage across the African meningitis belt and the impact of vaccination with a group A meningococcal conjugate vaccine. *The Journal of Infectious Diseases*. pii: jiv211.

Watkins, E. R. et al., 2014. Contrasting within- and between-host selection shapes *Neisseria* antigenic Opa repertoires. *Nature Scientific Reports*, 4: 6554.

Bennett, J.S. et al., 2014. Identifying *Neisseria* species using the 50S ribosomal protein L6 (rplF) gene. *Journal of Clinical Microbiology*, 5, pp. 1375-81.

Daugla, D.M. et al., 2014. Effect of a serogroup A meningococcal conjugate vaccine (PsA–TT) on serogroup A meningococcal meningitis and carriage in Chad: a community trial. *The Lancet*, 4, pp. 40-7.

Basta, N. et al., 2013. Methods for Identifying *Neisseria* Meningitidis Carriers: A Multi-Center Study in the African Meningitis Belt. *Plos One*, 8, e78336.

African Meningococcal Carriage Consortium, 2013. Meningococcal carriage in the African meningitis belt. *Tropical Medicine & International Health*, 18, pp. 968-978.

Basta, N. et al., 2012. Age-specific prevalence estimates and risk factors for asymptomatic *Neisseria meningitidis* carriage in Bamako, Mali. *The International Journal of Infectious Diseases*, 16, e211.

Watkins, E.R. & Maiden, M.C.J., 2012. Persistence of Hyperinvasive Meningococcal Strain Types during Global Spread as Recorded in the PubMLST Database. *Plos One*, 7.

Buckee, C.O. et al., 2011. Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108, pp.15504–15509.