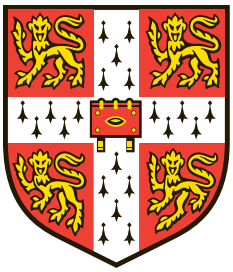
Host and pathogen genetics associated with pneumococcal meningitis

John Andrew Lees

Wellcome Trust Sanger Institute
Jesus College, University of Cambridge

July 2017



This dissertation is submitted for the degree of
Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed 60 000 words in length, as required by the School of Biological Sciences.

John Andrew Lees

July 2017

Acknowledgements

My main thanks have to go to Stephen Bentley, who has supervised me in exactly the way I would have wished. As well as being a steady hand on the tiller of my PhD, he has also shown me that it is possible to have a well-adjusted life, be kind to others and still be a great researcher. I will never forget all that he taught me (about Special Brew). Similar thanks must also go to my other supervisors. Jeff Barrett, who had no obligation to do so, fully accepted me into his research group. His group meetings and my conversations with him have really shaped this research and the way I think. Julian Parkhill has always been helpful and available for answering those tricky bacterial genomics questions no-one else could, and has given me many opportunities to present my work to other researchers (often in sunnier climes). Thanks too to Carl Anderson and John Welch, the other members of my thesis committee, whose freely contributed ideas have made this work better than it otherwise would have been.

Many collaborators have made this work possible. Nick Croucher is a pneumococcal master who I am lucky to know. Jukka Corander I am likewise lucky to know, and while explicitly involved with SEER I have felt his influence throughout this PhD. My Dutch friends Diederik, Matthijs, Philip, Arie and Bart have worked very hard on the unique dataset at the core of this thesis, and have been a pleasure to collaborate with. A special mention to Philip who even saw fit to invite me to his wedding. Paul Turner helped me throughout with the Maela dataset, and organised an excellent meeting in Cambodia without which chapter three wouldn't exist.

To the people who I have talked to about my PhD, thanks for your contributions (tangible and intangible) to everything I've done here: Sumana, Katie, Tom, Jeremy, Liam, Marcia, Theresa, Simon, James, Sophia, Becca, Leo, Izzy, Claire, Darryl and Alison, and all the other members of teams 81 and 143 past and present. Thanks too to all of pathogen informatics and the graduate office, who have doubtless helped me in many ways. Finally, I am very grateful to the Wellcome Trust and MRC for funding this research.

In the knowledge that this is the only page most readers of this document will look at, the pressure to be witty or memorable is greatest here. I guess you'll have to live with the Special Brew reference.

Summary

Host and pathogen genetics associated with pneumococcal meningitis

John Andrew Lees

Meningitis is an infection of the meninges, a layer of tissue surrounding the brain. In cases of pneumococcal meningitis (where the bacterium *Streptococcus pneumoniae* is the causative agent) this causes severe inflammation, requiring intensive care and rapid antibiotic treatment. The contribution of variation in host and pathogen genetics to pneumococcal meningitis is unknown. In this thesis I develop and apply statistical genetics techniques to identify genomic variation associated with the various stages of pneumococcal meningitis, including colonisation, invasion and severity.

I start by describing the development of a method to perform genome-wide association studies (GWAS) in bacteria, which can find variation in bacterial genomes associated with bacterial traits such as antibiotic resistance and virulence. I then applied this method to longitudinal samples from asymptomatic carriage, and found lineages and specific variants associated with altered duration of carriage. To assess meningitis versus carriage samples I applied similar analysis techniques, and found that the bacterial genome is crucial in determining invasive potential. As well as bacterial serotype, which I found to be the main effect, I discovered many independent sequence variants associated with disease. Separately, I analysed within host-diversity during the invasive phase of disease and found it to be of less relevance to disease progression.

Finally, I analysed host genotype data from four independent studies using GWAS and heritability estimates to determine the contribution of human sequence variation to pneumococcal meningitis. Host sequence accounted for some variation in susceptibility to and severity of meningitis. The work concludes with a combined analysis of pairs of bacterial and human sequences from meningitis cases, and finds variation correlated between the two.

Contents

List of Figures

List of Tables

Acronyms

AF allele frequency. 56, 57

AIC Akaike information criterion. 84, 85

ALF artificial life framework. 58, 71

AMP anti-microbial peptide. 21, 26

BAM binary sequence alignment/map. 111, 112

BAPS Bayesian analysis of population structure. 48, 54, 59, 61–63, 79, 180, 193

BFGS Broyden–Fletcher–Goldfarb–Shanno. 65, 66

CDS coding sequences. 138, 139, 142, 145

CFU colony forming unit. 135

CI confidence interval. 48, 118, 126

CMH Cochran–Mantel–Haenszel. 48, 54, 75, 79, 193

CNV copy number variant. 39, 108, 112, 124, 143

COG cluster of orthologous genes. 30, 46, 49, 55–57, 112, 120, 122, 188, 193

CPP closest phylogenetic-pairs. 118

CSF cerebrospinal fluid. 17–21, 23, 77, 108–110, 114, 117, 132, 134–152, 185, 186, 192, 196

CSV comma separated values. 176

d.f. degrees of freedom. 36, 56, 66

DSM distributed string mining. 55, 56, 71, 73

FWER family-wise error rate. 36, 67

GoNL The Genome of the Netherlands. 162

GOS Glasgow outcome score. 20, 118

GTR generalised time reversible. 58, 60

GWAS genome wide association study. 17, 33, 36–39, 41–52, 54, 55, 57, 63, 75, 77, 79, 81, 82, 98, 106, 108–112, 116, 119, 121, 125, 135, 147, 151, 152, 154–156, 161, 167–172, 183, 185–187, 189, 191, 193–197

H. influenzae *Haemophilus influenzae*. 20, 24, 155

HLA human leukocyte antigen. 43, 154, 175, 182

HMM hidden Markov model. 82, 84, 85, 95, 189

HPD highest posterior density. 147, 148

HRC haplotype reference consortium. 162, 163

HWE Hardy-Weinberg equilibrium. 158, 160, 162

ICE integrative conjugative element. 29, 31, 74, 91, 126, 129, 130

ICU intensive care unit. 156

IPD invasive pneumococcal disease. 18, 24

ivr inverting variable restriction. 32, 116, 117, 119, 131, 136, 146, 147, 252

JC Jukes-Cantor. 60

KC Kendall-Colijn. 60–62

L. monocytogenes *Listeria monocytogenes*. 20, 47, 58, 155

LD linkage disequilibrium. 30, 34–38, 42, 44–46, 49, 57, 73–75, 79, 88, 96, 98–100, 102, 161, 162, 164, 166, 177, 196

LMM linear mixed model. 39, 50, 86, 88–90, 93, 99, 102, 105, 120, 164, 187–189, 191, 193, 194, 246

LOD logarithm of odds. 33

LoF loss of function. 28, 39, 51, 124, 125, 128–131, 140, 141, 151, 191

LRT likelihood ratio test. 62, 66, 67, 90, 118, 164, 187

M. tuberculosis *Mycobacterium tuberculosis*. 43, 46, 47, 50, 128, 195

MAC membrane attack complex. 22, 27, 112

MAF minor allele frequency. 34, 36, 38, 39, 42, 55, 68, 71, 77, 96, 98, 99, 124, 128, 156, 158, 160–163, 165, 166, 170, 175–178, 180, 183

MCMC Markov-chain Monte Carlo. 116, 118, 132, 163

MDS multidimensional scaling. 63–65, 67, 68, 119, 176

MIC minimum inhibitory concentration. 93

MLST multi-locus sequence typing. 30, 47, 59, 61, 62, 108, 139, 143

MNP multiple nucleotide polymorphism. 110

MRCA most recent common ancestor. 58, 194

N. gonorrhoeae *Neisseria gonorrhoeae*. 66

N. meningitidis *Neisseria meningitidis*. 20, 21, 43, 46, 47, 99, 109, 135, 136, 138, 139, 142–146, 149, 150, 152, 155

NCD normalised compression distance. 60–62

NJ neighbour joining. 60–62

NT non-typable. 25, 31, 82, 85, 86, 90, 95

OR odds-ratio. 19, 45, 48, 49, 71, 72, 128, 165, 166, 170, 175, 178, 180, 183

OU Ornstein-Uhlenbeck. 118

pbp penicillin binding protein. 29, 49

PCA principal component analysis. 39, 115, 158, 178, 251

PCR polymerase chain reaction. 132, 147

PCV pneumococcal conjugate vaccine. 21, 31, 82, 195

PEER probabilistic estimation of expression residuals. 178–180

ply pneumolysin. 26, 195

QC quality control. 36, 109, 111, 155, 157, 160, 162, 163, 176

S. aureus *Staphlyococcus aureus*. 24

S. mitis *Streptococcus mitis*. 24, 58, 110

S. pneumoniae *Streptococcus pneumoniae*. 20–22, 24–28, 30–33, 43, 46, 48, 49, 56–58, 64, 67, 70, 71, 74, 75, 81, 86, 88, 95, 99, 105, 108–110, 121, 123, 135, 136, 138–140, 143–147, 149, 151, 152, 177, 179, 180, 185, 187, 189, 195, 196

S. pyogenes *Streptococcus pyogenes*. 9, 53, 56, 64, 70, 77–79, 187, 240

s.d. standard deviation. 113

SEER sequence element enrichment analysis. 53, 55, 59, 61–67, 69, 70, 74–77, 79, 81, 89, 102, 119, 120, 124, 176, 188, 193–195, 240

SFS site frequency spectrum. 56, 57, 124, 125, 141, 142, 157

SIR susceptible-infected-recovered. 195

SNP single nucleotide polymorphism. 31, 35–39, 46–49, 54–58, 61, 63, 70, 73, 75–79, 86–88, 91, 96–102, 108, 110, 111, 113, 120, 121, 137–139, 142, 145, 150, 156, 158, 161–166, 170, 175–177, 183, 188, 193, 194, 247, 248

SVM support vector machine. 115

VCF variant call format. 124, 176, 193

VEP variant effect predictor. 111, 128, 138

WHO World Health Organisation. 82

Chapter 1

Introduction

This thesis primarily concerns the application of a modern statistical genetics technique, the genome wide association study (GWAS), to determine how genetic variability of both host and pathogen contributes to invasive pneumococcal disease (particularly meningitis). Chapter 2 describes the issues with applying this technique to bacterial genomes, and a method I developed to overcome these difficulties. In chapters 3 and 4 I then applied this new technique, and others, to describe genetics associated with carriage duration (a pre-requisite for disease) and invasive disease respectively. Finally, in chapter 5, I performed a similar analysis of the association between host genetics and invasive disease, ending by jointly analysing both host and pathogen together in a genome-to-genome analysis.

These results are therefore tied together both through the disease studied, and the technique used to analyse genotype to phenotype associations. I start with an introduction to the disease: the clinical manifestations of bacterial meningitis, its cause and treatment are mentioned, with specific reference to the Netherlands where most of the new data analysed was obtained. As the focus is on pneumococcal meningitis I then give a background of pneumococcal genomics and pathogenesis. Though the results start with analysis of pathogen genomes, GWAS and its development is crucial throughout. This section of introduction starts with a short history of this method in the context of human genetics where it was first applied. The application to host susceptibility to infectious disease, while analysed last in this thesis, is discussed at the end of this first introductory section. I then go on to describe the application of GWAS to bacterial genomes.

1.1 Bacterial meningitis

Bacterial meningitis is a severe inflammation of the membranes surrounding the brain, the meninges, which is a response to the presence of bacteria in the cerebrospinal fluid (CSF) (Mook-Kanamori et al., 2011). This inflammation can compromise brain function, requiring immediate admission to hospital (Weisfelt et al., 2006). Other forms of meningitis

(viral, parasitic) are common, but are generally less severe than bacterial meningitis (Attia et al., 1999; Ginsberg, 2004). I also note early on two other terms related to this infection: bacteremia, which is bacteria in the blood, and invasive pneumococcal disease (IPD), which is bacteria in any normally sterile site, with the most serious disease caused when in the blood or CSF.

1.1.1 Diagnosis, epidemiology and treatment

Accurate diagnosis of meningitis is challenging (Attia et al., 1999; Brouwer, Tunkel & van de Beek, 2010) and requires clinical experience based on patient presentation as biomarkers, co-occurrences with other diseases and other routine patient data are uninformative (Khatib et al., 2016). Some symptoms such as headache, neck-stiffness, fever and altered mental state are usually required for a diagnosis of bacterial meningitis (van de Beek et al., 2006).

The 'gold-standard' for confirming bacteria as the causal agent is a positive culture from the CSF (Attia et al., 1999; van de Beek et al., 2004). Following successful culture, a range of microbiological techniques can be used to determine the organism (such as Gram staining, PCR or MALDI-TOF). While highly specific, the sensitivity of this technique relies on good antibiotic stewardship in the community, and a lumbar puncture (a sample of the CSF) being taken before treatment commences (Attia et al., 1999; van de Beek et al., 2006). In certain settings this may be impossible, and there is debate over situations where it may be dangerous due to increasing intra-cranial pressure (Hasbun et al., 2001; Winkler et al., 2002; Oliver et al., 2003).

It is also interesting to note the enormous effect of varying antibiotic use in the community and early lumbar puncture on the sensitivity of obtaining positive cultures, as this also affects the number of isolates which can be subjected to whole-genome sequencing using present methods. In the Netherlands, for example, antibiotic use in the community is well regulated and lumbar puncture is taken as standard upon admission to hospital and before antibiotic treatment commences: positive culture is obtained in 80-96% of suspected cases of bacterial meningitis (van de Beek et al., 2004; van de Beek et al., 2006) – an ideal location to set up a genomic study. When treatment occurred before lumbar puncture, positive culture rate lowered to 66-80% (Bohr et al., 1983; Nigrovic et al., 2008). As practices, and many other factors, vary by country, so do positive culture rates: in Brazil 67% (Bryan et al., 1990); UK 19% (Ragunathan et al., 2000); Kenya 1.7% (Knoll et al., 2009). In developing countries, where disease burden is highest, positive culture rates range from 0.8-19.4% (Levine et al., 2009).

The variability over the conditions which need to be met for a positive diagnosis leads to difficulty in obtaining accurate estimates for the prevalence of bacterial meningitis (Brouwer, Tunkel & van de Beek, 2010; Jafri et al., 2013). In European adults, the focus of this thesis, the best estimates for prevalence show that bacterial meningitis is now relatively

rare (prevalence of 0.94 cases per 100 000 per year in 2013-14) (Bijlsma et al., 2016).

In adults, defined throughout as >16 years, meningitis is more common in immunodeficient patients (Brouwer, Tunkel & van de Beek, 2010; Adriani et al., 2015). That is, people with other conditions which lower the efficacy of the immune system making them more prone to infectious diseases. For example HIV/AIDS, while rare in the Dutch population (incidence 0.13% in 2013 ('Monitoring Reports SHM', 2013)), represents 1% of patients diagnosed with bacterial meningitis (odds-ratio (OR) ~ 7.5). Pre-disposition to infection also occurs due to alcoholism, diabetes mellitus and splenectomy. For pneumococcal meningitis incidence increases with age: individuals >65 years are most at risk (OR ~ 6).

Once bacterial meningitis has been diagnosed, treatment is with broad-spectrum antibiotics administered two to three times a day (Tunkel & Scheld, 2002; Brouwer, Tunkel & van de Beek, 2010). After confirmation of the bacterial species causing the infection the antibiotic used may be changed to more effectively treat the infection, or in response to a measured or expected resistance. Meningitis progresses rapidly, with 47% of cases having <24 hours of symptoms, and all cases terminating within a week (Bijlsma et al., 2016). The disease usually rapidly worsens during this time, so rapid diagnosis and treatment is crucial for a favourable prognosis. In the Netherlands time from arrival to treatment is a median of four hours, and this delay has a major impact on the outcome of treatment (Aronin et al., 1998; Proulx et al., 2005).

The risks to the patient during the treatment is due to septic shock and acute inflammation of the meninges (Brandtzaeg, 1993). The former, more common in meningococcal meningitis, is due to blood infection (bacteremia) causing damage to organs which in turn leads to a dangerously lowered blood pressure (Pathan et al., 2003). This is the cause of the blotchy rash diagnosed by the 'tumbler test', and can lead to limb loss (perhaps the most common image of meningitis seen in the public sphere). Inflammation is caused by the innate immune response to bacterial infection, largely due to the action of neutrophils (Kolaczkowska & Kubes, 2013; Kruger et al., 2015). Even after death of the cell, the remaining material from the bacterium continues to promote further inflammation.

Inflammation of tissue is effective at, and usually essential for, clearing bacterial infection. However it is not good for the host if the tissue in question surrounds the brain. The expansion of tissue at the top of the cranium puts physical pressure on the brain itself, pushing it down towards the spinal column. The reduction of pressure of the CSF in the spinal column caused by a lumbar puncture can therefore in some cases increase this effect, so a CT or MRI scan of the head is first recommended in these circumstances to check for shift in position of the brain before this procedure is carried out (van de Beek et al., 2006). This pressure, if not relieved by treatment, leads to damage of the brain tissue, and death (Pathan et al., 2003; van de Beek et al., 2004).

In some circumstances it is therefore appropriate to seek to suppress the host immune system during treatment to limit the inflammation and damage to the brain it causes (de

Gans et al., 2002; Brouwer, Heckenberg et al., 2010). In the Netherlands, the use of such adjunctive therapy (dexamethasone) has been shown to reduce the rate of poor outcome (OR 0.54; 95% CI 0.39-0.73) (Bijlsma et al., 2016), and in particular reduce the number of patients who suffer long-term deafness or neurological effects after they have recovered from the infection (van de Beek et al., 2010; Brouwer et al., 2013). Of course, suppressing the action of the immune system when it is required to fight an acute infection may not be a good idea, and the trade-off between decreasing inflammation and decreasing the severity of infection must be considered. In immunocompromised patients such additional therapy is therefore inappropriate, nor is its use outside of the conditions where the randomised control trials of its efficacy took place (Molyneux et al., 2002; Mai et al., 2007).

These considerations also raise an interesting point about the strength of the host response, which causes the same trade-off between effectively clearing infection without causing extreme inflammation and damage to the meninges. If there is an intrinsic (most likely genetic) basis for strong immune response in some patients this would likely make them this group susceptible to contracting bacterial meningitis in the first place, but should meningitis occur they may suffer from a worse disease outcome. The converse would be true for naturally weaker immune responders.

The five-point Glasgow outcome score (GOS) is used to report the clinical outcome of cases: 5 is full recovery, 4 recovery with moderate disability, 3 recovery with severe disability, 2 persistent vegatative state, 1 is death (Jennett & Bond, 1975). Throughout, anything other than 5 is referred to as an unfavourable outcome. Sadly, despite advances in treatment and vaccination which have reduced incidence and disease severity, the serious nature of bacterial meningitis persists. In a recent Dutch study Bijlsma et al. (2016) estimated the case fatality rate in adults as 17% and unfavourable outcome in 38% of cases.

1.1.2 Causal organisms

Meningitis can be caused by CSF invasion from a wide range of bacterial species. In European countries the bacteria which most frequently cause meningitis are *Streptococcus pneumoniae* and *Neisseria meningitidis*, both of which are respiratory pathogens which normally exist as commensals in the upper respiratory tract of humans (Brouwer, Tunkel & van de Beek, 2010). In the past, serotype B *Haemophilus influenzae* caused the highest proportion of bacterial meningitis cases, but nationwide roll-out of an effective vaccine in a species for which serotype switching or replacement do not cause further disease have all but eliminated haemophilus meningitis (Schuchat et al., 1997; McIntyre et al., 2012). Recently an increase in *Listeria monocytogenes*, a food-borne pathogen, has been observed (Koopmans et al., 2017) which may be due to changes in use of antibacterial agents in the food-production chain (Kremer et al., 2017).

Vaccines have perturbed the populations of *S. pneumoniae* and *N. meningitidis*. In

the case of *S. pneumoniae*, first the 7-valent pneumococcal conjugate vaccine (PCV) and subsequently the 10- and 13-valent vaccines have immunised against the most invasive serotypes of *S. pneumoniae* in children, reducing the amount of carriage of in the population, and the amount of disease caused by these serotypes (Klugman, 2001; Knol et al., 2015). However, due to serotype switching and replacement allowing for vaccine escape, whether this vaccine has an overall effect on bacterial meningitis over longer time periods is yet to be determined (McIntyre et al., 2012) (section 1.2.3). For *N. meningitidis* there are now effective vaccines available against all invasive serogroups (A, B, C, W, X and Y) (Rouphael & Stephens, 2012), and though the B vaccine is expensive and therefore still has limited global coverage (Christensen et al., 2014), rates of meningococcal meningitis have fallen (McIntyre et al., 2012).

The route of infection varies depending on the species of bacteria, though in the majority of invasive cases the final stage is from blood to CSF (Mook-Kanamori et al., 2011). These respiratory pathogens are carried asymptomatically in the nasopharynx by a proportion of the population at any given time (Caugant et al., 1994; Hammitt et al., 2006). In a small number of cases commensal nasopharyngeal bacteria may invade the blood through a single cell bottleneck (bacteraemia) (Gerlini et al., 2014; Kono et al., 2016), then cross the blood-brain barrier into the CSF where they cause meningitis (Weisfelt et al., 2006). In some meningitis patients the CSF may be invaded directly due to CSF leakage or otitis media (Adriani et al., 2015), in which case the progression of bacteria after carriage is reversed: CSF to blood.

1.1.3 Immune response to pneumococcal meningitis

The host response to pneumococcal invasion mostly involves the innate immune system (Janoff et al., 1999; Paterson & Mitchell, 2006). Initial defence is through anti-microbial peptides (AMPs) such as lactoferrin and lysozyme which are secreted into mucosal surfaces and are active against a broad range of infectious agents (Brogden, 2005; André et al., 2015). Invading pneumococci are then detected by range of pattern recognition receptors (including the Toll-like receptors) which are primarily activated in response to their outer capsule but also other antigenic proteins such as pneumolysin (Paterson & Mitchell, 2006). The two most important signalling molecules in this process are TNF-α and IL-1 (Jones et al., 2005; Paterson & Orihuela, 2010), which are the first to be activated after infection (Takashima et al., 1997; Quinton et al., 2007). These receptors regulate the inflammatory response to infection (Koppe et al., 2012), causing recruitment of macrophages, which engulf and destroy the pneumococci (Janoff et al., 1999), and neutrophils, which as well as phagocytosis can release AMPs which cause inflammation and direct damage to the bacteria (Craig et al., 2009; Hyams et al., 2010).

This immune response is aided by the complement pathway, a system of over thirty

cascading proteins which aid the innate and adaptive immune responses (Walport, 2001a, 2001b). The pathway is activated in one of three ways (Serruto et al., 2010):

- Classical pathway – antibody recognition of of the bacteria, followed by binding of complement C1 to the pathogen's surface.
- Lectin pathway – recognises particular patterns of sugars on pathogen cell surfaces.
- Alternative pathway – constantly activated at low levels, positive feedback amplifies the response over time. Factor H binds to host cell surfaces to suppress the activity against self cells.

All three starting points end up with cleavage of C3 into C3a and C3b (Lambris et al., 2008). C3a triggers a pro-inflammatory response and enhances recruitment of immune cells to the region (through chemotaxis). C3b covalently bonds to the bacterial surfaces causing three further effects: making them more susceptible to phagocytosis (known as opsonisation); forming a C3 $\rightarrow$ C3a + C3b convertase on the cell surface, which amplifies the response through a positive feedback loop; cleavage of C5 to C5a and C5b near the cell surface. C5a fills a similar role to C3a and increases inflammation, whereas C5b causes a cascade of proteins through C6-C9. This results in formation of the membrane attack complex (MAC), which forms pores in the bacterial surface resulting in cell lysis and death.

Due to the rapid progression of disease, and the acute nature of symptoms, the adaptive immune system plays little role in fighting invasive infections (Paterson & Orihuela, 2010). However, in carriage, antibodies (immunoglobulins) produced by the adaptive immune system play a more important role. These antibodies increase opsonisation targeted phagocytosis, neutralise toxins, and inhibit adhesion of pneumococci to host tissue surfaces (Anttila et al., 1999; Janoff et al., 1999). In the nasopharynx the most abundant antibody type is IgA (Kett et al., 1986). This antibody type can bind *S. pneumoniae*, and through interaction with the complement pathway increases killing above the level of the innate immune system alone (Janoff et al., 1999). IgG plays a similar role, and is the type of antibody elicited by the pneumococcal vaccine against the capsule (McCool et al., 2002; Balmer et al., 2003; Croucher et al., 2017).

S. pneumoniae and humans have co-evolved, hence the pathogen has methods to evade each of the immune mechanisms discussed here (Lambris et al., 2008; Hyams et al., 2010). I discuss the mechanisms *S. pneumoniae* uses to evade these responses in more detail in section 1.2.2.

1.1.4 A nationwide Dutch cohort

The analysis presented in chapters 4 and 5 uses the MeninGene cohort: a prospective cohort running from 2006 onwards in the Netherlands (Bijlsma et al., 2016). The study

collects and combines data from cases of bacterial meningitis from across the Netherlands using a number of means. Firstly, the national reference laboratory for bacterial meningitis automatically receives blood and CSF isolates from about 85% of all culture-confirmed cases, along with limited metadata. This metadata allows the identification of adult cases along with the hospital the patient was treated at. The hospital is contacted, and the attending physician is invited to seek patient consent to fill out a report on their case. If the patient agrees to this, the physician also fills out more detailed information (treatment given, clinical course, neurological findings at discharge) which is submitted to the MeninGene database (http://www.meningitisamc.nl/en/inclusion-new-patient/meningene/). Bottles of wine in bespoke MeninGene wooden cases are sent from an AMC office to physicians each time they submit a patient, as an incentive to take part (fig. 1.1).

Figure 1.1: The incentive sent to physicians enrolling patients in the MeninGene study. Available in red or white.

To ensure the study focuses on the normal route of infection, patients are excluded if they have had neurosurgery or head trauma in the month prior to their meningitis, or if they have a neurosurgical device present in their central nervous system (for example a deep brain stimulation electrode). Patients who acquired bacterial meningitis nosicomially (occurring during a hospital stay, or within a week after) rather than in the community are also excluded. Around 200 cases not excluded for these reasons are added to the cohort each year, mostly during the winter.

The aim of this collection is to identify host and bacterial genetic variants which affect the susceptibility to and severity of bacterial meningitis. Consenting patients were genotyped (using human tissue collected during the lumbar puncture) and positive bacterial cultures whole-genome sequenced with the aim to link genetic variation to the extensive clinical metadata collected for the cohort. In this thesis I am primarily concerned with

pneumococcal meningitis: it was the largest and therefore most well powered part of the collection. Before describing the necessary background to this analysis I first consider the issues encountered when working with pneumococcal genomes.

1.2 Pneumococcal biology

In this section I first describe the basic biology of the pneumococcus, its pathogenesis and how genetic studies have increased our understanding of its evolution.

S. pneumoniae is a Gram-positive bacterium, only found in human hosts. It is normally a commensal in the nasopharynx, where it is challenged by host immune system (Paterson & Orihuela, 2010), other bacteria such as *H. influenzae* (Pericone et al., 2000; Lysenko et al., 2005) and *Staphlyococcus aureus* (Bogaert et al., 2004; Regev-Yochay et al., 2006)) and itself (Dawid et al., 2007; Cobey & Lipsitch, 2012). The closest relative to *S. pneumoniae* is *Streptococcus mitis*, a commensal with many, but not all, of the same virulence factors and a much higher intra-species diversity (Denapaite et al., 2010).

Pneumococcal carriage in the nasopharynx is asymptomatic. Estimates of carriage rates depend on the population, and the time of measurement (largely due to vaccination) but are high enough to suggest that most people will be exposed to the pathogen during their lifetime. Some examples of measured carriage rates in unvaccinated populations are: 66% in Kenyan children (Lipsitch et al., 2012); 68-84% in Karen infants on the Thailand-Myanmar border, 17-30% in Karen adults (P. Turner et al., 2012). In the Netherlands example estimates after vaccine introduction are: 69%-88% of children (Wyllie et al., 2014; Wyllie et al., 2016); 3-15% of adults (Spijkerman et al., 2011; Bosch et al., 2016). The duration of carriage ranges from a few days to many months (Abdullahi et al., 2012a; P. Turner et al., 2012), and generally decreases with age (P. C. Hill et al., 2010). Outside of the nasopharynx, *S. pneumoniae* infection can cause a variety of diseases. As well as causing IPD (meningitis and bacteremia), the pneumococcus can cause less serious diseases such as pneumonia and empyema (by entering the lungs), or sinusitis and otitis media (by entering the inner ear).

1.2.1 Importance of capsular serotype

One of the most important distinguishing factors between members of the pneumococcal species is their capsular type. The capsule is a polysaccharide structure which is bound to the outer pneumococcal cell wall (with the exception of serotypes 3 and 37 (Dillard et al., 1995; Llull et al., 1999)), and is important in most extra-cellular interactions. The capsule is immunogenic (AlonsoDeVelasco et al., 1995), defends against the host immune system (Hyams et al., 2010) and is likely required to survive in blood and so cause invasive disease (Kadioglu et al., 2008).

The different capsules are defined by their interaction with antisera (Lund & Henrichsen, 1978), though since the publication of the sequences of all known capsule loci by Bentley et al. (2006) the genome has increasingly been used to define the serotype of an isolate. This original publication consisted of 90 capsular types, however more are being discovered (Kapatai et al., 2017) and the current count stands at 98. Other than serotypes 3 and 37 the capsule locus consists of around 15 genes on the forward strand between *dexB* and *aliA* (Yother, 2011). Nucleotide variation within these genes, and structural variation of the locus leads to different antigenic serotypes.

The serotype is broadly correlated with the background genotype as the two are vertically inherited (Croucher, Finkelstein et al., 2013; Chewapreecha, Harris et al., 2014). However switching of serotype locus through recombination (horizontal inheritance) is possible (Croucher, Harris, Fraser et al., 2011), though usually happens within a serogroup (Croucher, Kagedan et al., 2015). Non-typable (NT) strains do not express capsule, either due to a complete or partial deletion of the capsule locus (Chewapreecha, Harris et al., 2014) or other surface proteins in its place (Salter et al., 2012; Park et al., 2012). They do not generally cause invasive disease, but are observed to be frequent donors of DNA in recombination events (Chewapreecha, Harris et al., 2014).

Serotypes have been shown to be associated with a number of important pneumococcal phenotypes, most notably invasive potential (Brueggemann et al., 2003). The exact mechanism is unknown, but capsular charge, thickness and expression seem to make a difference (Y. Li, Weinberger et al., 2013; Manso et al., 2014). Capsule type has also been shown to affect carriage duration (P. C. Hill et al., 2010; Abdullahi et al., 2012a; P. Turner et al., 2012), recombination frequency (Croucher, Kagedan et al., 2015; Chaguza et al., 2016), growth phenotype (Hathaway et al., 2012) and the ability to colonise the host (Trzciński et al., 2015).

Why over 90 different serotypes of pneumococci should be able to continue to coexist over long times when some have much higher fitness than others is puzzling (Lipsitch et al., 2009) – should the fitter serotypes not simply out-compete the less fit strains? Modelling work by Cobey and Lipsitch (2012) has suggested that serotype specific immunity working to stabilise competition, combined with acquired immunity to non-capsular antigens (section 1.2.2) reduces differences between fitness, allowing the continued prevalence of different serotypes and strains of *S. pneumoniae*.

1.2.2 Pneumococcal pathogenesis and immune evasion

As mentioned in section 1.2.1, the capsule is an important virulence factor, decreasing binding of complement (C3b) and IgG to the cell surface (Musher, 1992; Abeyta et al., 2003; Hyams et al., 2010). Its negative charge prevents phagocytosis (C. J. Lee et al., 1991), and reduces susceptibility to neutrophil extracellular traps (Wartha et al., 2007).

The pneumococcal genome encodes a variety of other proteins which directly interact with the host, mostly to enhance colonisation and avoid the host immune response (Kadioglu et al., 2008). Though the role of these antigens in colonisation and disease is known, whether sequence variation at these loci has an effect on pathogenesis in human disease remains unclear. Some antigens such as pneumolysin (*ply*) are essential for transmission and colonisation (Zafar et al., 2017; Rubins et al., 1998), whereas others such as *pspA* and *pspC* enhance virulence (Ogunniyi et al., 2007) but are not required for disease. These antigens can vary their sequence rapidly through recombination (Brooks-Walter et al., 1999; Iannelli et al., 2002; Lipsitch & O'Hagan, 2007; Croucher, Harris, Fraser et al., 2011) and are therefore highly variable. This mechanism may aid bacteria in evading detection by the immune system (Lambris et al., 2008).

In fig. 1.2 I review the immune system's response to pneumococcal infection (section 1.1.3), and the mechanisms the bacteria use to evade destruction. One of the first defences against pathogens is lactoferrin, encoded by the *LTF* gene. The core pneumococcal protein PspA binds lactoferrin strongly, preventing killing by this mechanism (Shaper et al., 2004; André et al., 2015). PspA has a further role in complement evasion, preventing deposition of C3b on the pneumococcal surface, and by inhibiting the formation of C3 convertases (Tu et al., 1999; Hyams et al., 2010).

The pneumococcal protein PspC also interacts with the complement system. PspC comes in two main forms, concordant with the genetic distances between their coding sequences, either with a choline binding domain or an LPXTG motif instead anchors them to the bacterial cell wall (Iannelli et al., 2002). PspC binds C3 using the choline binding domain, inhibiting this immune pathway in a similar way to PspA (Q. Cheng et al., 2000). On the bacterial cell surface, PspC can bind complement factor H (Janulczyk et al., 2000; Dave et al., 2001). This downregulates the alternative complement pathway in the vicinity of the cell, making the bacterial surface appear more like a host cell (Herbert et al., 2015).

To evade immunoglobulin, the pneumococcal genome encodes up to four proteases which cleaves the heavy chain of human IgA (*iga*/*zmpA*, *zmpB*, *zmpC*, *zmpD*) of which two (*zmpA* and *zmpB*) are core genes (Bek-Thomsen et al., 2012). This interaction inhibits the action of these antibodies on *S. pneumoniae*, primarily in the mucous membranes (Poulsen et al., 1996; Wani et al., 1996).

A number of other genes have been confidently implicated in pneumococcal virulence. Dlt, which causes D-alanylation of teichoic acids in the cell wall (Deininger et al., 2007) protects the cell against host AMPs (Kovács et al., 2006; Habets et al., 2012) and neutrophil extracellular traps (Wartha et al., 2007). *ply* is confined to the cell cytoplasm due to lack of a signal sequence, it is only released upon bacterial cell lysis. At low levels it can cause apoptosis, activate complement, and is pro-inflammatory (Kadioglu et al., 2002). Through inflammation this can increase shedding of *S. pneumoniae* during carriage, which is essential from transmission (Zafar et al., 2017). At higher levels pneumolysin forms

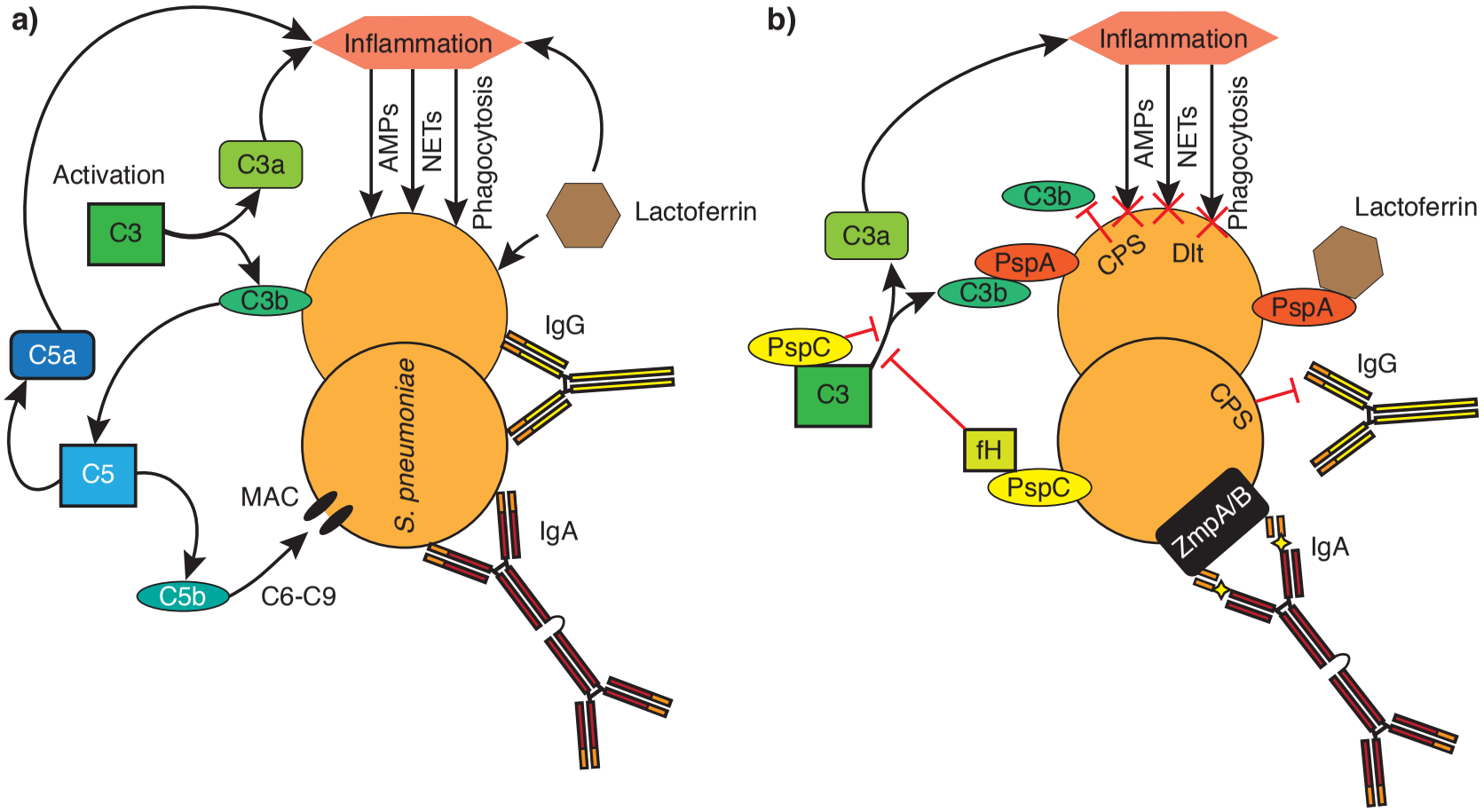


Figure 1.2: Interactions between the immune system and *S. pneumoniae*. **a)** Immune response to infection. Activation of the complement pathway by *S. pneumoniae* causes C3 to be cleaved into C3a and C3b. C3a increases inflammation; C3b binds to the bacterial surface, and leads C5 to be cleaved into C5a and C5b. Through activating the complement cascade C6-C9 this creates the pore-forming MAC. Lactoferrin (LTF) binds to the bacterial cell surface, binds iron needed for growth and cause inflammation. The antibodies IgG and IgA bind the bacterial cell. **b)** *S. pneumoniae* immune evasion. Capsule and D-alanylation of teichoic acids reduce effectiveness of inflammation mediated reponses, including IgG binding. PspA binds lactoferrin, and inhibits C3b binding. Surface bound PspC binds factor H, which inhibits C3 cleavage; secreted PspC binds C3 directly. The zinc-metalloproteases ZmpA and ZmpB cleave IgA.

pores in the membranes host cells, causing direct damage to the host tissues (Hirst et al., 2004; Harvey et al., 2011). LytA, an autolysin, was thought to enhance virulence through self-killing and release of pneumolysin (Berry & Paton, 2000), but has since been shown to be independently associated with virulence in a mouse model (Balachandran et al., 2001).

Other known virulence factors include metabolic genes such as *pflA* (Yesilkaya et al., 2009), adhesins allowing colonisation of host cell surfaces such as the Pht proteins (Khan & Pichichero, 2012; Plumptre et al., 2013) and *pclA* (Paterson et al., 2008), and the neuraminidases *nanA*/*nanB* which cleave sugars from host proteins contributing to adherence and immune evasion (S. J. King et al., 2004; Manco et al., 2006). An imaging-based localisation study has suggested that interaction between host factors pIgR and PECAM-1 with pneumococcal adhesins PspC and RrgA is involved in brain invasion during bacterial meningitis (Iovino et al., 2017).

Most of the studies confirming the effect of these proteins on virulence and the mechanism through which they do this have been by creating isogenic loss of function (LoF) knock-out mutants, which completely lack the protein of interest, and investigating variance in their ability to cause disease in a mouse (Ogunniyi et al., 2007). While this reveals interesting basic biology, and can be a useful approach for finding vaccine candidates which are immunogenic and required for invasive disease, the relevance of these virulence factors in clinical cases of disease (i.e. in humans) is currently unknown. More subtle variation within these genes, and its overall importance compared to other virulence factors is generally understudied, though some lab-based work has found capsular type to be more important than antigenic variation (Abeyta et al., 2003; Weinberger et al., 2009; Hyams et al., 2013) consistent with epidemiological studies (Weinberger et al., 2008; Weinberger, Harboe et al., 2011). Woehrl et al. (2011) showed that C5 cleavage affects the outcome of pneumococcal meningitis in a mouse model, but their sample size and statistical approach was insufficient to show similar relevance in clinical cases.

Complete knock-out of a gene is not naturally (or only rarely) occurring variation in the pneumococcal population due to the fitness cost it would incur. Rather than choosing candidate proteins and showing they have an effect on disease in an animal model, an alternative approach is to take a collection of clinical cases of disease and carriage and then agnostically test all naturally observed variants for association with each niche. Animal models can then lend further evidence to these results, and propose functional mechanisms. I discuss the power of this approach and its potential application to pneumococcal virulence in detail in sections 1.3 and 1.4.

Antibiotic resistance mechanisms

Since the introduction of antibiotics to treat *S. pneumoniae* infection, resistance has arisen to each treatment, in some cases through multiple mechanisms. The most effective

treatment in patients without allergies to penicillins are β-lactams, whose target is the penicillin binding proteins (*pbp*s). This disrupts cell-wall biosynthesis, leading to cell death and lysis. Variation of these target proteins, while at a general cost to fitness, gives rise to resistance to these antibiotics (Spratt, 1994b, 1994a).

Resistance to tetracycline and chloramphenicol are mediated through the *tetM* and *cat* genes respectively, which are carried on the integrative conjugative element (ICE) (Croucher et al., 2009). Erythromycin resistance can be gained through *ermB* which methylates the target ribosomal site, or the *mel*/*mef* efflux pump; both of these mechanisms are carried on transposable elements (Croucher, Harris, Fraser et al., 2011). Single base changes in *parC*, *parE* and *gyrA* cause fluoroquinolone resistance (Pletz et al., 2006), and single base changes in *rpoB* cause rifampicin resistance (Ferrándiz et al., 2005). Trimethoprim resistance is through the mutation I100L in *folA*/*dyr*, though it has been suggested other mutations in this gene can also contribute to resistance (Maskell et al., 2001).

As expected, there is an association between the amount of use of antibiotics and the levels of resistance in the population (Lipsitch, 2001; Samore et al., 2006). Similarly to the existence of multiple serotypes, the continued existence of both antibiotic resistant and sensitive pneumococci at a stable ratio over time is evolutionarily puzzling. In a simple model, when treatment is being applied the resistant bacteria should out-compete the sensitive, and when treatment is not being applied the sensitive bacteria should out-compete the resistant. More complex models proposing linkage with carriage duration modifying alleles (through altering carriage duration) or through including host structure and treatment frequency have been proposed to address this conundrum (Lehtinen et al., 2017; Cobey et al., 2017).

1.2.3 Population studies of *S. pneumoniae*

The first sequence of a pneumococcal genome was reported by Tettelin et al. (2001): the virulent TIGR4 (serotype 4) strain. It was found to be a singular circular chromosome of 2.16Mb, with a GC content of 39.7% encoding 2 236 genes. 84% of the genome was found to be protein coding. The authors noted that the genome contained a relatively high proportion of insertion sequence elements (5%), and the presence of a type I restriction-modification system. Various specificity domains invertible from upstream in the genome were found, which the authors hypothesised could allow rapid variation of the methylated motif, inhibiting DNA transfer between clonal strains. Despite its early discovery, it took another 13 years to fully describe the function and variation of this locus in the pneumococcal population (see below and section 4.3.2).

The publication of the TIGR4 genome was shortly followed by the avirulent (non-capsular) R6 strain (Hoskins et al., 2001). With more than one genome comparative

genomics within the species could be performed, using breaks in synteny to find differences in gene content or other variation between the sequences (Bentley & Parkhill, 2004). Lanie et al. (2007) added the sequence of the serotype 2 D39 strain, and were able to find different evolutionary rates in the three genomes, and further found that these mutations affected the expression of regulatory, virulence and metabolic genes. Further analysis of the sequence of a multidrug resistant clone using these techniques highlighted the role of mobile elements in the evolution of *S. pneumoniae* (Croucher et al., 2009).

In parallel to single complete genomes and comparisons between them, other studies based on the population genetics of the pneumococcus using a subset of the overall genomic variation were taking place. Early population genetic studies used the sequences of seven housekeeping genes to define a multi-locus sequence typing (MLST) scheme for *S. pneumoniae*, where a single base change in any of these genes defines a new allele, and any combination of alleles of the genes is a unique sequence type (Enright & Spratt, 1998). An advantage to this scheme is that a recombination event is more correctly counted as a single evolutionary change equivalent to a single base change, whereas counting the number of base changes itself would overestimate the distance from recombination events (Maiden et al., 1998). However, the designers of the scheme in *S. pneumoniae* later found it to be somewhat flawed: one of the chosen genes (*ddl*) is in linkage disequilibrium (LD) with the *pbp2b* gene, which is under diversifying selection due to its role in β-lactam resistance, driving excess diversity in *ddl* through hitch-hiking of mutations (Enright & Spratt, 1999).

Through the use of MLST schemes the genotype of *S. pneumoniae* could be defined for large numbers ($>$100) of isolates, allowing association between background genotype and traits such as serotype, resistance, virulence factors and recombination to be tested (Hanage et al., 2005; Hanage et al., 2009). It was not until the availability of high throughput sequencing that full length genomes of multiple isolates could be obtained, unifying the two approaches of studying bacterial genomics.

The importance of recombination and mobile elements

Hiller et al. (2007) performed one of the first multi-whole genome studies of *S. pneumoniae*, going beyond pairwise synteny comparisons between isolates. Using the whole genome sequences of 17 *S. pneumoniae* isolates, they aligned all 3 170 clusters of orthologous genes (COGs) and showed that there exists a 'core' of genes present in all isolates in a population, but that the majority of genes are 'accessory' and are only present in a subset of isolates. The mode frequency was presence in only one isolate (singleton genes). More recent estimates using a larger sample size of 616 genomes found 1 194 core genes from a total of 5 442 COGs (22%) (Croucher, Finkelstein et al., 2013).

The first large-scale study to fully unite techniques from both whole genome analysis

and bacterial population genetics sequenced 240 isolates from the PMEN1 serotype 23F multidrug resistant clone (variously referred to as Spain^{23F}, ST81 and ATCC 700669). Croucher, Harris, Fraser et al. (2011) were able to both find recombination events and map them to specific regions of the genome. These recombinations were found most frequently in antigens (*pspA*, *pspC* and *psrP*), prophage and a large ICE carrying drug-resistance conferring genes. They also found that the capsule locus itself is frequently involved in recombination events, leading to a switching of serotype; later work in a larger population quantified the selective constraints on serotype switching, finding most switches happen within a serogroup (Croucher, Kagedan et al., 2015). Overall, this showed that pneumococcal variation can occur on much shorter timescales than previously thought, allowing adaptation to environmental perturbations such as antibiotic use and vaccination.

The first high efficacy vaccine against *S. pneumoniae* was the seven-valent PCV, which offered protection against the seven most common disease causing serotypes in the US (Obaro et al., 1996; Klugman, 2001). Later vaccines have expanded this to ten and then thirteen serotypes. The vaccination of children successfully reduced carriage rates of these serotypes, and therefore disease. Since mass vaccination began the *S. pneumoniae* population has started to escape the vaccine through two mechanisms. At a population level, other serotypes not in the vaccine have less competition and are now found more frequently in carriage (Weinberger, Malley & Lipsitch, 2011). At a genomic level serotype switching to a non-vaccine type can directly aid vaccine escape (Croucher, Finkelstein et al., 2013).

The frequency and role of recombination in pneumococcal evolution has continued to be a theme in studies of population genetics. Subsequent work has quantified the length of recombinant DNA fragments, and found them most likely to be a mechanism to repair damaging mutations and guard against selfish mobile genetic elements rather than a mechanism to exchange accessory genes (Croucher et al., 2012; Croucher et al., 2016). A pneumococcal population can cease to be transformable due to a prophage inserting into the *comYC* gene, interrupting its competence machinery (Croucher, Hanage et al., 2014).

The role of single nucleotide polymorphism (SNP) variation compared to recombination in evolution differs by lineage (Croucher, Mitchell et al., 2013). In one of the first papers to move from analysis of a single lineage to a species-wide genomic analysis, Chewapreecha, Harris et al. (2014) calculated the ratio of recombination to mutation events r/m across the main lineages within the species: despite a similar number of mutations per site per year, they found estimates to vary between 0.06-0.25 depending on serotype. NT (unencapsulated) isolates had a significantly higher recombination rate than capsular strains (r/m = 0.3-0.35), and were more frequently donors of recombinant DNA. This suggested that NT serve as a reservoir for DNA, which is easily passed on without capsular polysaccharides providing steric hindrance.

Prophage sequence, viral DNA inserted into the bacterial host genome in the lysogenic

phase of replication, varies rapidly (Romero et al., 2009; Croucher, Coupland et al., 2014) and reduces host cell fitness (DeBardeleben et al., 2014). While in other species prophage can be found to carry 'cargo' genes which can advantage the host cell and partially offset the fitness reduction of carrying the phage, this is uncommon in *S. pneumoniae*. Exceptions are the phage MM1 which has been found to increase pneumococcal adherence (Loeffler & Fischetti, 2006), and the phage-carried virulence genes *pblB* and *vapE* (Romero et al., 2009).

The function of the inversions of the type I restriction-modification system, originally noted in the first pneumococcal genome sequence, could now be explained by these studies of population level variation. Despite the relatively rapid rate at which *S. pneumoniae* can vary its genome, the rate of variation in prophage inserted into pneumococcal genomes is much higher (Croucher, Coupland et al., 2014). The rapid phase variation of systems such as this inverting variable restriction (*ivr*) locus is therefore required to defend the host from foreign DNA. In parallel, *in vitro* work found that this phase variation also causes genome-wide methylation and transcriptional changes, which have been suggested to have knock-on effects on virulence (Manso et al., 2014; J. Li et al., 2016).

1.2.4 Within-host variation of *S. pneumoniae*

In the nasopharynx, evolution of *S. pneumoniae* is limited by a small effective population size (Y. Li, Thompson et al., 2013), which limits efficient selection or purging of mutations arising in the population . Combined with a single-cell bottleneck at transmission, likely due to the airborne route of infection (Gerlini et al., 2014; Kono et al., 2016), this means drift is the dominant evolutionary force within the host (Didelot et al., 2016).

Previously, it was thought that mutation rates in bacterial genomes were low, and as such there would be no change within a single host (Ochman et al., 1999). Through whole genome sequencing however, variation over the course of a single bacterial infection was found to exist (Mwangi et al., 2007; E. E. Smith et al., 2006). Additionally, many studies sequencing bacterial populations of various different species gave estimates of mutation rates three orders of magnitude higher than previously expected (Bryant et al., 2013; Morelli et al., 2010; Wilson et al., 2009). These new estimates of mutation rate were also supported by evidence that DNA sequence variation can occur over the course of a single infection (Eyre et al., 2013).

Such within-host variation has been shown to occur through a variety of mechanisms such as recombination (Kennemann et al., 2011), gene loss (Ehrlich et al., 2010; Rau et al., 2012) and variation in regulatory regions (J. Li et al., 2016; Manso et al., 2014; Marvig et al., 2014). The rapid variation that occurs in these regions of the genome can increase the population's fitness as the bacteria adapt to the host environment (Barrick et al., 2009; L. Yang et al., 2011), and potentially affect the course of disease (Young et al., 2012).

Previous studies in single patients have shown variation between strains even during the rapid clinical progression of bacterial meningitis (Croucher, Mitchell et al., 2013; Omer et al., 2011).

In mixed infections the main mechanism through which *S. pneumoniae* compete with each other is through the fitness effect of their capsule (Trzciński et al., 2015). A mechanism for intra-strain competition is the bacteroicins, encoded by a *blp* cassette (Dawid et al., 2007), though pneumococcal genomes are diverse in which combination of these bacteriocins they encode (Bogaardt et al., 2015). These produce peptides with antibacteriocidal activity against other strains, and the cell may also contain immunity proteins which protect against this (Moll et al., 1996). As there is a fitness defect from producing these toxins and anti-toxins this can lead to a number of different interactions affecting population dynamics (Miller et al., 2017). One example would be a 'rock-paper-scissors' interaction: bacteriocin producing bacteria are fitter than those not producing; those with the immunity protein are fitter than the bacteriocin producing bacteria; bacteria with neither are fitter than the immunity protein producing.

1.3 Association mapping in humans

Before going on to describe how GWAS can be applied to the problems in pneumococcal biology discussed in section 1.2, I first describe how this study design was first developed in human genetics and its application to host genetics affecting pneumococcal meningitis.

It has long been a goal of genetics to map heritable traits to the genes which affect them. Early attempts to map genetic regions to traits focused on simple Mendelian inheritance within families. Mendelian traits are those which are caused by a single, fully penetrant, allele. Dominant traits require just a single copy of the allele to mainfest the phenotype, whereas recessive traits require both the maternal and paternal chromosomes to carry the causal allele. The inheritance pattern within a family can determine whether a trait is fully Mendelian, or if the alleles are likely to display incomplete penetrance (there is a probability of an allele carrier having the trait, rather than certainty).

Given a family with a known pedigree where all members have been phenotyped for a trait of interest, if a candidate allele is genotyped one can then calculate the logarithm of odds (LOD) score which can be used to assess whether the allele co-segregates with the trait (Morton, 1955). If it does, then the allele is either associated with the trait or closely linked to an associated allele. How then, to choose the candidate allele? Some first attempts were based on speculation and known biology, but an approach able to test all genes was desired. By exploiting the linkage structure of the genome this became possible.

During meiosis, the maternal and paternal chromosomes undergo recombination, exchanging the order of alleles on each inherited chromosome. The recombination frequency

varies along each chromosome and is more likely at certain positions. Sites with a small physical distance between them are unlikely to have had a recombination event between them, and are inherited as a single piece of DNA. When averaged over a population, this results in high LD (which can be thought of as correlation between alleles at two different sites) between nearby sites, an approximately exponential decay of LD moving away from the site, and perfect linkage equilibrium (no correlation) between alleles on different chromosomes (Reich et al., 2001).

Botstein et al. (1980) were the first to map linkage across the human genome, finding linkage blocks which are inherited as a single unit and polymorphic loci which can be used to determine which of these blocks an individual has. Complementary DNA probes which genotype an allele can then determine the linkage block present. These 'linkage' studies were the first attempts at searching the whole genome for association with a trait of interest, and had a number of successes in rare diseases (Gusella et al., 1983; Siddique et al., 1991).

However, despite methodological improvements (Spielman et al., 1993), they suffered from a number of fundamental issues in association mapping for common traits. Firstly, they are designed to find associations between highly penetrant variants tending towards the Mendelian case, so for less penetrant variants quickly loses power. This is well suited for rare disease, but did not appear to be working for common diseases. A second, more practical limitation is that it is difficult to collect entire families of affected cases and genotype and phenotype every member of the pedigree – it would be much easier to collected affected cases and unaffected controls opportunistically.

Testing every linkage block in the genome for co-segregation with a trait leads to many thousands of tests, necessitating a heavy multiple testing correction burden (Lander & Kruglyak, 1995). Risch and Merikangas (1996) showed that under this multiple testing burden even a fairly penetrant common allele (OR = 2; minor allele frequency (MAF) = 13%) would require around 12 000 families to map the association. The lack of linkage based associations was providing increasing evidence that common traits were affected by multiple alleles with smaller individual effect sizes, this was good evidence that the linkage study was not the right design for discovering complex disease genes. In other animals linkage studies can still be a powerful approach, thanks to the ability to create and design crosses rather than having to rely on observed natural pedigrees. For the study of rare disease linkage studies can also be useful, as whole genome-sequencing has been able to increase their association mapping specificity (Ott et al., 2015).

In the same paper, Risch and Merikangas (1996) calculated that a population study would only need 640 samples to find the association. It had previously been proposed that by sampling affected and unaffected individuals from a population, association between an allele and the trait could be found by simple correlation. Population structure was known to confound such studies, as alleles are present at different frequencies in different populations due to their demographic history (for example, passing through a population

bottleneck can cause alleles to be lost from the new population, and previously rare alleles to become common). Therefore if there are uneven numbers of cases and controls from different populations, allele frequency will appear to associate with case status. However, sampling cases and controls from a single population can be used to address this issue (Hirschhorn & Daly, 2005).

The real barrier to the proposal of performing population association studies of common diseases was therefore the lack of knowledge about the human genome, and of human genetic variation (Hirschhorn & Daly, 2005). The low throughput resequencing available at the time was also an issue, and limited sample size and the number of markers tested. 'Candidate gene' studies had to guess a gene or region which may be associated with the trait, and then performed an analysis of correlation between the trait and polymorphisms in the gene. This initial guess was difficult to make, and not conducive to discovering association of genes where little prior biological knowledge is available. Despite well-known statistical guidelines for reporting associations (Lander & Kruglyak, 1995), many candidate gene studies did not follow the correct multiple testing correction, leading to very few results replicating in independent samples (Altshuler et al., 2008).

Such results have appeared between candidate genes and susceptibility to bacterial meningitis (Khor et al., 2007; Woehrl et al., 2011), however I do not review them here. Instead I quote a line from the review of Brouwer et al. (2009), whose meta-analysis was unable to confirm any of the published results: 'Results of the 44 case–control studies were hampered by methodological flaws. First, and most importantly, sample sizes were inadequate, preventing robust conclusions on the influence of the studied genetic variants . . . control populations were heterogeneously selected and often not matched for age and sex . . . quality control procedures for DNA extraction and genotyping were rarely done . . . most studies that assessed multiple polymorphisms did not correct for multiple testing'. It is perhaps surprising that over twenty years later similar mistakes are still being made, and published (Stessman et al., 2017; Barrett et al., 2017).

1.3.1 Genome-wide association studies

A better design for genetic mapping with a common trait was therefore a population study using all polymorphisms present in the population: this could test, in an unbiased manner, every gene and region of the genome for association with the trait (Hirschhorn & Daly, 2005; Altshuler et al., 2008). The first steps towards this goal were the sequencing of the human genome (Lander et al., 2001), and the genome-wide discovery of SNPs it facilitated (Sachidanandam et al., 2001). These efforts led to an improved mapping of linkage blocks in globally distributed populations, and the design of arrays which could genotype hundreds of thousands of SNPs in a high-throughput manner, with the SNPs chosen to capture variation across the entire genome through LD (International HapMap

Consortium, 2005). Using whole-genome sequencing these population maps of variation were later expanded in terms of variant frequency range, variant types, number and diversity of samples (1000 Genomes Project Consortium et al., 2012).

Using these advances Klein et al. (2005) performed the first GWAS in 96 cases and 50 controls, mapping an association between age-related macular degeneration and the *CFH* gene – narrowing the association to a region of a chromosome known from linkage based studies to a single gene, and showing this method could be used to understand complex trait genetics. The first large scale GWAS was the Welcome Trust Case-Control Consortium, which was performed on seven common diseases, using 2 000 cases for each and a shared set of 3 000 controls (Burton et al., 2007). The study was particularly successful in finding genetic loci associated with autoimmune disorders, and also set out the methodology for future studies.

I refer here to binary traits of interest (cases and controls), which can easily be generalised to multi-level or continuous traits. First, cases and controls are collected and genotyped together on arrays. The arrays have green and red fluorescent probes which bind to one of the two possible alleles (A and B, with B the effect/minor allele here) at each SNP location, so by clustering based on intensity of each colour samples can be called as AA, AB or BB. Crucially these SNPs were chosen to be roughly equally and densely spaced across the genome, be common (MAF $>5\%$) in the study population, and 'tag' nearby untyped variants through LD. This design later allowed for the incorporation of population level variation to gain greater information at untyped sites using genotype imputation.

After careful quality control (QC) of the genotype called on the samples, a test for association is performed independently at every site. The test for association is, at its simplest, a 3x2 contingency table between the genotypes and phenotypes with significance tested using a χ^2 test with two degrees of freedom (d.f.). Regression of the phenotype against the genotype gives similar results, but can also include covariates (often age and sex) or priors in the association. Most studies test for additive effects, where each extra copy of the effect allele has an equal effect on the phenotype. Recessive effects can be modelled by instead combining the AA and AB genotypes, and dominant effects by combining the AB and BB genotypes. A p-value against the null hypothesis of no association is generated at every site, and plotted on a log-scale against physical location on a 'Manhattan plot'. Association of a locus is usually declared when $p < 5 \times 10^{-8}$, which is a family-wise error rate (FWER) of 0.05 with a Bonferroni correction for multiple testing using the number of independent linkage blocks as the number of multiple tests. Figure 1.3 shows the overall study design of a GWAS based on these methods, and the methods are described in more detail when applied to the MeninGene cohort in chapter 5.

With the main technological limitations overcome, and the fact that a simple regression model works well for the analysis of GWAS data, finding more associations has mostly

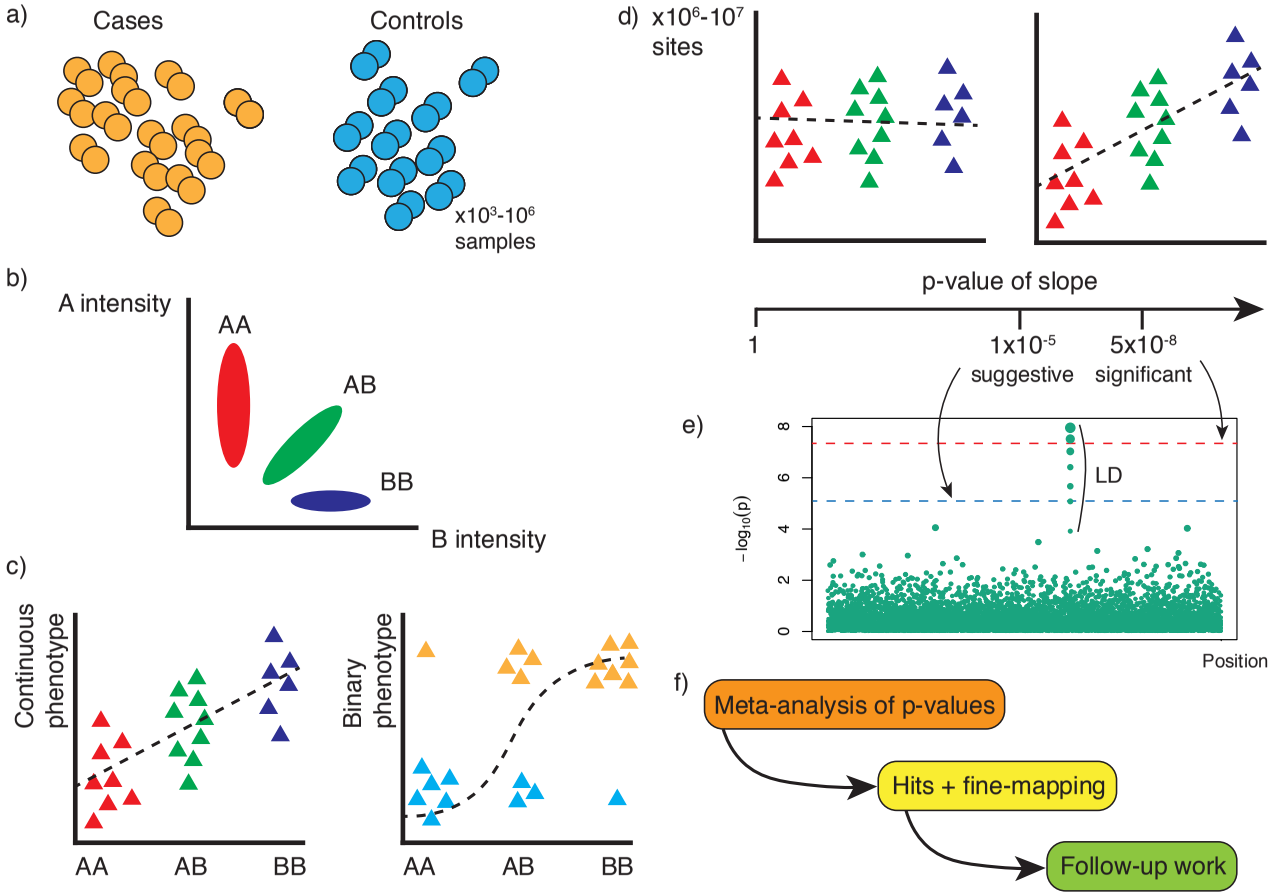


Figure 1.3: An overview of GWAS study design. **a)** Collect cases and control samples; discovery power increases with number of samples. Most successful GWAS studies in humans use at least 10^3 samples, with more recent studies reaching of the order of 10^6 samples. **b)** Pool samples together and genotype at every site. A SNP on a microarray is depicted here, but whole-genome sequencing is equally applicable. **c)** Perform a regression of genotype against phenotype, linear if continuous (left), logistic if binary (right). **d)** Calculate the p-value of the slope for every site. Stronger associations have smaller p-values. **e)** Plot all p-values on a log scale versus their position, those exceeding multiple testing adjusted significance thresholds are 'hits'. LD between nearby sites will make multiple variants appear associated in a locus, though usually only one is causal. **f)** Hits should be replicated in an independent study through meta-analysis. These form the foundation of further work and validation.

been a case of increasing the number of samples. The discovery power of GWAS is a function of MAF, effect size and sample size – an increase in any of these increases power. As MAF and effect size are determined by underlying biology and population history, increasing the number of cases (and controls, though as the number of GWAS studies has increased more samples have become available to use as shared controls) is how GWAS study design has progressed from the first successes. Meta-analysis, where separate GWAS studies are pooled in a combined analysis, both increases discovery power and makes discoveries less likely to be artefacts due to technical noise in a single cohort (Altshuler et al., 2008; A. Franke et al., 2010). Some studies, to minimise cost, genotype only their top p-value markers in a second cohort using 'MASSARRAY'. This uses mass spectrometry to genotype a small number of specifically designed probes, so unlike running a whole genotyping array this only allows validation at the chosen markers. Of course, evidence from an orthogonal approach (functional analysis in an animal model for example) that relates an associated locus/gene to the phenotype will also increase confidence that the association is not an artefact of the specific cohort. A meta-analysis can be performed using just the p-values, effect size and direction and sample size at each site (known collectively as 'summary statistics') and does not require the full genotype of every sample. By sharing this data at each incremental increase in sample size, GWAS consortia have greatly increased the number of loci associated with a range of common diseases (Liu & Anderson, 2014; de Lange & Barrett, 2015).

Due to LD between nearby variants, signals of association are not found to a single SNP. Usually a set of between a few and hundreds of genotyped or imputed SNPs in the region of the signal will be associated with the trait (albeit with different p-values), so interpretation of the chain of causation from genetic variant to effect on phenotype is not simple. However, with enough samples methods do exist to assign a probability of being the causal variant (Spain & Barrett, 2015). In coding regions knowledge of the codon table can predict the effect on proteins of genetic changes (McLaren et al., 2010), and analysis of conversation of amino acids across species can predict the effect of amino acid changes on protein function (Ng & Henikoff, 2003; Kircher et al., 2014) which can help fill in more of the chain of causation. In some cases an associated locus may contain multiple causal variants, in which case conditional analysis can be used to determine which variants are independently associated.

GWAS in humans has gone from strength to strength, and as of June 2017 2 500 studies have found over 40 000 significant associations (MacArthur et al., 2017).

Methodological advances

The issue of population structure driving association effects was initially dealt with by sampling participants from a single country, and excluding individuals found to have

divergent ancestry (which given their genotype can be determined). A. L. Price et al. (2006) showed that performing principal component analysis (PCA) on study participants' genotypes, and then including the leading principal components as fixed-effect covariates in the association model could correct for this effect without as much power loss as completely excluding samples. By instead including the kinship (relatedness) matrix as random effects in a linear mixed model (LMM) type II error rate can be controlled when combining samples of any ancestry, maximising sample size and discovery power (A. L. Price, Zaitlen et al., 2010). Subsequent computational improvements and approximations have made it possible to apply this to the millions of regressions needed when using imputed variants (Lippert et al., 2011; Zhou & Stephens, 2012; Loh et al., 2015).

The availability of lower cost high throughput whole-genome sequencing has not increased discovery power for common variants or enhanced the ability to fine-map association signals. Money is best spent on obtaining many samples at the lower price-point of genotyping arrays, rather than many sites. Whole-genome sequencing instead increases the range of the allele frequency spectrum which can be tested for association with a trait.

The design of GWAS genotyping arrays and tag-SNPs, when combined with improved imputation panels and techniques, has been very successful in discovering loci down to lower MAFs than originally thought possible (1%) (de Lange & Barrett, 2015; de Lange et al., 2017). In the case of uncommon (0.1% <MAF <5%) variants, which are less well tagged and are therefore poorly imputed (The Genome of the Netherlands Consortium, 2014), and rare variants (MAF <0.1%), which are not even present at a population level in current reference panels, direct sequencing of these variants can help find new associations. More complex rare variants, such as copy number variants (CNVs), long insertions or deletions (INDELs) and structural variants, which were not included on genotyping arrays can be tested using whole-genome sequencing. Very rare variants appearing in a single sample (singletons) or two samples (doubletons) are the mode variant frequency in the human genome (1000 Genomes Project Consortium et al., 2012). Without time for them to become common in the population, strong selection may not arise against their potential fitness defects. They may therefore play a role in determining complex trait phenotypes. These variants are challenging to genotype from low coverage sequencing data as population level variation cannot inform the genotype call, and they are difficult to distinguish from sequencing errors (particularly at heterozygous sites). In the future, cheaper high coverage whole genomes will help deal with some of these challenges.

While there is not enough information at a single site to perform a regression against the phenotype, by grouping sets of these variants by their predicted functional effect sufficient power to perform association tests can be reached (S. Lee et al., 2014). Rare variants can be grouped for example by LoF of a gene or any element in an entire pathway, or within a region around a gene or haplotype. The simplest association test of these

variant sets is a burden test, which works best when the variants are causal and their effect sizes are in the same direction. More complex tests relaxing these assumptions, such as `SKAT-O`, are available (Wu et al., 2011; S. Lee et al., 2012). It therefore has been possible to discover the role of rare variation in common auto-immune disorders such as type II diabetes and inflammatory bowel disease using whole genome sequencing and newer methods (Fuchsberger et al., 2016; Luo et al., 2017).

As well as expansion in terms of genotyping space, recent efforts have been made to expand the phenotype space. The compilation of large biobanks containing hundreds of thousands of genotyped individuals each with thousands of phenotypic measurements (usually through electronic health records) has inspired the creation of 'PheWAS' (phenome-wide association study), in which the focus is instead on variants and the spectrum of diseases and traits they are associated with (Denny et al., 2013; Bush et al., 2016). By association of many diseases in the same set of individuals, the overlap in genetic architecture and co-heritability between phenotypes can be assessed (Ge et al., 2017).

By exploiting the unidirectional causality of genetics on phenotype, the causality of association between phenotypes can be determined using Mendelian randomisation (Davey Smith & Hemani, 2014). Current efforts are being made to exploit the known hierarchical relation between phenotypes to increase the power of PheWAS studies given their increased multiple-testing burden, and also incorporate self-reported phenotype information (Cortes et al., 2017).

1.3.2 Heritability

Heritability is a classical concept in quantitative genetics which represents the amount of variation in a trait which can be ascribed to genetics (and is therefore inherited between generations) versus other environmental factors (Lynch & Walsh, 1998). Fisher (1919) was the first to reconcile Mendelian inheritance patterns, which are fully penetrant, with normal variance about the mean observed in most human traits by proposing multiple inherited genetic mechanisms each with their own variance components. Wright (1920) applied this theory to guinea pig coat patterning, and so defined heritability H^2 as the proportion of variance in a phenotype σ^2_{P} which can be attributed to genetics σ^2_{G}, compared to the environment σ^2_{E}:

$$\sigma^2_{\mathrm{P}} = \sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{E}}$$
$$H^2 = \frac{\sigma^2_{\mathrm{G}}}{\sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{E}}}$$

The proportion of heritability which can be ascribed to additive variation σ^2_A as opposed to dominant σ^2_D or epistatic σ^2_I interaction is known as the narrow-sense heritability h^2:

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_I$$
$$h^2 = \frac{\sigma^2_A}{\sigma^2_P}$$

If a trait is not heritable then one will not be able to find genetic variation associated with it, but even significant evidence for small but non-zero heritability may have additive genetic variants associated. Heritability does not however tell us about the distribution of effect sizes of associated variants, nor is it constant between populations (Visscher et al., 2008). Heritability is therefore an important parameter in estimating the power of GWAS, and can also be used to describe the proportion of overall variance described by sets of variants in the genome.

Before the availability of sequencing, known genetic relationships could be exploited to determine H^2. For example, monozygotic twins have an identical genetic sequence, whereas dizygotic twins share only half of their sequence. However both cases share a similar environment, so by comparing the correlation between phenotype of these two cases with the overall phenotypic variance then H^2 can be calculated (Lynch & Walsh, 1998).

The availability of genomic data has allowed calculation of the narrow-sense heritability h^2 directly from genetic variation detected in unrelated individuals. Taking the significantly associated variants from GWAS and regressing them against the phenotype to calculate the variance explained (R^2) directly gives the heritability. However, these estimates are systematically lower than estimates from twin studies across a range of human traits, leading to the coining of the phrase 'missing heritability' (Manolio et al., 2009; Eichler et al., 2010). Various reasons that heritability is being missed have been proposed (untyped rare variants, structural variants, non-additive inheritance such as epistasis), but the inclusion of weak effects which do not reach significance in GWAS has been shown to be important (S. H. Lee et al., 2011).

To include all variants, a regression could be performed between all genotyped or imputed sites and the phenotype to calculate the variance explained (so $h^2 = R^2$). However the number of variants vastly exceeds the available number of samples, meaning this regression cannot be directly performed. By instead assuming that effect sizes of genetic variants on the trait are normally distributed with a mean of zero and variance of $\frac{\sigma^2_G}{m}$ (where m is the number of markers) a linear mixed model can be fitted by restricted maximum likelihood to determine h^2. In analogy with classic methods of heritability estimation, this uses the kinship (amount of shared sequence) estimate from the sequence to determine the relatedness of samples in the study. This is known as the 'GCTA' model (J. Yang,

Lee et al., 2011) and has been successfully used to narrow the gap between heritability estimates for human height from genomic and twin studies (J. Yang et al., 2010). This technique has been shown to be robust to deviations from the model assumptions, with the exception of varying LD between predictors (Speed et al., 2012), genotype certainty and inclusion of predictors across the MAF spectrum. These issues which have been addressed in recent advances by Speed et al. (2017). Including sets of predictors in this model, known as 'genomic partitioning', has been shown to fulfil the desire to attribute part of the overall h^2 to selected pathways and/or regions of the genome (J. Yang, Manolio et al., 2011).

1.3.3 Host susceptibility to infectious disease

While GWAS has enjoyed great success at finding loci associated with auto-immune disorders and anthropometric traits such as height and body-mass index, far fewer associations with susceptibility to infectious disease have been found (Newport & Finan, 2011; Ko & Urban, 2013). Twin-study and epidemiology based estimates of H^2 have convincingly shown that there is a genetic component to host susceptibility to a range of infectious diseases (Jepson, 1998; Burgner et al., 2006), so why are associations hard to find?

Firstly, candidate gene studies ensnared the study of infectious disease association studies for a number of years, without producing many reproducible findings (Abel & Dessein, 1997, 1997; Brouwer et al., 2009). When GWAS became feasible, infectious disease phenotypes began to be used. However, potential variability in exposure to the pathogen being studied (in some cases making it difficult to find equally exposed controls), difficulty of determining the exact pathogen causing a disease and lack of funding leading to lack of samples have been suggested as reasons why associated loci have been hard to find (Chapman & Hill, 2012).

An interesting debate continues over the genetic architecture of infectious disease susceptibility (A. Hill, 2012). In human history, susceptibility to infectious disease (especially in childhood) would be associated with a serious fitness disadvantage, given the lack of effective treatment. Given a sufficient effective population size these damaging variants would therefore be purged from the population. However, autoimmune disease would have had a small fitness cost, and recent changes in environment combined with population bottlenecks allowing relatively rare alleles to become common may explain the relative ease of finding these GWAS hits (Amos & Hoffman, 2010; Schraiber & Akey, 2015). It has therefore been suggested that common variants which explain infectious disease susceptibility may not exist, with variation in susceptibility caused by single variants unique to each patient (monogenic cause) (Casanova, 2015).

Most likely, as in other complex traits, both modes of causation are possible in some proportion. In bacterial infections, Zhang et al. (2009) performed a successful common variant GWAS on leprosy susceptibility, and common variants in the *ASAP1* gene and

the human leukocyte antigen (HLA) have since been associated with susceptibility to *Mycobacterium tuberculosis* infection (Curtis et al., 2015; Sveinbjornsson et al., 2016). Similar results have been found for viral and parasitic infections (Fellay et al., 2007; Jallow et al., 2009; Khor et al., 2011).

Host genetics of meningitis

Meningitis has been a relative success story for infectious disease GWAS. Davila et al. (2010) performed one of the first successful studies on a bacterial infection, and found variants in the *CFH* region to be associated with susceptibility to meningococcal meningitis in 1 443 European children. In a similar manner to *S. pneumoniae*, *N. meningitidis* is known to bind factor H with fHBP to inhibit activation of the alternative complement pathway (McNeil et al., 2013). The minor alleles were found to be protective, so the authors hypothesised that these less common forms of fH were more weakly bound by fHBP, increasing the effectiveness of the host immune response.

Rautanen et al. (2016) performed a GWAS in 542 cases of pneumococcal bacteremia in Kenyan children. They found variants on chromosome 17 in a long intergenic non-coding RNA gene (AC011288.2) to be associated with doubled susceptibility to invasive disease. The variants are specific to African populations so would not be found in a GWAS of a European population. Expression of these gene was found only in neutrophils, a cell type involved in the innate immune response to *S. pneumoniae* infection.

Finally, Davenport et al. (2016) assayed both genomic and transcriptomic variation in 384 British adults with sepsis. They found two classes of gene expression as response to infection, activated depending on whether the patient was immunodeficient or not. They were then able to map genetic variants which affected these transcriptional networks, defining sepsis related eQTLs.

1.4 Association mapping in bacteria

The trend of scaling from a single genome to represent a bacterial species, to performing comparative genomics between two genomes to analysis of populations of whole genomes was seen not just in *S. pneumoniae* (section 1.2.3), but most pathogens deemed important enough to undergo the first sequencing attempts. There has been increasing availability of whole-genome sequence data from populations of bacteria along with phenotypes such as antibiotic resistance, virulence and host specificity. A natural question is therefore which pathogen variation, if any, contributes to these traits. The move to whole genomes of populations occurred well after GWAS had been established in human genetics, yet the first bacterial GWAS only started to appear years later. Falush and Bowden (2006) were the first to formally address this disparity. There are three main issues which frustrate the

simple study design so successful in the study of human complex traits: strong population structure, greater variation of the pan-genome and low sample sizes.

1.4.1 The effect of population structure

The strong population structure of bacteria is both a technical limitation to be addressed by the association model, and a fundamental limitation to the resolution of association mapping. Humans are diploid eukaryotes which recombine during meiosis every generation. Over a population, this shuffling of alleles makes separate variants independent, with the exception of nearby variants where LD is only partially broken by meiosis causes some level of correlation. Bacteria are haploid prokaryotes, where between generations the entire chromosome is clonally copied to the daughter cells, meaning all sites across the entire genome are perfectly correlated. If a set of mutations are introduced de novo over time, one of which is causal for the phenotype of interest, a naive association will find the entire set of mutations to be associated with the phenotype (i.e. the causal mutation, and the genetic background). While this is locally true around causal variants in the human genome, the exponential LD decay still allows mapping the association to a single region. However in bacteria LD extends across the entire genome and does not quickly decay over the chromosome (P. E. Chen & Shapiro, 2015; Earle et al., 2016), so the set of associations will also be genome-wide, preventing mapping of the causal association to a specific region.

Another way to understand the issue of population structure is through the more bacteria-centric idea of phylogeny (fig. 1.4). If a mutation which is causal for a phenotype has arisen on an ancestral branch, the descendants will be more likely to have the phenotype and the variant will be positively associated with the phenotype. However, any other mutation on that branch (potentially thousands, depending on the branch length) will appear equally associated. Again, these associations will not map to a single region of the genome.

Such associations, variants correlated with a specific genetic background and the phenotype, are known as ‘lineage’ associations. The best bacterial GWAS can reasonably hope to achieve with such associations is to identify them as such (and not treat them as potentially causal), and prioritise sets of associated variants for study by other means. Alongside the formal use of GWAS, genomic epidemiology studies have investigated the properties of clonal lineages with the phenotype of interest, using comparative genomics to identify possible sets of genes or other variants which differ between phenotype positive and negative clones (Shea et al., 2011; De Chiara et al., 2014; Cleary et al., 2016). Some studies explicitly followed a GWAS of frequency differences between genes without adjusting for population structure, and were lucky enough to find sets of only a handful of variants associated (Holt et al., 2015).

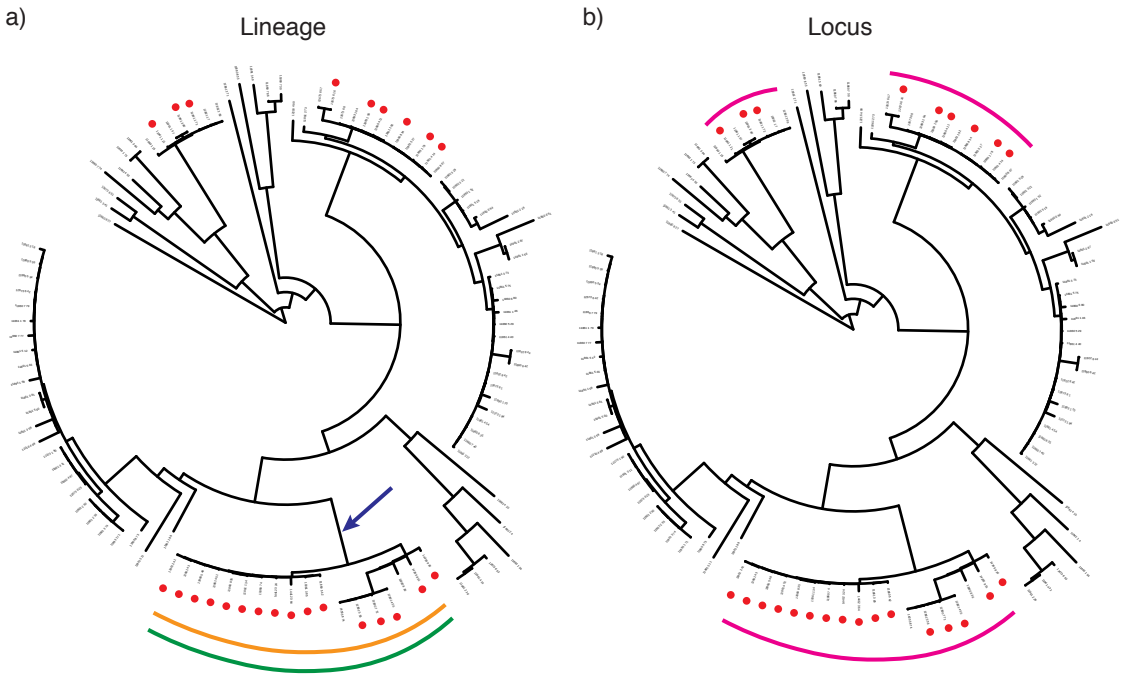


Figure 1.4: Phylogenetic illustration of lineage and locus variants. Depicted is an example phylogeny, with cases identified by red dots at the tips, and controls without dots. Variant presence is shown as coloured arcs. **a)**: The yellow variant is a causal lineage variant, and will be associated with the phenotype in a naive analysis. However the green variant, present in the same clade, is not causal but will also appear associated at the same level of significance. Indeed, any mutation that has occurred on the branch indicated by the arrow will appear associated, hindering association mapping. **b)**: The magenta variant has arisen independently in three separate clades containing cases, giving more independence from genetic background and more evidence for association with the phenotype. The association should have a higher p-value, and slightly lower OR than the green and yellow variants due to the reduced penetrance observed.

However, it is possible for variants to be associated with a phenotype independent of genetic background. These 'locus' variants can be mapped to a region of the genome, and are currently the main focus of bacterial GWAS studies. This is not because they are less important than lineage variants (both types of variant may explain any amount of the heritability), but are easier to find and map.

The phylogeny picture described above also allows us to understand two mechanisms by which locus associations may arise. Firstly, if a causal variant has happened more than once, that is independently on multiple ancestral branches, it will remain associated with the phenotype but now be uncorrelated with genetic background. These are homoplasic variants, which are likely to occur when there is selection for the phenotype across the species, for example with antibiotic use. Similarly, recombination between strains causes horizontal inheritance of DNA which cannot be represented by a phylogeny (which only represents vertical inheritance). Variants introduced by recombination are independent of genetic background, and may be associated with the phenotype across the tree. In the LD picture both these mechanisms break the correlation between variants and the rest of the genome, though not in a simple way. I note that I have only explicitly considered ancestral mutations so far. Mutations at the tips of the tree, if they have happened multiple times, are

valid homoplasies. However, if they have only happened at a handful of tips, even if they are causal, standard association will lack power to detect them regardless of population structure.

The relative prevalence and importance of recombination and homoplasy varies by the species and population of interest (as different selection pressures may have acted on different populations over time). In highly diverse and recombinogenic species such as *S. pneumoniae* and *N. meningitidis*, a phylogeny-based adjustment for population structure is likely to be the wrong approach as this will cause the tree to be inaccurate (Croucher, Page et al., 2015). However, the recombination makes genome-wide LD of the population less prevalent and somewhat more like the human genome, so a suitable regression approach may be used instead. In a clonal species such as *M. tuberculosis*, the availability of an accurate phylogeny and the huge levels of LD make direct identification of homoplasy more applicable than regression methods (Farhat et al., 2013; P. E. Chen & Shapiro, 2015).

1.4.2 More variation and fewer samples

Most human genetic variation is due to small variants which can be detected by resequencing and mapping to a reference from a single population (1000 Genomes Project Consortium et al., 2015). Though some variation is lost by considering a single reference, the contribution of pan-genomic variation is small (~1% of the overall sequence) (R. Li et al., 2010). In bacteria short variants in core genes are undoubtedly important, but the presence of an accessory genome not covered by simple SNP mapping, not to mention variation within accessory genes, is a significant source of variation (McInerney et al., 2017).

A successful bacterial GWAS therefore needs to assess not only SNP and INDEL variation, but also gene level variation. A simple way this can be achieved with modern techniques (Page et al., 2015) is by associating the presence and absence of common accessory COGs against the phenotype. This of course does not account for variation within the accessory genes unless multiple alleles are clustered separately, however adjusting this tradeoff of specificity and sensitivity in pan-genome estimation is difficult to tailor specifically to GWAS.

An alignment-free method of variant detection is therefore ideal, as the computational burden of multiple reference mappings, the bias of available references and the issue of varying levels of missing calls across the genome makes alignment generally less suitable than in human genomes. Genome assembly uses sequence words of length k, called k-mers, to align sequence internally within a sample without requiring use of a reference (Zerbino & Birney, 2008; Compeau et al., 2011). Further work has been able to co-assemble multiple samples calling variation across the pan-genome in a reference free manner (Iqbal et al., 2012), or call variation directly from k-mers in sequence reads (Gardner & Hall,

2013). One of the first bacterial GWAS studies used k-mers as the variant to perform a pan-genome-wide association study (Sheppard et al., 2013) (see section 1.4.3), and in chapter 2 I will propose this as the unit of variation in bacterial GWAS.

The pan-genome and strong population structure makes it difficult to design genotyping arrays of tag SNPs, especially as microbiologists do not have the luxury of an entire field being able to focus on a single organism (albeit a fascinating and complex one). MLST schemes can be used to define population structure with less sequencing effort, but do not have sufficient precision to perform GWAS. Without the possibility of relatively cheap genotyping arrays, bacterial sequencing has necessarily been whole-genome. The expense of this sequencing, as well as the difficulty inherent in obtaining clinically relevant bacterial samples has therefore limited sample sizes. Compounding this, the high level of variation in bacteria despite their relatively short genome size increases the multiple testing burden, necessitating large sample collections. Only recently were the first studies with thousands of phenotyped genomes published (Shea et al., 2011; Chewapreecha, Harris et al., 2014), with well powered GWAS studies following closely behind (Chewapreecha, Marttinen et al., 2014).

1.4.3 Early successes

In perhaps the first bacterial GWAS, Bille et al. (2005) were able to develop a gene-based microarray for *N. meningitidis*, and look for frequency differences between carriage and invasive isolates deliberately chosen to cover the diversity of the species. Without explicitly adjusting for population structure and only assaying a single form of variation they were able to find a phage associated with hypervirulence (Bille et al., 2008).

By equally representing isolates from different genetic backgrounds, as defined by MLST, in both cases and controls Bille et al. (2005) implicitly controlled for population structure. If the representation of different genetic backgrounds was unequal in cases and controls, in an identical way to human population structure this would confound the results. A more direct method to inform sampling before sequencing is to take pairs of phylogenetically close but phenotypically discordant isolates across the tree (Farhat et al., 2014). While it would of course increase study power to simply sequence the entire collection and adjust for population structure during analysis, the existing availability of MLST of very large isolate collections can be used to perform this targeted approach at a lower cost. Despite the limited resolution of MLST to determine genetic background, this approach has been able to find functionally confirmed associations for *L. monocytogenes* virulence (Maury et al., 2016) and *M. tuberculosis* transmissibility (Nebenzahl-Guimaraes et al., 2016).

Sheppard et al. (2013) performed a ground-breaking bacterial GWAS, which was the first to properly account for population structure and assay variation across the pan-genome

using k-mers. The authors used k-mers of length 30 to test for association of genetic variation in 29 *Campylobacter jejuni* and *Campylobacter coli* isolates with host specificity. A Monte Carlo simulation of characters on the tree was used to define a null distribution of the association test statistic when following the correlation structure of the phylogeny, thus adjusting for population structure. K-mers which were significantly associated with presence in isolates from cattle rather than isolates from birds were found to map to a seven gene cluster, which included genes coding for vitamin B_5 synthesis, a molecule present in grains but not grasses. While an important leap forward methodologically, the Monte Carlo simulation method was unfortunately not scalable to the large collections of isolates needed for greater study power, and the reliance on a recombination removed phylogeny is restrictive in many settings. The association found had a very large effect size (OR 95% confidence interval (CI) $28 - \infty$), hence the ability to find it using a small number of samples.

It is worth noting that a similar issue with population structure exists with viral GWAS, though in RNA viruses the high mutation rate and within-host diversity makes it a generally weaker effect than in bacteria. Viral sequences are (almost always) shorter than bacterial sequences, and though calling variation for association testing faces different challenges, the eventual multiple testing burden is lower. By using principal components to adjust for population structure, like in early human GWAS (A. L. Price et al., 2006), Bartha et al. (2013) performed an association between HIV-1 amino acid changes and viral load. Though they did not find any hits, this showed human genetics derived methods could control type I error rate. This study was notable for being the first genome-to-genome analysis of host and pathogen (section 5.3 covers this in more detail).

GWAS in *S. pneumoniae*

Given the high recombination rate and relatively high availability of samples, *S. pneumoniae* is a good candidate for bacterial GWAS. Chewapreecha, Marttinen et al. (2014) therefore performed the first well powered bacterial GWAS, using 3 085 genomes from pneumococcal carriage in an unvaccinated population to associate core SNPs called against a single reference with resistance to β-lactams. With this many species-wide isolates a phylogeny-independent method was required, and the authors opted to use the Cochran–Mantel–Haenszel (CMH) test to control for population structure. Using 188 discrete population clusters defined by Bayesian analysis of population structure (BAPS) as groups, this essentially performs a χ^2 test for association within each clonal group, and then meta-analyses the results from each cluster. This gave an overinflated test statistic, though substantially lower inflation than the use of 35 less finely resolved clusters. Though both have clearly been successful, the power and false positive rate of using discrete population clusters through the CMH test or as binary covariates in a regression, versus the use of

continuous covariates such as principal components remains unknown.

While they did not perform a formal meta-analysis, the results were validated in a second population of 616 carriage isolates from children in Massachusetts (Croucher, Finkelstein et al., 2013) finding 303 SNPs in the intersection of significant hits. Though mosaic alleles of the *pbp* genes are known to cause resistance (section 1.2.2), the authors aimed to identify the individual SNPs causal for resistance. However extensive and complex LD across these regions stymied this inferential aim. The lowest OR of detected hits in this study was around 2, a substantial improvement on previous smaller studies.

Aside from antibiotic resistance, only a single study has reported a GWAS for an association between pneumococcal variation and a clinical outcome. Tunjungputri et al. (2017) used an identical association model but tested COGs for association with 30-day mortality in 349 cases of bacteremia, finding that the platelet binding protein *pblB* (Bensing et al., 2001) was associated with increased mortality.

1.4.4 Phylogenetic methods

Having discussed the issues facing bacterial GWAS compared to human GWAS, and how they were approached by early studies I will now cover the state-of-the-art methods and analysis currently available for bacterial GWAS. As mentioned above these broadly fall into two categories: phylogenetic methods and regression methods.

Phylogenetic methods offer precise control of type I error rate when accounting for population structure, but rely on having a trusted phylogeny; not tainted by recombination and with good branch supports. This is possible for small collections of isolates where recombination can be removed (Croucher, Page et al., 2015; Didelot & Wilson, 2015; Mostowy et al., 2017), but not feasible across a diverse species such as *S. pneumoniae*. In some cases a posterior of trees can be used as input rather than a single representative, which can partly account for poorly supported branch splits at the expense of a greater computational burden. The total computational burden of these methods is generally high, especially if they use Monte Carlo simulations, and they are therefore unlikely to scale to millions of tests needed to assay variation across the entire pan-genome. Hence application has mostly been limited to analysis of accessory COGs, or species/clades with limited levels of SNP variation.

The history of these methods is rooted in assessing correlations between traits measured across different species (Garland & Ives, 2000). Felsenstein (1985) first proposed the use of independent contrasts, motivated by a Brownian motion model of trait evolution on the tree, using the difference in phenotype between phylogenetic sister isolates and their branch lengths to adjust for expected correlations between species (which has echoes of the approach of Farhat et al. (2014)). A tool has been written to apply this instead to binary traits using this form of approach (Brynildsrud et al., 2016). It associates COGs with

phenotypes in a naive manner, then also uses pairwise comparisons (A. F. Read & Nee, 1995) on the phylogeny to estimate the number of times the trait has evolved independently. However this model does not offer a way of combining the test of evolutionary convergence with phenotypic association.

An alternative approach is to use a generalised least squares regression, but instead of assuming independent and identically distributed error terms they use the phylogeny to estimate covariances between error terms in the model (Pagel, 1997). Desjardins et al. (2016) used this approach to test for correlated evolution between antibiotic resistance and genetic variants in *M. tuberculosis*, which in conjunction with a naive association was found to improve type II error rate without affecting type I rate in a handful of cases.

It is possible to simulate the null distribution of test statistics accounting phylogenetic correlations using Monte Carlo simulations (Martins & Garland, 1991), which was the method used by Sheppard et al. (2013) with the correlation between phenotype and genetic variants at tips of the tree as the test statistic. A recently proposed extension specific to bacterial GWAS also calculates test statistics which capture variants with correlated evolution with the phenotype through changes at nodes, and integrating across branches and therefore evolutionary history (Collins & Didelot, 2017).

1.4.5 Regression methods

In contrast to phylogenetic methods regression based methods are fast, do not require an accurate phylogeny (and therefore may also be alignment-free) and are more in-sync with the active development of human GWAS methods. They are therefore more scalable with the large sample sizes needed for high powered GWAS studies, and the high number of variants which must be tested across the pan-genome. However, compared to well-calibrated phylogenetic methods these methods may have an elevated type I error rate. Regression methods with similar control of the type I error rate have recently appeared, but are generally restricted to the discovery of locus variants, and can only test association at the tips of the tree rather than over the evolutionary history of the bacteria.

Following the approach of using principal components as fixed effects in a regression, variants associated with phenotypes such as drug resistance and virulence have successfully been found in a number of species other than those mentioned above (Laabei et al., 2014; Alam et al., 2014; Salipante et al., 2015). This method is fast, and has been successfully scaled to analysis of k-mer variants across the pan-genome (Weinert et al., 2015). The first attempt to improve upon this method in terms of population structure control leveraged the efficiency boosts in LMMs being used for trans-ethnic human GWAS studies. By applying an efficient LMM, using the relationship between strains as random effects, to their top variants from a naive association test, Earle et al. (2016) were able to find locus variants affecting antibiotic resistance while controlling type I error from population structure.

Within their model they were also able to identify potential lineage associations which were associated with both the phenotype and the population structure components, albeit with greatly reduced power.

Advances in expanding the variant space tested using regression methods have included k-mers being assembled over a sample collection into unitigs – high confidence contigs extracted from the de Bruijn graph without needing repeat resolution – thereby giving larger haplotype-like variants to test (Jaillard et al., 2017). The inclusion of rare variants by grouping LoF variants in genes has also been successful (Desjardins et al., 2016).

1.5 Conclusions

Since it became possible, GWAS has become the first step in the genetic analysis of complex traits, taking an agnostic association approach across the entire genome to generate a hypothesis for further work. By meta-analysis of data with other cohorts these associations can be asserted with more confidence. With enough samples the association can be fine-mapped, and in some cases the specific causal variant discovered. The focus of the field of human genetics on this method has led to many methodological advances, which have made this analysis more routine and more powerful.

The simple study design makes it relatively easy to collect large sample sizes, giving high power for association mapping of polygenic traits. Compared to a lab-based or *in vivo* assay, where a bottom-up approach of knocking out a gene and then testing for an effect on phenotype may well be followed, GWAS has four potential advantages:

1. The top-down approach tests all regions of the genome simultaneously, and can find associations which necessarily have any effect on phenotype without the need for any prior biological hypothesis.

2. The variation tested occurs naturally in the study population, where more subtle effects than a gene knock-out are likely important, and do not rely on a potentially inaccurate animal model.

3. The phenotype tested can be anything quantifiable. This allows investigation of important traits such as invasiveness or transmissibility which can't be determined in the lab.

4. Genetics has one way causation on phenotype, so in some cases successful association mapping can be used to determine a causal link without worrying about other epidemiological confounders. This can also be used to determine causal correlations using Mendelian randomisation.

These advantages, and the likely heritable and polygenic nature of bacterial meningitis noted so far, therefore make it an ideal technique to discover more about genetic risk factors for pneumococcal meningitis susceptibility and severity. Historically, studies have been held back by only assessing candidate genes, and current studies have not had large enough sample sizes or well-defined phenotypes in bacterial meningitis. The availability of the MeninGene cohort addresses this by adding many more samples of culture-proven pneumococcal meningitis, along with clinical outcomes.

The same benefits apply to traits in bacteria as well as humans, however issues of strong population structure, pan-genomic variation and limited sample sizes make these studies more difficult. Recent methods have successfully addressed a subset of these concerns, but an approach which deals with all of these issues and is broadly applicable is still lacking. Given the large sample sizes becoming available, a well-designed GWAS in bacteria is a promising avenue for research. In the next chapter, I will start by developing and testing a new method to perform bacterial GWAS in an efficient manner, which simultaneously addresses the difficulties listed above.

Chapter 2

Bacterial genome-wide association studies

Declaration of contributions

Jukka Corander, Stephen Bentley and Julian Parkhill supervised this work. The fsm-lite k-mer counting software was written by Niko Välimäki. The initial generation of p-values of k-mers associated with antibiotic resistance using sequence element enrichment analysis (SEER) was performed by Minna Vehkala. Mark Davies, Andrew Steer and Stephen Tong collected the *S. pyogenes* isolates. I performed all other analyses, design, coding and maintenance of SEER, and generated all the figures.

Publication

The following has been published as:
Lees J. A., Vehkala M., Välimäki N., Harris S. R., Chewapreecha C., Croucher N. J., Pekka M., Davies M. R., Steer A. C., Tong S. Y. C., Honkela A., Parkhill J., Bentley S. D., Corander J. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*. 2016; 7, 12797. http://dx.doi.org/10.1101/038463

and prepared for publication as:
Lees J. A., Parkhill J., Harris S. R., Bentley SD. An evaluation of phylogenetic reconstruction using bacterial whole genomes.

2.1 Introduction

The goal of GWAS is to determine which genetic variants, anywhere in the genome, are associated with a trait of interest. For a binary phenotype, DNA from unrelated cases and controls are collected (ideally in the ratio 1:1 to maximise power). The simplicity of sample collection and the power of the resulting test has made GWAS a compelling study design in human genetics. In this I present work I undertook to apply this study to populations of bacterial genomes.

I wished to overcome the following issues, which were yet to be simultaneously solved by existing methods:

- Account for strong clonal population structure.
- A test which works for both complex and Mendelian-like traits.
- Test variation in the entire pan-genome.
- A computationally tractable method, implemented in a form others can use.

The first issue requires the development of an appropriate association test. The simplest test between a variant and binary phenotype is a χ^2 test based on the difference between observed and expected counts in a 2x2 contingency table comparing the proportion of case isolates an element is present in to the proportion of control isolates an element is present in. This does not account for population structure described in section 2.3, leading to many non-causal lineage associated variants reaching significance. Chewapreecha, Marttinen et al. (2014) showed that performing this test separately in each discrete defined population cluster, then combining the results (i.e. the CMH test) can mitigate this problem.

However, the definition of these clusters requires a core genome alignment and running external software (BAPS). The former may not always be available, and the latter can be computationally prohibitive to run. Additionally, when there are many population clusters compared to the total number of samples, power may be reduced. I first investigated the accuracy and computational requirements of a number of methods which represent bacterial population structure, with the goal of finding one which is fast to run and does not require a core genome alignment. Given such a definition of population structure, this could then included as fixed effects in a logistic regression. This is similar to a χ^2 test, but allows covariates to be included in the model fit, in this case to account for clonal population structure. I additionally gave consideration to the performance of this test when a single highly penetrant variant causes the phenotype, as for many antibiotic resistance determinants. This is closer to a Mendelian-like trait, as opposed to a complex trait which is affected by many lower penetrance variants.

The issue of assaying variation in the bacterial pan-genome relates to what variant is used as the predictor in these tests. Taking SNPs in the core genome, as in early human

GWAS, will miss phenotypes caused by diverse forms of variation. This can include indels, recombinations, variable promoter architecture, and differences in gene content as well as capturing these variations in regions not present in all genomes. I compared calling variation in terms of SNPs and COGs with k-mers – short words of DNA of length k, that have the potential to capture all these forms of variation. In the present chapter only common ($\geq 1\%$ MAF) variants are considered. The testing of rare variants ($< 1\%$) is underpowered in the sample sizes used here. The use of burden testing to approach this issue is discussed and performed in section 4.4.

Finally, after coming up with a test framework to overcome these issues, I designed the software package SEER to implement it. I used object oriented C++ code for speed and maintainability, as well as access to efficient linear algebra and optimisation packages (Sanderson, 2010; Sanderson & Curtin, 2016; D. E. King, 2009). I released SEER on github (https://github.com/johnlees/seer), where user comments have contributed to continued improvement and maintenance of the software.

The following sections describe how I dealt with each of these issues in turn. Section 2.6 then describes how the finished method was then applied to three datasets: on simulated data to compare its performance to existing methods, and two real datasets. The first real dataset tested whether known associations with antibiotic resistance can be recapitulated, and the second attempted to find new associations with virulence.

2.2 K-mers as a generalised variant

K-mers have the potential to allow simultaneous discovery of both short genetic variants and entire genes associated with a phenotype. Longer k-mers provide higher specificity but less sensitivity than shorter k-mers (Ondov et al., 2016). Rather than arbitrarily selecting a length prior to analysis or having to count k-mers at multiple lengths and combine the results, I wished to count all k-mers at lengths over nine bases long (as below this mapping specificity is poor).

Over all N samples, all k-mers over 9 bases long that occur in more than one sample are counted. All non-informative k-mers are omitted from the output; a k-mer X is not informative if any one base extension to the left (aX) or right (Xa) has exactly the same frequency support vector as X. The frequency support vector has N entries, each being the number of occurrences of k-mer X in each sample. Further filtering conditions are explained in section 2.2.1 below.

I used three different methods to count informative k-mers from all samples in a study. For very large studies, or for counting directly from reads rather than assemblies, I used an implementation of distributed string mining (DSM) (Välimäki & Puglisi, 2012; Seth et al., 2014) which limits maximum memory usage per core, but requires a large cluster to run.

DSM parallelises to as much as one sample per core, and either 16 or 64 master server processes. DSM includes an optional entropy-filtering setting that filters the output k-mers based on both number of samples present and frequency distribution. On 3 069 simulated genomes this took 2 hrs 38 min on 16 cores, and used 1Gb RAM per core. The distributed approach is applicable up to terabytes of short-read data (Seth et al., 2014), but requires a cluster environment to run.

For data sets up to around 5 000 sample assemblies (gigabyte-scale data) we implemented a single core version, fsm-lite, which is easier to install and run. We based fsm-lite on a succinct data structure library (Gog et al., 2014) to produce the same output as DSM. On 675 *S. pyogenes* genomes this took 3hrs 44min and used 22.3Gb RAM.

For comparison with older datasets, or where resources do not allow the storage of the entire k-mer index in memory, I used DSK (Rizk et al., 2013) to count a single k-mer length in each sample individually, then combined the results. I wrote the program combineKmers using an associative array in C++ to combine the results from DSK in memory. I concatenated results from k-mer lengths of 21, 31 and 41, as in Sheppard et al. (2013). This could in future be scaled to larger genome numbers by instead using external sorting to avoid storing the entire array in memory.

To get an idea of how much of the total genomic variance of the population each type of variant (gene, SNP or word) captured, I compared the site frequency spectrum (SFS) of informative k-mers with COGs and SNPs. Figure 2.1 shows this comparison for the 1 144 *S. pneumoniae* genomes described in chapter 4. The k-mer SFS is a similar distribution to the SNP SFS, though there are in total two orders of magnitude more words. There are also more fixed k-mers ($> 99\%$ allele frequency (AF)) – these are due to the core COGs seen in the final row. Removing rare variants which are not tested for association, the k-mer SFS remains representative of the two other variation types, and appears to be capturing both.

2.2.1 Filtering k-mers

Before testing for association, I filtered k-mers based on their frequency and unadjusted p-value. This reduced false positives from testing underpowered k-mers and reduce computational time. If not biologically plausible, k-mers with negative effect sizes are filtered at this point.

K-mers are filtered if either they appear in $< 1\%$ or $> 99\%$ of samples, or are over 100 bases long when counted by DSM. I also first test if the p-value of association in a simple χ^2 test (with 1 d.f.) is less than 10^{-5}, and remove it otherwise. In the case of a continuous phenotype a two-sample t-test is used instead. The effect of these filters is discussed in section 2.4.1.

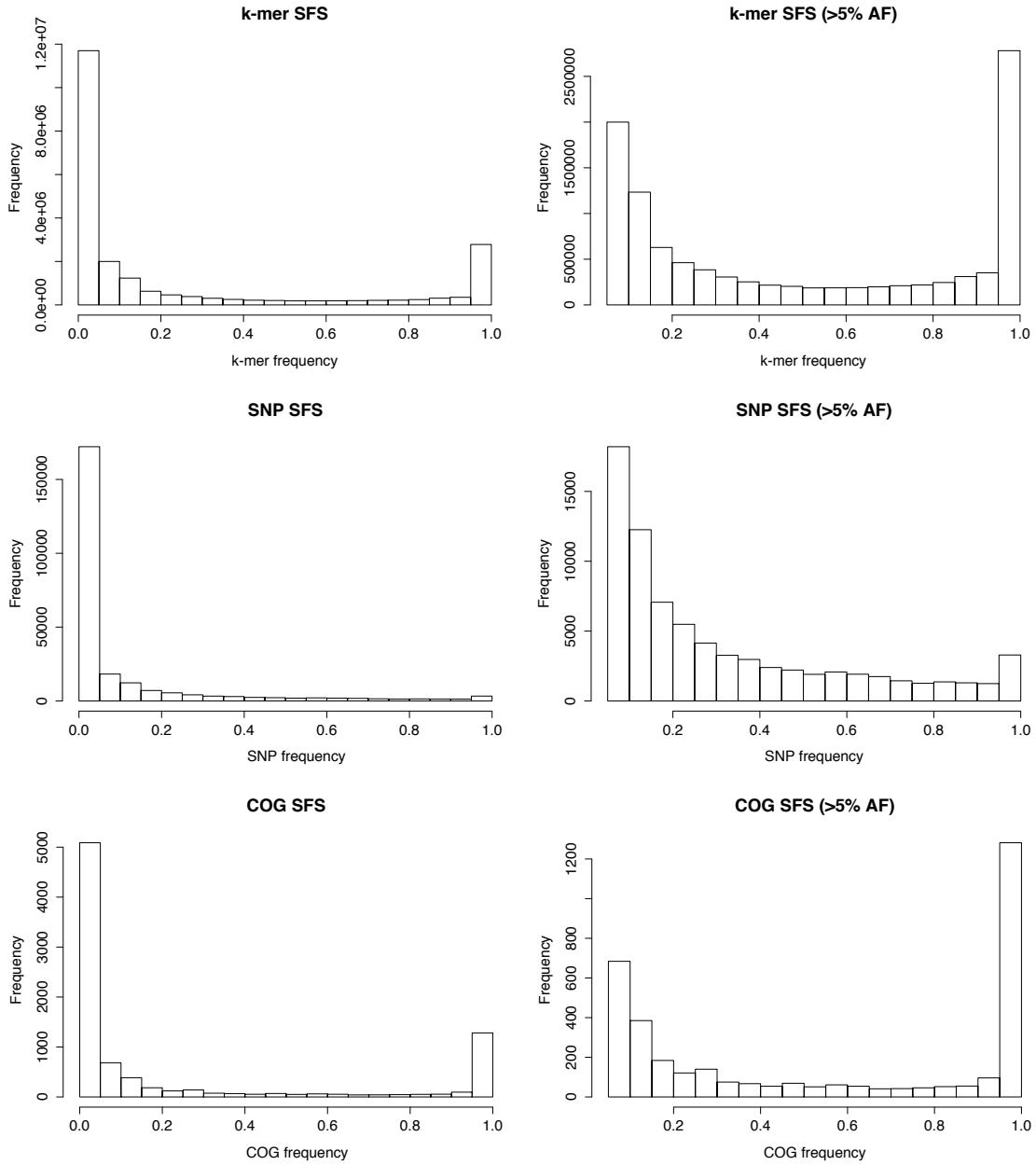


Figure 2.1: The SFS of 1 144 *S. pneumoniae* genomes. The x-axis is AF, the y-axis is the number of variants with allele-frequencies in that bin. Each row uses different sites: the first row shows k-mer presence, the second row SNPs as the sites (with respect to the ATCC 700669 reference), the third COGs. The first column shows all sites, the second column only common sites with $> 5\%$ AF.

2.3 Accounting for population structure

Due to the clonal reproduction of bacteria, rather than eukaryotic sexual reproduction resulting in recombination every generation, the genomes from a sampled population will usually be highly related. This leads to extensive LD across the chromosome, and a simple GWAS will therefore find many variants reaching significance due to their correlation with causal variants. The relatedness between all the bacteria in the study must therefore be quantified, and then appropriately used in the association model to control for this effect.

In this section I detail ways in which the population structure may be quantified, then in section 2.4 I explain how this is incorporated into an appropriate association test.

2.3.1 Phylogenetic simulation of genomes

To test the accuracy of population structure estimation, I simulated realistic data with a known phylogenetic relationship. I then used a suite of methods that infer this phylogeny from the resulting genome sequence assemblies or alignments, and evaluated them in terms of accuracy, efficiency and ease of implementation. The use of simulated data under a realistic model was desirable, as using a tree inferred from real read data as the true tree would be circular, and would necessarily result in the model that was used to infer the tree in the first place as being the most accurate.

I used artificial life framework (ALF) (Dalquen et al., 2012) to simulate evolution along a given phylogenetic tree, using the 2 232 coding sequences in the ATCC 700669 genome as the most recent common ancestor (MRCA). I used a phylogeny (fig. 2.2), originally produced by Kremer et al. (2017) from a core genome alignment of 96 *L. monocytogenes* genomes from patients with bacterial meningitis, possessing a number of qualities I wished to be able to reproduce: two distinct lineages, several clonal groups within each lineage, long branches and a polyphyletic cluster. I define N as the number of strains in the study and M as the number of aligned sites.

To estimate rates in the generalised time reversible (GTR) matrix and the size distribution of insertions and deletions, I aligned *S. pneumoniae* strains R6 (AE007317), 19F (CP000921) and *S. mitis* B6 (FN568063.) using Progressive Cactus (Paten et al., 2011). I used previously determined parameters for the rate of codon evolution (Kosiol et al., 2007), relative rate of SNPs to indels in coding regions (J. Q. Chen et al., 2009), rates of gene loss and horizontal gene transfer (Chewapreecha, Harris et al., 2014) when running the simulation. In parallel, I used DAWG (Cartwright, 2005) to simulate evolution of intergenic regions using the same GTR matrix parameters and previously estimated intergenic SNP to indel rate (J. Q. Chen et al., 2009). I combined the resulting sequences of coding and non-coding regions at tips of the phylogeny while accounting for gene loss and transfer, and finally generated error prone Illumina reads from these sequences using pIRS (Hu et al., 2012).

To generate input to phylogenetic inference algorithms, I created assemblies and alignments from the simulated reads. I assembled the simulated reads into contigs with velvet (Zerbino & Birney, 2008), then improved and annotated the resulting scaffolds (Page et al., 2016). I generated alignments by mapping reads to the TIGR4 reference using bwa-mem with default settings (H. Li, 2013), and called variants from these alignments using samtools mpileup and bcftools call (H. Li, 2011). I used Roary (Page et al., 2015) with a 95% BLAST ID cutoff to construct a pan-genome from the annotated assemblies,

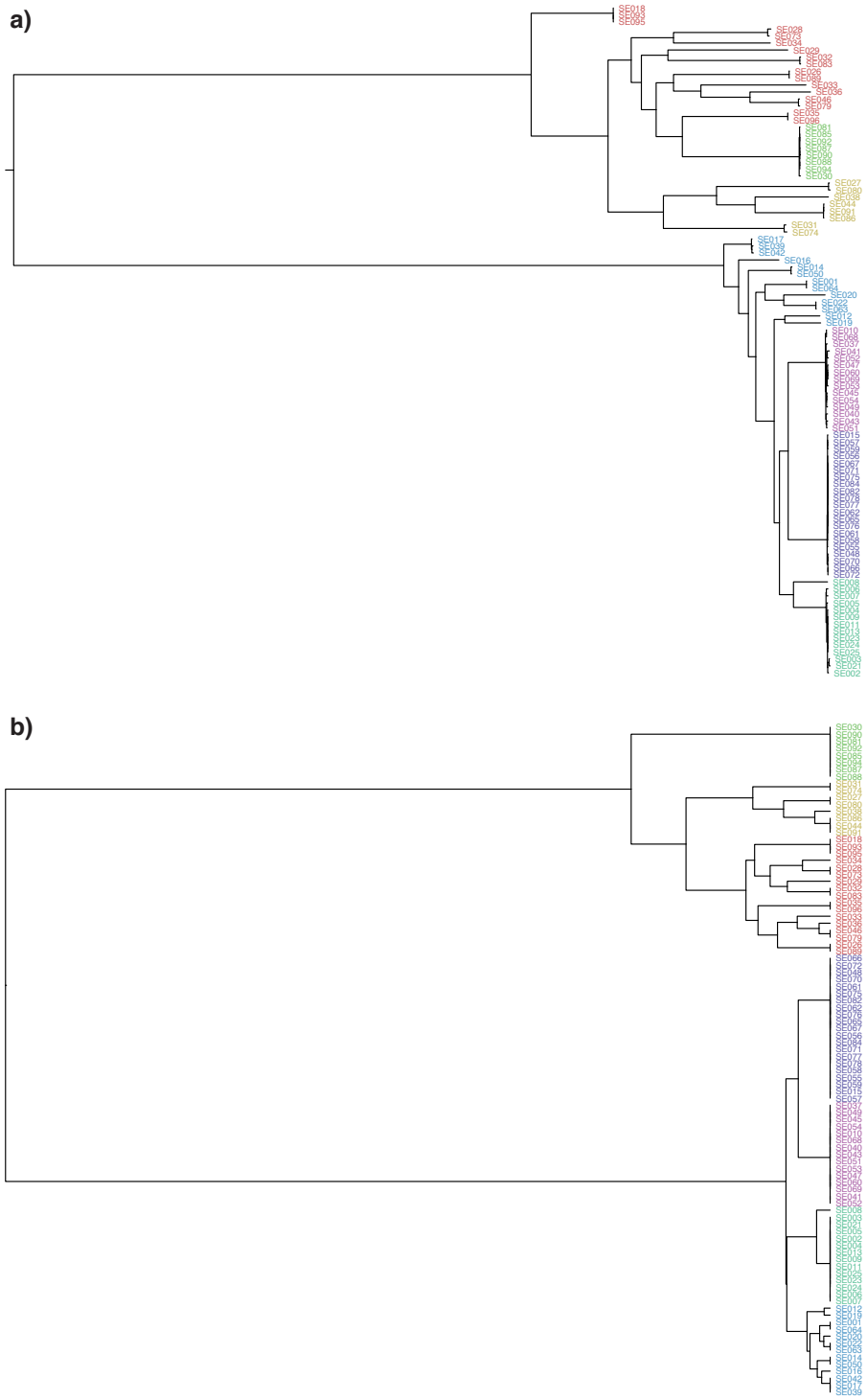


Figure 2.2: a) The phylogeny inferred by Kremer et al. (2017) used as the true tree in simulations. Tips are coloured by BAPS cluster inferred from the core genome alignment. **b)** The UPGMA tree using k-mer distances as used by SEER; tip colours are the original BAPS clusters shown in a).

from which a core gene alignment was extracted. I then created alignments by two further methods. For a MLST alignment I selected seven genes at random from the core alignment (present in all strains) which had not been involved in horizontal transfer events. For a Progressive Cactus alignment, I ran the software on the assemblies using default settings, and extracted regions aligned between all genomes from the hierarchical alignment file

and concatenated them.

Using the nucleotide alignments described above as input, I ran the following phylogenetic inference methods:

- RAxML 7.8.6 (Stamatakis, 2014) with a GTR+gamma model (-m GTRGAMMA).
- RAxML 7.8.6 with a binary+gamma sites model (-m BINGAMMA).
- FastTree 2.1.9 (M. N. Price et al., 2009) using the GTR model (denoted slow) and using the -pseudo and -fastest options (denoted fast).
- Parsnp 1.2 (Treangen et al., 2014) on all assemblies using the -c and -x options (removing recombination with PhiPack).

I also created pairwise distance matrices using:

- Mash 1.0 (Ondov et al., 2016) (default settings) between assemblies.
- Andi 0.9.2 (Haubold et al., 2015) (default settings) between assemblies.
- Hamming distance between informative k-mers using a subsample of 1% of counted k-mers from assemblies.
- Hamming distance between rows of the gene presence/absence matrix produced by Roary (using 95% blast ID cutoff).
- Jukes-Cantor (JC) and logdet distances between sequences in the alignment, as implemented in SeaView 4.0 (Gouy et al., 2010).
- Distances between core gene alleles (add a distance of zero for each core gene with identical sequence, add a distance of one if non-identical), as used in the BIGSdb genome comparator module (Jolley & Maiden, 2010).
- Normalised compression distance (NCD) (Vitányi et al., 2009), using PPMZ as the compression tool (Alfonseca et al., 2005).

For all the above distance matrix methods I then constructed a neighbour joining (NJ) tree, a BIONJ tree (Gascuel, 1997) using the R package ape, and an UPGMA tree using the R package phangorn. In the comparison I retained the tree building method from these three with the lowest Kendall-Colijn (KC) distance from the true tree.

To measure the differences in topology between the produced trees (either between the true tree and an inferred tree, or between all different inferred trees) I used two measures. As a sensitive measure of changes in topology I used the metric proposed by Kendall and Colijn (2016) with $\lambda = 0$ (ignoring branch length differences). I compared the true tree

against midpoint rooted random trees giving 286 (95% CI 276-293) as an upper limit on poor topology inference.

For trees distant from the true tree by the KC metric it was useful to test whether the tree was accurate overall and only a few clade structures were poorly resolved, or whether the tree failed to capture important clusters at all. I therefore used a measure of the clustering of the BAPS clusters from the true alignment on each inferred tree. For each pair of isolates in a BAPS cluster, a one is added to the score if any children of their most recent common ancestor is from a different cluster. I applied this to both the primary BAPS cluster, which separates the two main lineages, and the secondary BAPS clusters which define finer structure in the data. For the primary BAPS cluster a score of 0 was achieved by the true tree, which maintained these clusters, and 2437 (95% CI 2401-2457) for random trees. For the secondary BAPS clusters (excluding the 'bin' cluster) a score of 63 was achieved by the true tree, as one cluster is polyphyletic (removing this cluster gives a score of 0 to the true tree), and 535 to random trees (95% CI 531-539).

Method	KC	BAPS 1	BAPS 2	CPU time	Memory	Overheads	Parallelisability	Accessory genome?
	(0-286)	(0-2437)	(0-535)					
RAxML + close reference alignment	4.63	0	63	806.5 minutes	2.7 Gb	Mapped alignment	Pthreads	No
RAxML + alignment	11.2	0	63	587 minutes	3 Gb	Mapped alignment	Pthreads	No
Parsnp	14.0	0	63	42.5 minutes	2.6 Gb	Assemblies	Threads	No
FastTree + alignment	16.0	0	63	189 minutes	10.6 Gb	Mapped alignment	Threads (up to 4)	No
RAxML + core gene alignment	18.6	0	63	29.2 minutes	0.15 Gb	Core gene alignment	Pthreads	No
NJ + SNP alignment	20.5	0	63	Negligible	Negligible	Mapped alignment	No	No
BIONJ + mash distances	51.7	0	63	0.75 minutes	10 Mb	Assembly	Embarrassingly	Yes
RAxML + MLST alignment	62.6	0	63	1.4 minutes	19 Mb	Assembly	Pthreads	No
BIONJ + andi distances	66.0	0	60	7.48 minutes	290 Mb	Assembly	Embarrassingly	Yes
RAxML + Cactus alignment	67.2	0	63	9 600 minutes	37.4 Gb	Assembly	Threads	No
RAxML + gene presence/absence	77.3	0	57	4.28 minutes	20 Mb	Core gene alignment	Threads	Yes
BIONJ + k-mer distances	89.6	0	63	37.3 minutes	180 Mb	Assembly	Threads	Yes
BIONJ + BIGSdb	149.8	0	22	0.48 minutes	Negligible	Assembly	Embarrassingly	No
UPGMA + NCD	210	0	627	1 040 minutes	Negligible	Assembly	Embarrassingly	Yes

Table 2.1: Accuracy and resource usage of phylogenetic reconstruction methods, ordered by KC metric score. The method lists the best combinations of all alignment with phylogenetic method, and distance matrices with phylogenetic methods. Three scores of accuracy of the phylogeny are shown; values in the header are the range the values can take. Parallelisability shown is that built into the software, 'embarrassingly' is when every value in a distance matrix is independent so can be parallelised up to N^2 times.

Table 2.1 and fig. 2.3 show the results of my simulations. I used these simulations to guide the population structure correction to use in SEER bearing in mind the criteria laid out above, and also for efficiency/accuracy tradeoffs when constructing phylogenies

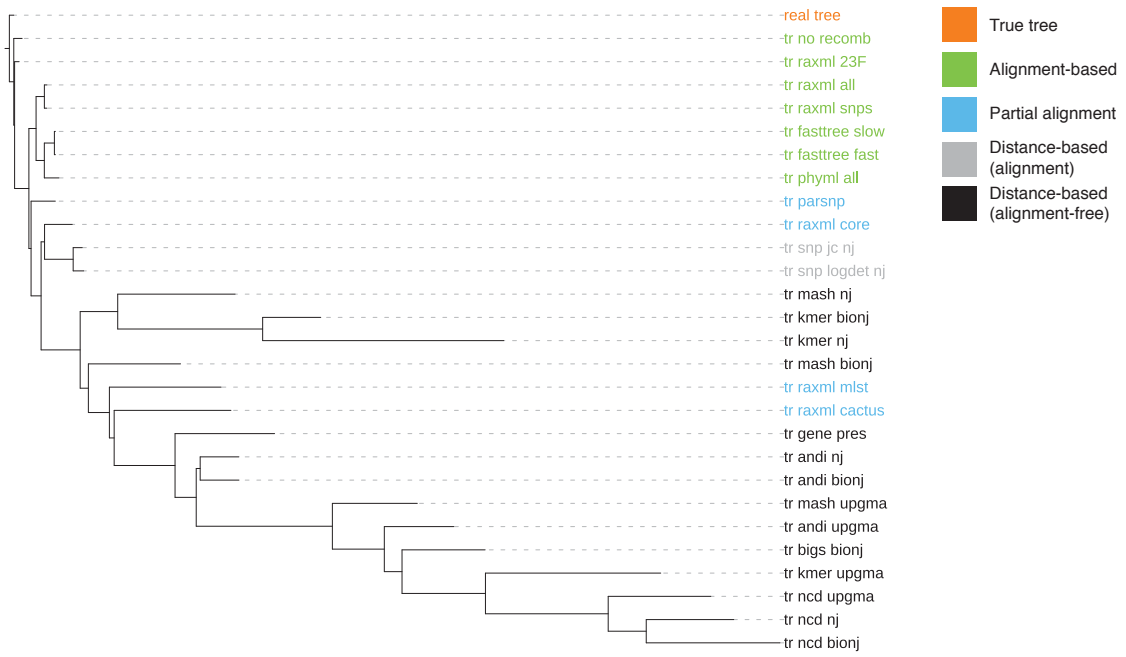


Figure 2.3: Using the KC metric between all the inferred phylogenies in table 2.1 to create a pairwise distance matrix, then an NJ tree from this matrix. This shows how the topologies from all methods are related to each other (a tree-of-trees, or supertree). The true tree is in orange and was used to root the tree, and four classes of method are labelled.

throughout the rest of this thesis.

Firstly I note that all methods except for the NCD were able to recapitulate the population clusters as defined by BAPS. Therefore for analyses which require identifying clusters on the phylogeny, but not finer scale topology, quicker but less accurate methods are sufficient. For construction of a maximum likelihood tree RAxML is currently the most efficient software available. This was the most accurate method tested, and also the most resource heavy. RAxML's model fits the way the data was generated, and is expected to be a good model of evolution. There was no significant difference in fit between the inferred tree and the true tree (likelihood ratio test (LRT) = 2.34; p = 0.13). When applied to an alignment with a reference genome more distant from the root, this method was still the most accurate. Using a core genome alignment slightly reduces the accuracy, as the number of sites M used in the inference was reduced compared to the pseudo-alignment from mapping. Using an MLST alignment of seven genes reduces the accuracy greatly, as only a small proportion of the genomic variants are now used the the inference.

I found parsnp and FastTree on a whole genome alignment to be the methods which, while slightly less accurate than RAxML, were able to produce a good quality phylogeny rapidly. This is useful for alignments with large N and M. Distance matrix and NJ methods generally performed more poorly, but were still able to resolve large scale population structure differences.

I now discuss in detail a method which fulfilled the criteria for SEER's population

structure correction: it accurately represented the BAPS clusters without needing a core-genome alignment, used only the information already needed to perform and association test on k-mers, could be efficiently implemented in C++ with the rest of SEER, and could be used to provide covariates for a logistic or linear regression rather than using discrete clusters or a phylogeny.

2.3.2 K-mer distance method producing covariates to control for population structure

Compared with modelling SNP variation, the use of k-mers as variable sequence elements has been previously shown to accurately estimate bacterial population structure (Tasoulis et al., 2014). As k-mers are going to be used as the input to the association test, it would be convenient if they could also be used to control for population strucutre. I defined the k-mer distance in table 2.1 as follows. First I take a random sample of between 0.1% and 1% of k-mers appearing in between 5-95% of isolates. I then construct a pairwise distance matrix $\boldsymbol{D}$, with each element being equal to a sum over all m sampled k-mers:

$$d_{ij} = \sum_{m} ||k_{im} - k_{jm}|| \tag{2.1}$$

where k_{im} is 1 if the mth sampled k-mer is present in sample i, and 0 otherwise. Each element d_{ij} is therefore an estimate of the number of non-shared k-mers between a pair of samples i and j, and furthermore is proportional to the Jaccard distance between the samples (Levandowsky & Winter, 1971). When I clustered samples using these distances, I got the same results as clustering core alignment SNPs using hierBAPS (L. Cheng et al., 2013) as shown in fig. 2.2b). These clusters have been used in previous bacterial GWAS studies to correct for population structure (Chewapreecha, Marttinen et al., 2014). However, this distance matrix has the clear advantage that no core gene alignment or SNP calling is needed, so it can be directly applied to the the k-mer counting result.

In an analogous way to the standard method used in human genetics of using principal components of the SNP matrix to correct for divergent ancestry (A. L. Price et al., 2006; Chengsong & Jianming, 2009), I then wrote C++ code to perform metric multidimensional scaling (MDS) on $\boldsymbol{D}$, projecting these distances into a reduced number of dimensions. The normalised eigenvectors of each dimension of this projection can then be used as covariates in the regression model, where the number of dimensions used is a user-adjustable parameter, and can be evaluated by the goodness-of-fit and the magnitude of the eigenvalues. For the tree shown in fig. 2.2, one dimension was sufficient as a population control (fig. 2.4a), whereas for the larger collection of 3 069 isolates 10-15 dimensions were needed to give tight control (fig. 2.4b). The small collection has much of the variance explained by the first dimension/eigenvector, as there is a large separation

between two main lineages. In the other collections there is a strain structure with multiple lineages, so more dimensions must be included to capture this. Over all the studies I tested, generally three dimensions appeared a good trade-off between sensitivity and specificity, but I automatically provide a scree plot as output so users can choose an appropriate number of dimensions to retain.

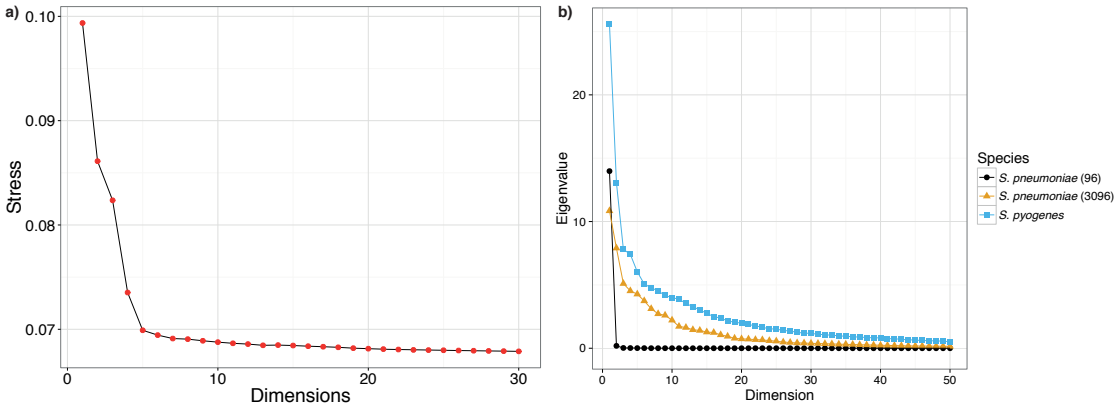


Figure 2.4: a) Stress against first thirty dimensions, calculated for the *S. pneumoniae* simulations in section 2.6.1 (orange in panel b). Stress is defined as $S^2 = 1 - R^2$, where the R^2 statistic is calculated from a regression between the upper triangle of entries in the distance matrix (i.e. pairwise between all samples) and the Euclidean distance between samples in the reduced dimension space. **b)** Eigenvalues for the first fifty dimensions of the 96 simulated *S. pneumoniae* isolates in black (section 2.3.1), 3 069 *S. pneumoniae* isolates in orange (section 2.6.1), and 675 *S. pyogenes* isolates (section 2.6.3) in blue.

I noted above that the distance used to approximate bacterial population structure is an estimate of the k-mer Jaccard distance. After the first version of SEER, the software mash was developed. This instead uses the MinHash algorithm on k-mers to estimate the Jaccard distance between sequences in a highly efficient manner (Ondov et al., 2016). As shown in table 2.1 and fig. 2.3 this distance matrix is considerably more computationally efficient than the subsampling proposed above, works from the same input data, and produced a more accurate version of the tree topology in tests. Since version `acc4bc1` I have recommended the use of mash over the above calculation I implemented in SEER, and provide scripts to run mash and MDS in a manner compatible with the rest of the package.

2.4 Association testing

Using k-mers as a generalised variant and the above population structure definition I used general linear models with fixed effects to test for association between genetic variation and phenotypes. For each k-mer, I wrote code to fit a logistic curve to binary phenotype data, and a linear model to continuous data. I took care to use time efficient optimisation routines to allow testing of all k-mers. Bacteria can be subject to extremely strong selection pressures, producing common variants with very large effect sizes, such as antibiotics

inducing resistance-conferring variants. This can make the data perfectly separable, and consequently the maximum likelihood estimate ceases to exist for the logistic model. Firth regression has been used to obtain results in these cases (Heinze & Ploner, 2003).

In detail, the SEER association testing code does the following. For samples with binary outcome vector $\boldsymbol{y}$, it fits a logistic model to each k-mer:

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{I}-\boldsymbol{y}}\right) = \boldsymbol{X}\boldsymbol{\beta} \tag{2.2}$$

where absence and presence for each k-mer are coded as 0 and 1 respectively in column 2 of the design matrix $\boldsymbol{X}$ (column 1 is a vector of ones, giving an intercept term). Subsequent columns j of $\boldsymbol{X}$ contain the eigenvectors of the MDS projection, any input categorical covariates (automatically dummy encoded), and quantitative covariates (automatically normalised). I used the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm to maximise the log likelihood $\mathscr{L}$ in terms of the gradient vector $\boldsymbol{\beta}$ (using an analytic expression for $\mathrm{d}(\log\mathscr{L})/\mathrm{d}\boldsymbol{\beta}$):

$$\log(\mathscr{L}) \propto \sum_i [y_i \cdot \log(\mathrm{sig}(\boldsymbol{X}\boldsymbol{\beta})_i) + (1-y_i)\cdot\log(\mathrm{sig}(1-\boldsymbol{X}\boldsymbol{\beta})_i)] \tag{2.3}$$

where sig is the sigmoid function. If this fails to converge, n Newton-Raphson iterations are applied to $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + [-\mathscr{L}''(\boldsymbol{\beta}_n)]^{-1}\cdot\mathscr{L}'(\boldsymbol{\beta}_n) \tag{2.4}$$

from a starting point using the mean phenotype as the intercept, and the root-mean squared beta from a test of k-mers passing filtering:

$$\beta_{0,0} = \frac{\sum y_i}{n}$$
$$\beta_{0,j>0} = 0.1$$

This is slower than using BFGS, but has a higher success rate.

If any entries for the observed counts in the contingency table were one or zero, or if two counts were five or less then Firth logistic regression is used instead. This regression is also used if after 1 000 Newton-Raphson iterations convergence is not reached, due to the observed points being separable, or the standard error of the slope is greater than 3 (which empirically indicated almost separable data). Firth regression adds an adjustment to $\log(\mathscr{L})$:

$$\log[\mathscr{L}(\boldsymbol{\beta})]^* = \log[\mathscr{L}(\boldsymbol{\beta})] + \frac{1}{2}\cdot\left\{\frac{d^2\mathscr{L}}{d\boldsymbol{\beta}^2}(\boldsymbol{\beta})\right\} \tag{2.5}$$

using which I applied Newton-Raphson iterations as above.

In the case of a continuous phenotype a linear model is fitted:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \tag{2.6}$$

to find $\boldsymbol{\beta}$, I used the BFGS algorithm to minimise the squared distance $U(\boldsymbol{\beta})$:

$$U(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \tag{2.7}$$

If this fails to converge then the solution is instead obtained by orthogonal decomposition of the design matrix:

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} \tag{2.8}$$

then back-solving for beta in:

$$\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{Q}^T\boldsymbol{y} \tag{2.9}$$

For both the logistic and linear model the standard error on the slope β_1 is calculated by inverting the Fisher information matrix $\mathrm{d}^2\mathscr{L}/d\boldsymbol{\beta}^2$ to obtain the variance-covariance matrix. Inversions are performed using the Cholesky decomposition, or if this fails due to the matrix being almost singular I used the Moore-Penrose pseudoinverse. In the initial version of SEER, I used the Wald statistic to test the probability null hypothesis of no association ($\beta_1 = 0$)

$$W = \frac{\beta_1}{\mathrm{SE}(\beta_1)} \tag{2.10}$$

which is the test statistic of a χ^2 distribution with 1 d.f. This is equivalent to the positive tail of a standard normal distribution, one minus the integral of which gives the p-value.

The Wald test loses power when large effect sizes are tested (Agresti, 2015); I observed this when testing k-mers of a mosaic *penA* allele which are known to be causal for cephalosporin resistance in *Neisseria gonorrhoeae* (Unemo & Shafer, 2014). A χ^2 test gave a p-value of 3.5×10^{-181} whereas a logistic regression using the Wald test gave a p-value of 1.9×10^{-45}, less significant than some non-causal k-mers. A better test is the LRT: in this case, the LRT of the logistic model gave a p-value of 8.4×10^{-190}, making these k-mers the top hit.

Here, the LRT test statistic D is defined as

$$\begin{aligned} D &= -2 \cdot \log\left(\frac{\mathscr{L}(\text{alternative model})}{\mathscr{L}(\text{null model})}\right) \\ &= 2 \cdot [\log\{\mathscr{L}(\beta_1 = \beta_{\text{fit}})\} - \log\{\mathscr{L}(\beta_1 = 0)\}] \end{aligned}$$

using eq. (2.3) as the likelihood. The distribution of D is χ^2 with $\mathrm{df}_{\text{alt}} - \mathrm{df}_{\text{null}}$. In this case, two times the difference between the log-likelihood at the fitted value and the log-likelihood of a fit where the k-mer presence/absence column is removed from the design

matrix is tested using a χ^2 distribution with one degree of freedom. Since version `038c4cd` of SEER the p-value for logistic regression is instead calculated using the LRT by default, though the Wald test p-value is still reported for backwards compatibility.

2.4.1 Significance cut-off

For the basal cut-off for significance I used $p < 0.05$, with which I used the conservative Bonferroni correction for multiple testing to give the threshold 1×10^{-8} based on every position in the *S. pneumoniae* genome having three possible mutations (Ford et al., 2013), and all this variation being uncorrelated. This is a strict cut-off level that prevents a large number of false-positives due to the extensive amount of k-mers being tested, but does not over-penalise by correcting directly on the basis of the number of k-mers counted. To calculate an empirical significance testing cut-off for the p-value under multiple correlated tests, I generated the distribution of p-values from 100 random permutations of phenotype. For the 3 069 Maela genomes setting the FWER at 0.05 gave a cut-off of 1.4×10^{-8}, supporting the above reasoning.

In general, the number of k-mers and the correlations between their frequency vectors will vary depending on the species and specific samples in the study, so the p-value cut-off should be chosen in this manner (either by considering possible variation given the genome length, or by permutation testing) for individual studies. I have also included association effect size and p-value of the MDS components in the output of SEER, to compare lineage and variant effects on the phenotype variation.

The effect the initial χ^2 filtering step can be seen by plotting the unadjusted and adjusted p-values of the k-mers from the simulated data set described in section 2.6.1 against each other (fig. 2.5). 430 k-mers of 12.7M passing frequency filtering have an unadjusted p-value which fail to meet the χ^2 significance threshold, but would be significant using the adjusted test (and have a positive direction of effect). These k-mers were all short words (10-21 bases; median 12) that appear multiple times per sample, and therefore are of low specificity. When I tested the top p-value k-mer in this set it showed a strong association of the presence/absence vector with three population structure covariates used ($p = 1.4 \times 10^{-24}$; $p = 1.2 \times 10^{-46}$; $p = 1.5 \times 10^{-9}$ respectively). Using lasso regression, the second population structure covariate has a higher effect in the model than the k-mer frequency vector. Together, this suggested that these filtered k-mers are associated to a lineage related to the phenotype, but are unlikely to be causal for the phenotype themselves. To confirm this, I mapped these k-mers back to the reference sequence. None of these k-mers map to the gene causal to the phenotype.

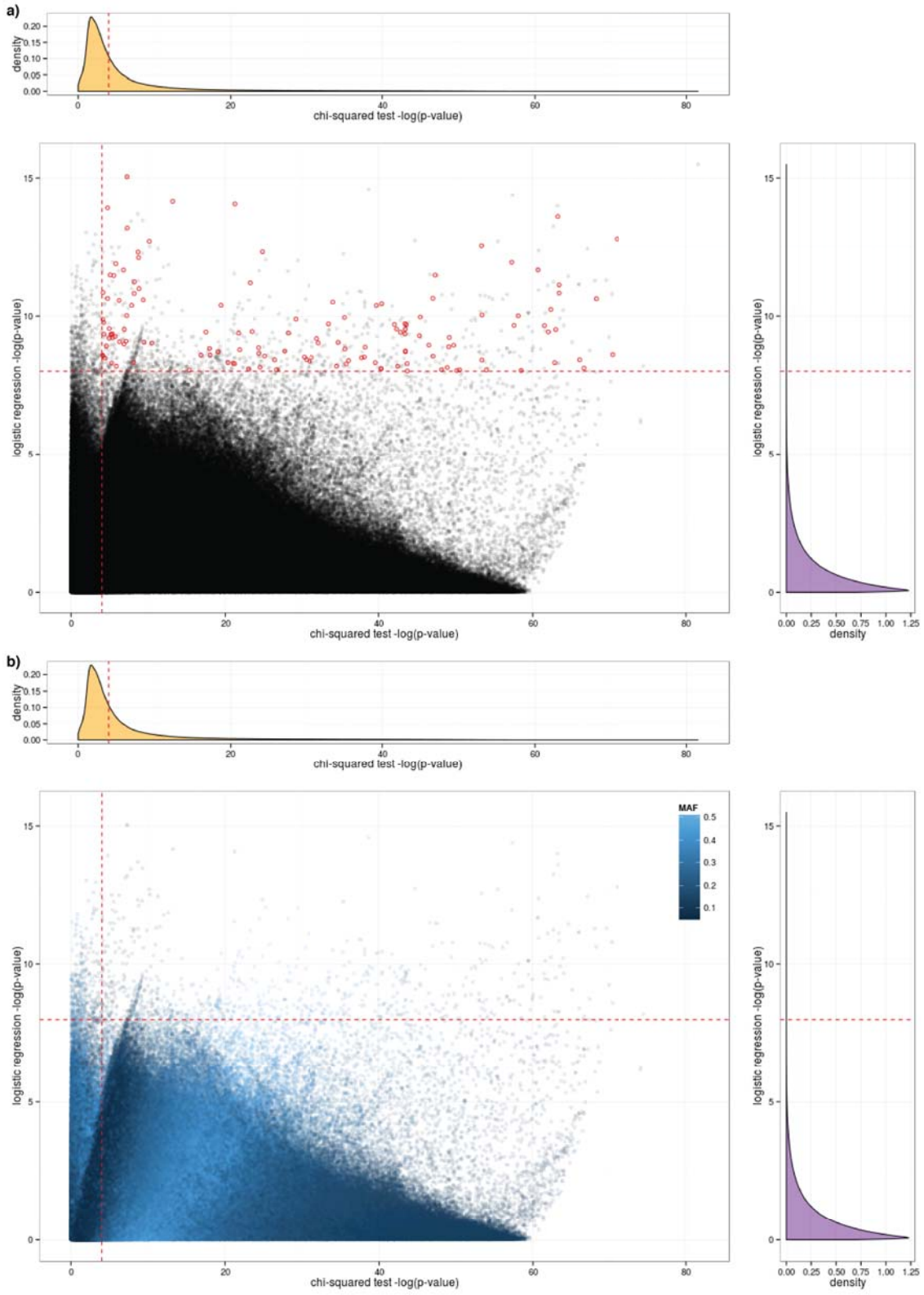


Figure 2.5: The $-\log_{10}$ p-values from a χ^2 test against the p-value from a logistic regression using the first three MDS components as covariates. The points are from all the simulated k-mers passing frequency filtering. The cut-offs used for each test are shown as red dashed lines. Top panel: marginal distribution of χ^2 p-values. Right panel: marginal distribution of logistic regression p-values. **a)** k-mers meeting the threshold for significance (a cut-off of 1×10^{-8}) in the logistic regression which map to the causal gene are coloured in red. **b)** shading of each point is by MAF. Most of the k-mers with a high χ^2 p-value and low logistic regression p-value are at low frequency, as are those with equal p-values from each test.

2.4.2 Downstream interpretation of significant k-mers

Significant k-mers can be interpreted directly through mapping to annotated genomes, or by assembling them first. Assembly may be better at searching for gene clusters associated with phenotype as longer and more specific k-mers will be generated. I assembled significant k-mers assembled using Velvet (Zerbino & Birney, 2008) choosing a smaller sub-k-mer size which maximises longest contig length of the final assembly. K-mers in the output which are substrings of other longer significant k-mers are removed.

I used BLAT (Kent, 2002) with a step size of 2 and minimum match size of 15 to find inexact but close matches to a well annotated reference sequence. Small k-mers are more likely than full reads to map equally well to multiple places in the reference genome, so reporting both mappings increases the sensitivity. For the tested dataset an average of 21% of k-mers significantly associated with antibiotic resistance report secondary mappings. These k-mers are short (median 15bp), and therefore have low specificity and high sensitivity as expected. I wrote a script which combines the p-values from SEER and co-ordinates from mapping of the significant k-mers into a .plot file, which can be loaded into visualisation software http://jameshadfield.github.io/phandango/ to create a Manhattan plot.

When k-mers do not map to a reference genome, I wrote the C++ program map_back to help interpret these. This reads in all the tested assemblies from which the k-mers were generated into memory, and threads are spawned which search for k-mers (and their reverse complement) by exact string match. Using the mapped co-ordinates, annotations of features in these regions can examined for overlap of function.

2.5 Development of SEER

I implemented SEER in C++ using the armadillo linear algebra library (Sanderson, 2010; Sanderson & Curtin, 2016), and dlib optimisation library (D. E. King, 2009). When the code was stable, I profiled its execution over a test dataset of 1 000 k-mers. Most of the processing time was spent evaluating the exp() function, which is required O(N) times per k-mer when calculating the likelihood function and its gradient during the logistic fit, where N is number of samples. I was satisfied that this demonstrated an efficient usage of CPU time, and further did not identify any memory leaks when profiling with valgrind.

For ease of deployment on non-cluster machines I also threaded each filtered k-mer's fitting routine; on four cores this achieved a 2.1 times speedup. While this could probably be improved by increasing the number of k-mers handled by each thread, the algorithm is embarrassingly parallel – in practise I split the k-mer file into 16 and ran an independent process on each one. I also threaded the calculation of entries in the distance matrix D, using mutex locks to ensure only one process wrote an entry to the matrix at a time. This

was over 99% efficient.

On my simulation of 3 069 diverse 0.4Mb genomes described in section 2.6.1, 143M k-mers were counted by DSM and 25M 31-mers by DSK. On the largest DSM set, using 16 cores and subsampling 0.3M k-mers (0.2% of the total), calculating population covariates took 6hr 42min and 8.33GB RAM. This step is $O(N^2M)$ where M is number of k-mers, but can be parallelised across up to N^2 cores.

Processing all 143M informative k-mers as described took 69min 44s and 23MB RAM on 16 cores. This step is $O(NM)$ and can be parallelised across up to M cores.

After the initial release I added the following features, fixes and improvements in response to user comments on github:

- Convergence errors and the type of regression used are added in a comment field for each k-mer.
- Created a virtual machine with SEER installed, without the requirement for further dependencies.
- Statically complied version (includes libraries in executable).
- Add scripts to map significant k-mers and create a Manhattan plot.
- An alternative implementation of the population structure correction, written in R.
- Tests of all features of SEER, and continuous integration of these through travis.
- Improved installation and usage instructions, including a self-contained tutorial.

2.6 Benchmarking SEER

I benchmarked the performance of SEER on three datasets. The first was a large simulated set of *S. pneumoniae* genomes where I was able to define the associated element and set its effect size manually – this allowed me to calculate the discovery power of SEER for different sample sizes under different situations. The second dataset was 3 069 real *S. pneumoniae* genomes with five antibiotic resistance phenotypes available which helped me evaluate whether SEER could capture both gene and SNP mediated resistances (which have large effect sizes, and are often homoplasic, so should be easy to find), and how SEER compares to previous methods. Finally, I tested SEER on 675 *S. pyogenes* genomes from invasive and non-invasive samples to see if SEER could discover any new associations with a clinically relevant phenotype other than resistance.

2.6.1 Simulated data

I used a framework similar to that described in section 2.3.1 to simulate genetic sequences. To make running the simulation tractable for such a large population size, I took a random subset of 450 genes from the *S. pneumoniae* ATCC 70066916 strain as the starting genome for ALF (Dalquen et al., 2012). Using the same parameters as in section 2.3.1 I simulated 3 069 final genomes along the phylogeny observed in a Thai refugee camp (Chewapreecha, Harris et al., 2014). pIRS (Hu et al., 2012) was again used to simulate error-prone reads from genomes at the tips of the tree, which I then assembled by Velvet (Zerbino & Birney, 2008). DSM was used to count k-mers from these de novo assemblies. I counted 143M informative k-mers from this simulated data, though on the real dataset of full length genomes only 68M informative k-mers were counted.

I used a gamma plus invariant sites model as the distribution of rate heterogeneity among sites. As I did not have estimates for the parameters of this distribution directly from the data, I used the estimate given by ALF. The resulting gamma distribution must have a longer tail than the real data, as some sites vary at high frequency. This created many low-frequency k-mers. As the simulation is computationally very expensive to run, I decided that rather than running it lots of times with different parameters until a k-mer distribution identical to the observed data was reached it would be sufficient to use the original result. The excess of low frequency k-mers would be filtered out in the common variation associations I am testing. 24.7M k-mers passed frequency filtering from the real data, whereas 12.7M passed from the simulated data – while this wasn't quite the linear scaling expected with genome length (which would predict around 7M k-mers) the amount of common variation at the gene level was similar to real data. For the purpose I used the simulations for, a gene driven association at different ORs, this result was still an appropriate test.

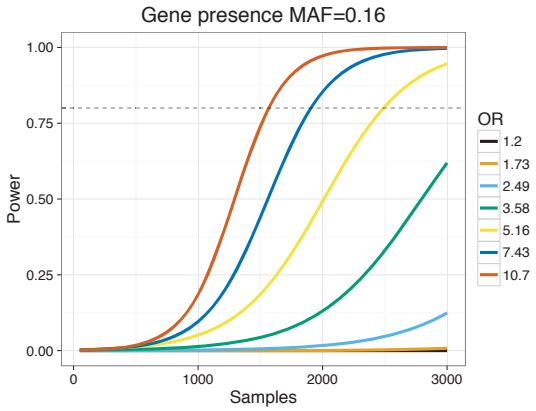
I then simulated the phenotype based on the genetic sequence. I set the ratio of cases to controls in the population (S_R) at 50% to represent typical antibiotic resistance, and designated a single variant (which could be either gene presence/absence or a SNP) as causal. MAF in the population is set from the simulation of genomes, and OR can be varied. The number of cases D_E is then the solution to a quadratic equation (Newman, 2003), which is related to probability of a sample being a case by

$$P(\text{case}|\text{major allele}) = \frac{D_E}{\text{MAF}} \tag{2.11}$$

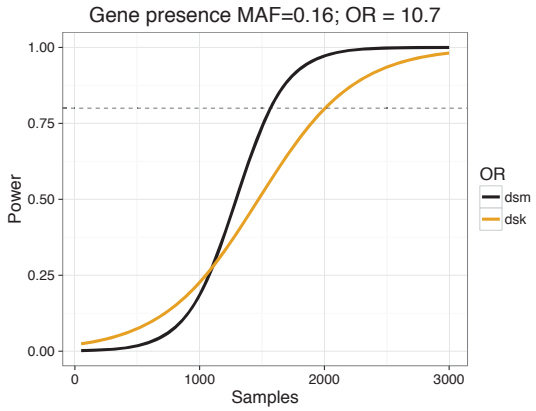
$$P(\text{case}|\text{minor allele}) = \frac{\frac{S_R}{S_R+1} - D_E}{1-\text{MAF}} \tag{2.12}$$

I generated random subsamples of the population 100 times at a range of sample sizes below the total, with case and control status assigned for each run using these formulae. I defined power by the proportion of runs that had at least one k-mer in the gene significantly

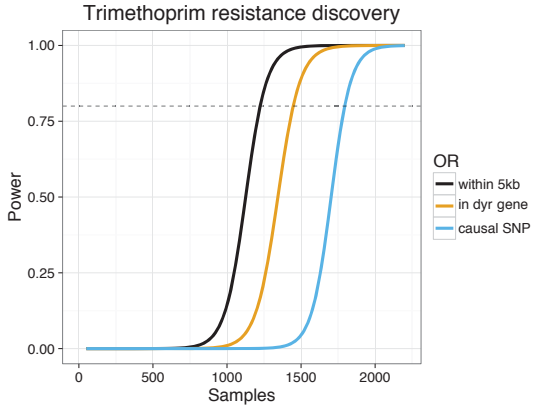
associated with the phenotype.



(a) Gene presence/absence at different odds-ratios.



(b) Using all informative k-mers versus a single length for gene mediated association.



(c) Detecting k-mers near, in the correct gene, or containing the causal variant for trimethoprim resistance.

Figure 2.6: Using simulations and subsamples of the population as described, power for detecting associations. All curves are logistic fits to the mean power over 100 subsamples.

Having knowledge of the true alignments, I then artificially associated an accessory gene with a phenotype over a range of odds-ratios and evaluated power at different sample sizes (fig. 2.6a). The expected pattern for this power calculation is seen, with higher odds-ratio effects being easier to detect. Currently detected associations in bacteria have had large effect sizes (OR $>$ 28 host-specificity (Sheppard et al., 2013); OR $>$ 3 beta-lactam resistance (Chewapreecha, Marttinen et al., 2014)), and the required sample sizes predicted are consistent with these discoveries.

The large k-mer diversity, along with the population stratification of gene loss, makes the simulated estimate of the sample size required to reach the stated power conservative. Convergent evolution along multiple branches of a phylogeny for a real population reacting to selection pressures will reduce the required sample size (Farhat et al., 2013).

I also compared the performance when using k-mers counted at constant lengths by DSK (Rizk et al., 2013) to perform the gene presence/absence association. Counting all informative k-mers rather than a pre-defined k-mer length gave greater power to detect

associations, with 80% power being reached at around 1 500 samples, compared with 2 000 samples required by 31-mers (fig. 2.6b). The slightly lower power at low sample numbers is due to a stricter Bonferroni adjustment being applied to the larger number of DSM k-mers over the DSK k-mers. This is exactly the expected advantage from including shorter k-mers to increase sensitivity, but as k-mers are correlated with each other due to evolving along the same phylogeny, using the same Bonferroni correction for multiple testing does not decrease specificity.

The strong LD caused by the clonal reproduction of bacterial populations means that non-causal k-mers may also appear to be associated. This is well documented in human genetics; non-causal variants tag the causal variant increasing discovery power, but make it more difficult to fine-map the true link between genotype and phenotype (Spain & Barrett, 2015). In simulations it is difficult to replicate the LD patterns observed in real populations, as recombination maps for specific bacterial lineages are not yet known. To evaluate the power of fine-mapping and associated locus to the single causal SNP I instead used the real sequence data and the effect size of a known causal variant, and evaluated the physical distance of significant k-mers from the variant site.

I tested the 68M k-mers from DSM for association with trimethoprim resistance: 2 639 k-mers reached significance, were mapped to a reference genome, and were found to cover most of the genome with a peak at the causal variant (fig. 2.7). I placed mapped k-mers near the correct physical location into three categories: those containing the causal variant I100L (10 k-mers), those within the same gene (74 k-mers), or those within 2.5kb in either direction (207 k-mers). Figure 2.6c shows the resulting power when random subsamples of the population are taken. As expected, power is higher when not specifying that the causal variant must be hit, as there are many more k-mers which are in LD with the SNP than directly overlapping it, thus increasing sensitivity.

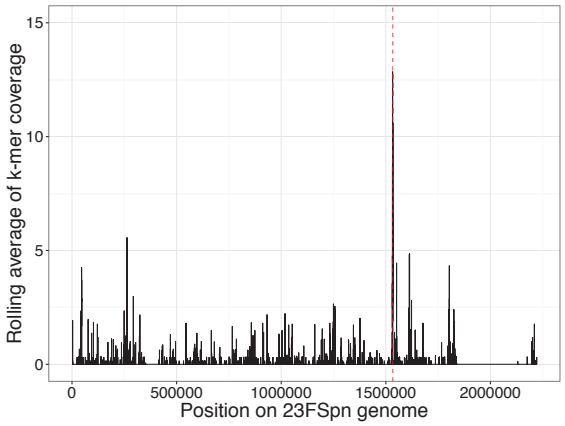


Figure 2.7: K-mers are mapped to the ATCC 700669 reference genome. Plotted coverage is the rolling average over 100bp windows over the genome. The red dashed line at 1 533 003bp shows the location of the causal variant, overlapping with the peak in coverage.

2.6.2 Antibiotic resistance in pneumococcal carriage

I then applied SEER to the sequenced genomes from the study described in section 2.6.1 (Chewapreecha, Harris et al., 2014), using measured resistance to five different antibiotics as the phenotypes: chloramphenicol, erythromycin, β−lactams, tetracycline and trimethoprim. Chloramphenicol resistance is conferred by the *cat* gene, and tetracycline resistance is conferred by the *tetM* gene, both carried on the ICE ICESp23FST81 in the *S. pneumoniae* ATCC 700669 chromosome (Croucher et al., 2009). For both of these drug resistance phenotypes the ICE contained 99% of the significant k-mers, and the causal genes rank highly within the clusters (table 2.2).

Antibiotic	Resistant samples	Number of significant k-mers			
		Total	Mapped to reference	Highest coverage annotation	Causal element
Chloramphenicol	204 (7%)	1 526	1 526	1 508 – ICE	166 – *cat*
				288 – ORF (UniParc B8ZK82)	
				206 – *rep*	
				166 – *cat*	
Erythromycin	803 (26%)	1 154	112	10 – permease (UniParc B8ZKV5)	4 – mega element
				8 – *prfC*	2 – *mef*
				6 – *gatA*	2 – omega element
				4 – ICE	
β−lactams	1 563 (51%)	23 876	17 453	381 – ICE	47 – *pbp2x*
				145 – prophage MM1	20 – *pbp2b*
				50 – SPN23F15110 (UniParc B8ZLE7)	8 – *pbp1a*
				49 – ICE *orf16*	
Tetracycline	1 958 (64%)	962	962	962 – ICE	96 – *tetM*
				136 – ICE *orf16*	
				121 – ICE *orf15*	
				96 – *tetM*	
Trimethoprim	2 553 (83%)	2 639	210	21 – *dyr*	21 – *dyr*

Table 2.2: Results from SEER for antibiotic resistance binary outcome on a population of 3 069 *S. pneumoniae* genomes. Significant k-mers were first interpreted by mapping to the ATCC 700669 reference genome. Up to the first four highest covered annotations are shown, and if the known mechanism is amongst these it is highlighted in orange. The ICE is the top hit in three analyses, as it carries multiple drug-resistance elements and is commonly found in multi-drug resistant strains (Croucher et al., 2009).

Resistance to erythromycin is also conferred by presence of a gene, but there are multiple genes that can be causal for this resistance: *ermB* causes resistance by methylating rRNA whereas *mef/mel* is an efflux pump system (Croucher, Harris, Fraser et al., 2011). In this population, this phenotype was strongly associated with two large lineages, making the task of disentangling association with a lineage versus a specific locus more difficult. I mapped some of the significant k-mers to the mega and omega cassettes, which carry the *mel/mef* and *ermB* resistance elements respectively.

I also mapped hits to other sites within the ICE, a permease directly upstream of *folP*, *prfC* and *gatA*. Macrolide resistance cassettes frequently insert into the ICE in *S. pneumoniae*, so it is in LD with the genes discussed above. In sulphamethoxazole resistance *folP* is modified by small insertions, with which the adjacent permease is in LD

with. Finally, *prfC* and *gatA* are both involved in translation, so could conceivably contain compensatory mutations when *ermB* mediated resistance is present. Further evidence of these compensatory mutations would be required to rule out the k-mers mapping to them simply being false positives driven by population structure.

Some k-mers did not map to the reference, as they are due to lineage specific associations with genetic elements not found in the reference strain. This highlighted both the need to map to a close reference or draft assembly to interpret hits described in section 2.4.2, as well as the importance of functional follow-up to validate potential hits from GWAS methods such as SEER.

Multiple mechanisms of resistance to $\beta-$lactams are possible (Chewapreecha, Marttinen et al., 2014). I considered just the most important (i.e. highest effect size) mutations, which are SNPs in the penicillin binding proteins *pbp2x*, *pbp2b* and *pbp1a*. In this case ranking annotations by highest coverage found these genes ranked top, but this was not sufficient evidence for discovery as so many k-mers were significant – either due to other mechanisms of resistance, physical linkage with causal variants or co-selection for resistance conferring mutations. Instead, I looked at the k-mers with the most significant p-values: the top four hit loci were *pbp2b* ($p = 10^{-132}$), *pbp2x* ($p = 10^{-96}$), putative RNA pseudouridylate synthase – UniParc B8ZPU5 ($p = 10^{-92}$) and *pbp1a* ($p = 10^{-89}$). The non-*pbp* hit is a homologue of a gene in linkage disequilibrium with *pbp2b*, which would suggest mismapping rather than causation of resistance.

Trimethoprim resistance in *S. pneumoniae* is conferred by the I100L mutation in the *folA*/*dyr* gene (Maskell et al., 2001). The *dpr* and *dyr* genes, which are adjacent in the genome, had the highest coverage of significant k-mers (fig. 2.8). To try and find the specific variant causal for the phenotype (i.e fine-mapping) I used the BLAT mapping of significant k-mers to a reference sequence, and called SNPs using bcftools (H. Li, 2011). I set quality scores for a read to be identical, as the Phred-scaled Holm-adjusted p-values from association. I then filtered for high quality ($QUAL > 100$) SNPs, and then annotated the predicted effect using SnpEff (Cingolani et al., 2012). I finally ranked the effect of missense SNPs on protein function using SIFT, which uses whether sites are conserved across the protein family to predict whether amino acid changes will alter protein function (Ng & Henikoff, 2003). Following this fine-mapping procedure, I called four high-confidence mutations that are predicted to be non-synonymous SNPs. One is the causal SNP, and the others appear to be hitchhikers in LD with I100L. The SIFT ranking places the known causal SNP top, showing that in this case fine-mapping is possible using the output from SEER.

I compared the performance of SEER to two existing methods. Chewapreecha, Marttinen et al. (2014) tested variants from a core-genome SNP mapping using plink (Purcell et al., 2007); population clusters were used to perform a CMH test to control for population structure. Sheppard et al. (2013) used fixed k-mer lengths of 21, 31 and 41 as counted

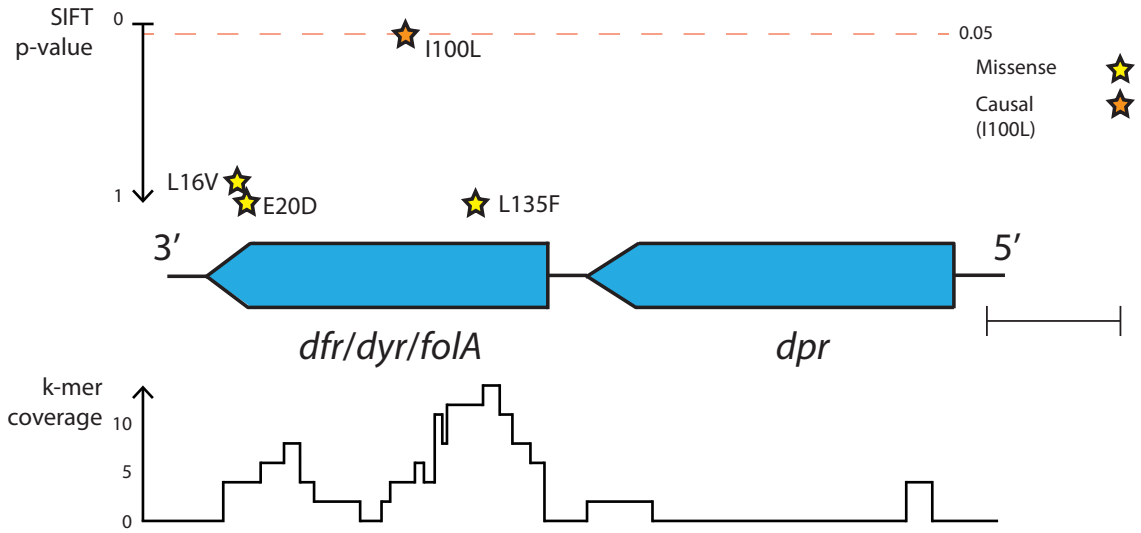


Figure 2.8: Fine mapping the causal variant for trimethoprim resistance. The locus pictured contains 72 significant k-mers, the most of any gene cluster (fig. 2.7). Coverage over the locus is pictured at the bottom of the figure. Shown above the genes are high quality missense SNPs, plotted using their p-value for affecting protein function as predicted by SIFT. Scale bar is 200 base pairs.

by DSK (Rizk et al., 2013), with a Monte Carlo phylogeny-based population control. As the second method is not scalable to this population size, I used the SEER population control as calculated from all genomes in the population and a subsample of 100 samples to calculate association statistics, which is roughly the number computationally accessible by this method. In both cases, the same Bonferroni correction is used as for SEER.

Antibiotic	Causal variant	Significant sites		Near correct site			Notes
		plink	dsk	plink			
Tetracycline	ICE, *tetM*	8 029	0	*tetM* – 124	ICE – 2240		
Chloramphenicol	ICE, *cat*	5 310	0	*cat* – 0	ICE – 1137		
β−lactams	*pbp2x*, *pbp1a*, *pbp2b*	858	0	*pbp2x* – 210	*pbp1a* – 113	*pbp2b* – 81	
Trimethoprim	*dyr* (I100L)	4 009	0	*dyr* – 47	*dpr* – 53		Causal SNP ranked 22nd
Erythromycin	*ermB*, *mef*, *mel*, *mefA*	8 469	0	None			Element not present in reference

Table 2.3: The power to find genetic associations with antibiotic resistance in the Maela study using existing methods. For each of the five antibiotics, the true causal variant is listed, as are the number of hits passing the significance threshold for each method (plink and DSK) and the number which map to the correct region.

Both SEER and association by core mapping of SNPs (using plink) identified resistances caused by presence of a gene, when it was present in the reference used for mapping (table 2.3). Both produced their most significant p-values in the causal element, though SEER appeared to have a lower false-positive rate. However, as demonstrated by chloramphenicol resistance, if not enough SNP calls are made in the causal gene this hinders fine-mapping. SNP-mediated resistance showed the same pattern since many other SNPs were ranked above the causal variant. In the case of β−lactam resistance both methods seem to perform equally well, likely due to the higher rate of recombination and the creation of mosaic *pbp* genes.

Additionally, as for erythromycin resistance, when an element is not present in the

reference it is not detectable in SNP-based association analysis. In such cases multiple mappings against other reference genomes would have to be made, which is a tedious and computationally costly procedure. Since the k-mer results from SEER are reference-free, the computational cost of mapping reads to different reference genomes is minimised as only the significant k-mers are mapped to all available references. Alternatively, the significant k-mers can be mapped to all draft assemblies in the study, at least one of which is guaranteed to contain the k-mer, to check if any annotations are overlapped.

The small sample, combined with fixed length 31-mer, approach did not lead to any words reaching significance for chloramphenicol, tetracycline or trimethoprim as the effect size of any k-mer is too small to be detected in the number of samples accessible by the method. I found 19 307 hits for erythromycin, and 419 hits for $\beta-$lactams, at between 1-2% MAF which are all false positives that would likely have been excluded by a fully robust population structure correction method such as the one the authors originally used.

2.6.3 Virulence of *Streptococcus pyogenes*

Most bacterial GWAS studies to date have searched for genotypic variants that contribute towards or completely explain antibiotic resistance phenotypes. As a proof of principle that SEER could be used for the discovery stage of sequence elements associated with other clinically important phenotypes, I applied the tool to 675 *S. pyogenes* (group A *Streptococcus*) genomes obtained from population diversity studies for genetic signatures of invasive propensity.

347 isolates of *S. pyogenes* collected from Fiji (Steer et al., 2009) were sequenced on the Illumina HiSeq platform, which I then combined with 328 existing sequences from Kilifi, Kenya (Seale et al., 2016). I defined those isolated from blood, CSF or bronchopulmonary aspirate as invasive (n = 185), and those isolated from throat, skin or urine as non-invasive (n = 490). I then ran SEER to determine k-mers significantly associated with invasion, followed by a BLAST of the k-mers with the nr/nt database to determine a suitable reference for mapping purposes.

After this preliminary analysis, I found the top hit was the *tetM* gene from a conjugative transposon (*Tn*916) carried by 23% of isolates (fig. 2.9a). These elements are known to be variably present in the chromosome of *S. pyogenes* (Roberts & Mullany, 2009), and the lack of co-segregation with population structure explained the power to discover the association. However, as a different proportion of the isolates from each collection were invasive (Fiji – 13%; Kilifi – 43%), the significant k-mers will also include elements specific to the Kilifi dataset. Indeed, I found that this version of *Tn*916 was never present in genomes collected from Fiji. To correct for this geographic bias, I repeated the SEER analysis by including country of origin as a covariate in the regression. This analysis removed *tetM* as being significantly associated with invasiveness, and highlighted the

importance of such covariate considerations in performing association studies on large bacterial populations.

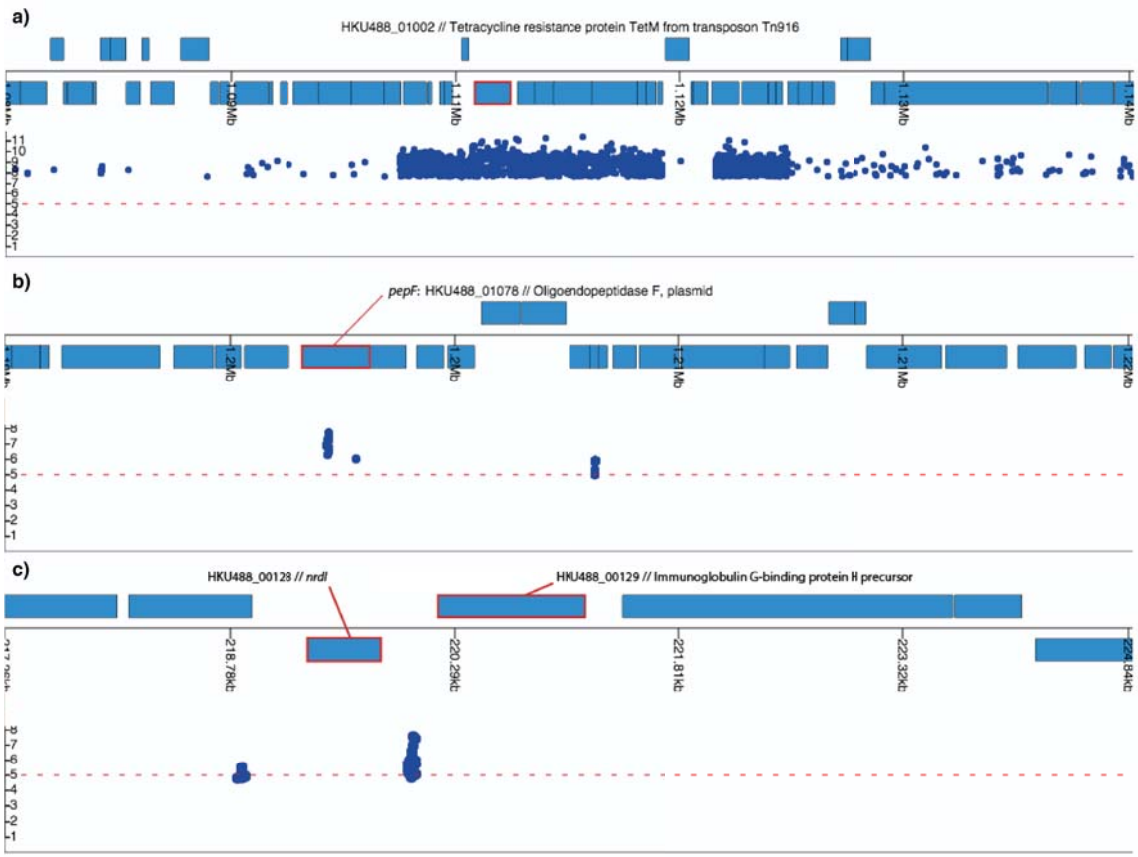


Figure 2.9: Phandango view of *S. pyogenes* HKU488 reference genome (blue blocks at top genes on forward and reverse strands, *tetM* highlighted in red) and Manhattan plot of start positions of significant k-mers: **a)** associated with invasiveness when not adjusted for country of origin; **b)** and **c)** adjusted for country of isolation.

After applying this correction, I identified two significant hits (fig. 2.9b,c). The first corresponded to SNPs associating a specific allele of *pepF* (Oligoendopeptidase F; UniProt P54124) with invasive isolates. This could indicate a recombination event, due to the high SNP density and discordance with vertical evolution with respect to the inferred phylogeny (Dubnau, 1999; Lefébure & Stanhope, 2007). The second hit represented SNPs in the intergenic region upstream of both IgG-binding protein H (*sph*) and *nrdI* (ribonucleotide reductase). In support of these findings, previous work in murine models have found differential expression of *sph* during invasive disease (Raeder & Boyle, 1993, 1995; T. C. Smith et al., 2003b), but little to no expression outside of this niche (T. C. Smith et al., 2003a). If these k-mers were found to affect expression of the IgG-binding protein, this would be a plausible genetic mechanism affecting pathogenesis and invasive propensity (Walker et al., 2014). The association of both of these variations would have to be validated either in vitro or a within a replication cohort, and functional follow-up such as RNA-seq may also help with determining the role of these genetic variants in

S. pyogenes pathogenesis.

In contrast, when I applied existing association methods described above (plink and DSK) to this *S. pyogenes* population dataset I found no sites significantly associated with invasiveness. The CMH test (stratified by BAPS cluster) that uses SNPs called against a reference sequence failed to identify the *tetM* gene and transposon as these elements are not found in the reference sequence. Furthermore, the population structure of this dataset is so diverse that 88 different BAPS clusters were found, which overcorrected for population structure when using the DSK method, leaving too few samples within each group to provide the power to discover associations.

2.7 Conclusions

SEER is a reference-independent, scalable pipeline capable of finding bacterial sequence elements associated with a range of phenotypes while controlling for clonal population structure. The sequence elements can be interpreted in terms of protein function using sequence databases, and I have shown that even single causal variants can be fine-mapped using the SEER output.

My use of all informative k-mers less than 100 bases long, a robust regression protocol and the ability to analyse very large sample sizes showed improved sensitivity over existing methods. This provides a generic approach capable of analysing the rapidly increasing number of bacterial whole genome sequences linked with a range of different phenotypes. The output can readily be used in a meta-analysis of sequence elements to facilitate the combination of new studies with published data, increasing both discovery power and confirming the significance of results.

As with all association methods, the approach is limited by the amount of recombination and convergent evolution that occurs in the observed population, since the discovery of causal sequence elements is principally constrained by the extent of LD. However, by introducing improved computational scalability and statistical sensitivity SEER improved on previous GWAS methods for answering important biologically and medically relevant questions.

In subsequent chapters I will start by using the GWAS techniques developed here to assess the contribution of bacterial variation to various stages of pneumococcal infection.

Chapter 3

Variation in duration of asymptomatic pneumococcal carriage

Declaration of contributions

Stephen Bentley, Paul Turner, Nicholas Croucher and Julian Parkhill supervised this work. Paul and Claudia Turner designed and ran the Maela study on which this work is based, and provided the swabbing data from which carriage duration was inferred. Susannah Salter for provided data on non-typable culture positive rates. I performed all analyses.

Publication

The following has been submitted as:
Lees, J. A., Croucher N. J., Goldblatt D., Nosten F., Parkhill J., Turner C., Turner P., Bentley, S. D. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife*.
Deposited on *bioRxiv*: 107086. https://doi.org/10.1101/107086

3.1 Introduction

In chapter 2 I developed a method and piece of software to perform GWAS on bacterial populations. The main test of SEER was finding known antibiotic resistance determinants. These are one of the easiest GWASs to perform in bacteria, as the effect size of these variants is so high (close to fully penetrant, hence the need to use Firth regression in some cases) and the selection pressure over time has led to the causal variants being homoplasic and broadly spread evenly across the population. In this chapter I test the method on a phenotype likely to be polygenic in origin, with causal variants that are both population stratified (lineage effects) and independent of population structure (locus effects) (Earle et al., 2016).

S. pneumoniae spends most of the transmission cycle in the nasopharynx, and so understanding and predicting the amount of time spent in this niche is critical for understanding this bacterium's epidemiology, and therefore controlling transmission (Abdullahi et al., 2012a; Melegaro et al., 2007). The nasopharynx is a complex niche in which each pneumococcal genotype must tackle a wide range of factors including host immune defence (McCool et al., 2002), other bacterial species (Pericone et al., 2000), and other pneumococcal lineages (Auranen et al., 2010; Cobey & Lipsitch, 2012) in order to maintain the genotype's population. The average nasopharyngeal duration period is therefore affected by a large number of factors, which may, themselves, interact.

A major potential advantage of GWAS in bacteria is the ability to test association with less well defined phenotypes, for example transmissibility (Nebenzahl-Guimaraes et al., 2016), or phenotypes which would be difficult to test in a lab. Here I assess genetic variation associated with pneumococcal carriage duration. Traditionally this would be difficult to assess due to the complexity of the nasopharyngeal niche, and the length of time experiments would need to be run for.

One factor that is known to strongly associate with carriage duration is serotype: as capsular polysaccharides are important in bacterial physiology and determining host immune response, different serotypes have different clearance and acquisition rates (Abdullahi et al., 2012a; P. C. Hill et al., 2010; Högberg et al., 2007; Melegaro et al., 2007; P. Turner et al., 2012). Additionally, a range of other proteins have been identified as critical to the colonisation process (Kadioglu et al., 2008), some of which exhibit similar levels of diversity to the capsule polysaccharide synthesis locus (Iannelli et al., 2002; Jedrzejas et al., 2001). However, the overall and relative contributions of these sequence variations to carriage rate have not yet been characterised. In addition variation of pathogen protein sequence, accessory genes and interaction effects between genetic elements may also have as yet unknown effects on carriage duration.

Changes in average carriage duration have been shown to be linked with recombination rate (Chaguza et al., 2016), which has been found to correlate with antibiotic resistance

(Hanage et al., 2009) and invasive potential (Chaguza et al., 2016). The carriage duration by different serotypes is widely used in models of pneumococcal epidemiology, and consequently is important in evaluating the efficacy of the PCV (Melegaro et al., 2007; Weinberger, Harboe et al., 2011). Additionally, modelling work has proposed that if alleles exist which alter carriage duration, these explain the long standing puzzle of how antibiotic-resistant and sensitive strains stably coexist in the population (Lehtinen et al., 2017). Measurement of carriage duration and the analysis of its variance beyond the resolution of serotype will have important consequences for these models.

I sought to determine the overall importance of the pathogen genotype in carriage duration in a human population, and to identify and quantify the elements of the genome responsible for the variation in carriage duration using GWAS. By combining epidemiological modelling of longitudinal swab data with and genome wide association study methods on the connected sequences, I made heritability estimates for carriage duration. I further partitioned the heritability into contributions from lineage and locus effects to quantify the variation caused by each individual factor.

3.2 Ascertainment of carriage episode duration using epidemiological modelling

I first estimated carriage duration from longitudinal swab data available for the study population. For 598 unvaccinated children up to 24 swabs taken over a two year period were available. The study population was a subset of infants from the Maela longitudinal birth cohort (C. Turner et al., 2013), and was split into two cohorts. In the 'routine' cohort, 364 infants were swabbed monthly from birth, 24 times in total. All swabs had been cultured and serotyped using the latex sweep method (P. Turner et al., 2013). In the 'immunology' cohort 234 infants were swabbed on the same time schedule, but cultured and serotyped following the World Health Organisation (WHO) method (P. Turner et al., 2012). NT pneumococci had been confirmed by bile solubility, optochin susceptibility and Omniserum Quellung negative.

I only considered swabs from infants in the study, as mothers did not have sufficient sampling resolution relative to their average length of carriage to determine carriage duration. Furthermore, the immune response of mothers to bacterial pathogens is different to children (Maródi, 2006), leading to shorter carriage durations (Gritzfeld et al., 2014).

To estimate carriage duration from the longitudinal swab data I constructed a set of hidden Markov models (HMMs) with hidden states corresponding to whether a child was carrying a serotype at a given time point, and observed states corresponding to whether a positive swab was observed for this serotype at this time point. The most general model for the swab data would be a vector with an entry of 0 or 1 for every possible serotype (of

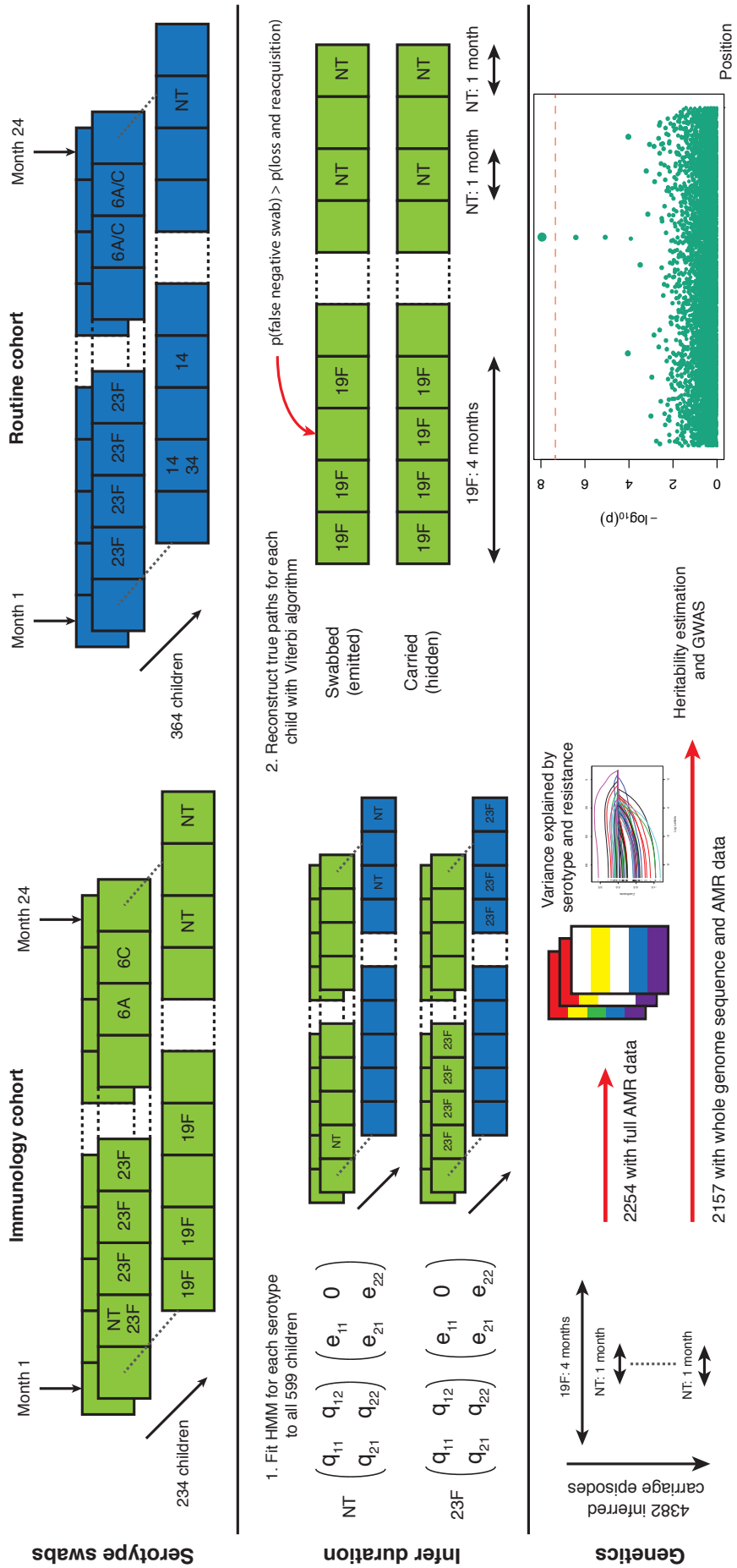


Figure 3.1: Swabbing and sequencing study design. I started with serotype swab data on 598 children from two cohorts, taken every month after birth for two years. For all samples I fitted the transition and emission probabilities of a continuous time hidden Markov model for each serotype. Then, for each child, I used these parameters were then used to infer the most likely carriage durations. I matched carriage episodes with resistance and genomic data for 2 157 episodes to draw conclusions on the basis of variation in this epidemiological parameter.

56 observed in the population), corresponding to whether each serotype was observed in the swab at each time point. However, the number of parameters to estimate in this model (with over 6 million states) is much larger than the number of data points (around 14000), and in particular some serotypes have very few positive observations. Instead, I modelled each serotype separately.

The models fitted, and their permitted transitions and emissions are shown in fig. 3.2. In model one, observation i emits state 2 if positively swabbed for the serotype, and state 1 otherwise. The unobserved states correspond to the child 'carrying' and being 'clear' of the serotype respectively. I assumed swabs have a specificity of one, so do not show positive culture when the child is clear of the carried serotype; I therefore set the coefficient for the chance of observing positive culture when no bacteria are present to zero ($e_{21} = 0$ in the emission matrix). Model two added a third state of 'multiple carriage' which is occupied when the serotype and at least one other are being carried. Both models were compared with a version which allows the parameters to covary with whether the child has carried pneumococcus previously. In model three I accounted for this explicitly by having separate states and emissions based on whether carriage has previously been observed.

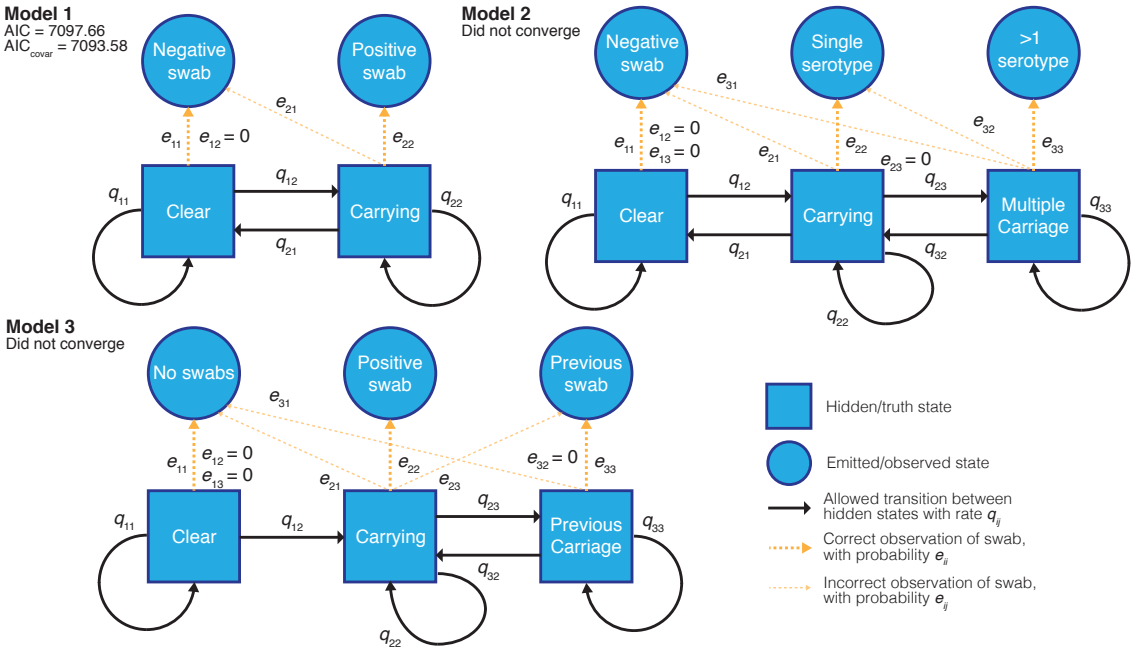


Figure 3.2: HMMs of swab time series, and their goodness-of-fit. I fitted three different models to the processed time-series data with states, allowed transitions and emissions as shown. I refitted each model allowing the transitions probabilities to covary with the age of the child and whether the child had carried pneumococcus previously. For the converged model the Akaike information criterion (AIC) is shown for the original fit, and when including these covariates (AIC_{covar}).

I modelled the time series of swab data using a continuous-time HMM, as implemented in the R package `msm` (Jackson, 2011). Unobserved (true) states correspond to whether the child is carrying bacteria in their nasopharynx, and observed (emitted) states correspond to whether a positive swab was seen at each point. Transition probabilities between each state

Q and the emission probabilities **E** were jointly estimated by maximum likelihood using the BOBYQA algorithm. To get a good fit of the HMM, I normalised observation times for each sample. Defining infant birth as $t = 0$, subsequent sampling times t_i were measured in days, and normalised to have a variance of one. I then constructed the most likely path through the unobserved states for each child using the Viterbi algorithm (Forney, 1973) with the observed data and estimated model parameters. Assuming that continuous occupation of the carried state corresponded to a single carriage episode, I calculated the duration for each such episode from the inferred true states.

I applied all three models to 19F carriage episodes, as these had the most data available, and calculated the AIC (Akaike, 1974) for each model that converged. Only the simplest model (model one) converged, as judged by having a positive-definite Hessian and a converged BOBYQA run. The more complex models had lower log-likelihoods: as extensions of the simpler model they should have higher log-likelihoods, so this result was not consistent with model convergence. I tried fitting models two and three using a fixed false positive values slightly greater than zero: this lead to better log-likelihoods, but the models still didn't converge. This failure of the more complex models is probably because most children in the study immediately enter the carrying state, and episodes of dual carriage (when split up by serotype) are rare. Therefore there were not enough events between these carriage states to estimate to the transition and emission intensities, without sensitivity to initial conditions during the fitting.

I then fitted the best performing model in this test for all serotypes separately. Latex sweeps could not differentiate 6A and 6C serotypes, so I treated these as a single serotype (in WHO serotyping PCR was used to differentiate these serotypes, but I still combined them for consistency across the two cohorts). 15B and 15C serotypes spontaneously interconvert, so were combined. I also removed two duplicated swabs (08B09098 from the immunology cohort; 09B02164 from the routine observation cohort). The models for 19F, 23F, 6A/C, 6B, 14 and NT converged, but other serotypes did not have enough observations to successfully fit the parameters of the model. For these less prevalent serotypes I used the transition and emission parameters from the 19F model fitted with the correct observations when reconstructing the most likely route taken through the hidden states. I manually inspected the results to ensure this did not cause systematic overestimation when compared with previous studies.

I found that the fit for NT swabs produced results which overestimated carriage duration when compared to previously reported estimates. The best fit to the model overestimated the e_{21} parameter, which measures the false negative rate of swabbing, in favour of reduced transition intensities. I therefore fitted the model again, fixing this rate at 0.12. I based this figure on non-typable *S. pneumoniae* abundance as defined by 16S survey sequencing. At 1% proportional abundance in the sample, 12% came out as culture negative (table 3.1).

Abundance	Culture positive	Number
>1%	Cultured	361
>1%	Not cultured	44
<1%	Cultured	56
<1%	Not cultured	54

Table 3.1: Success of culturing unencapsulated *S. pneumoniae*. Based on having >1% abundance of 16S reads showing the bacteria as being present, 44/361 true positive swabs were not successfully cultured.

3.2.1 Combining epidemiological data with genomic data

From all the swab data, I estimated that there were a total of 4 382 carriage episodes (7.3 per child), of which 2 254 had a complete set of AMR data available (fig. 3.3). After removing ten outlier observations (fig. A.3) from swabs taken accidentally during disease, I was able to match 2 157 sequenced genomes with a carriage duration.

As I aimed to fit a multiple linear regression model to the carriage duration y against binary lineage associated predictors, I first ensured the data was appropriate for this model. The phenotype distribution was positively skewed, with an approximately exponential distribution. Residuals were therefore non-normally distributed, potentially reducing power (McCulloch, 2003). In the regression setting, a monotonic function can be applied to transform the response variable to avoid this problem. I first took the natural logarithm of the carriage duration

$$\hat{\mathbf{y}} = \ln(\mathbf{y})$$

which led to the residuals being much closer to being normally distributed (figs. 3.3 and A.2). I applied the same transformation to child age, when it was used as a covariate in association. For association with a LMM I instead took a monotonic transform of the carriage duration using `warped-lmm` (Fusi et al., 2014) to maximise the study's power to discover associations and estimate heritability (figs. A.1 and A.2). This used a sum over three nonlinear step functions, plus a linear term, to transform the residuals into Gaussians (Snelson et al., 2004).

For each isolate with an inferred carriage duration I extracted SNPs from the previously generated alignment against the ATCC 700669 genome (Chewapreecha, Marttinen et al.,

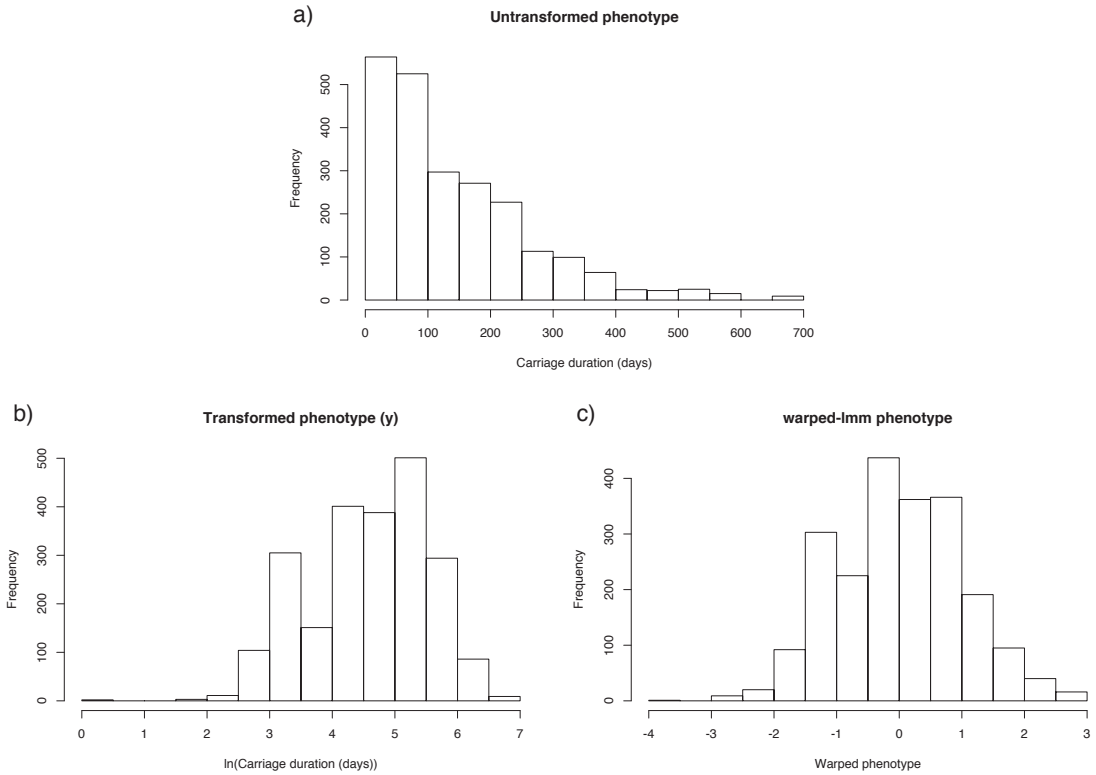


Figure 3.3: Distribution of carriage duration, and effect of monotonic transformation. Panel **a)** shows a histogram of the inferred carriage duration, **b)** shows this result after the natural logarithm is taken, and **c)** after the warping function is applied.

2014). Consequences of SNPs were annotated with `VEP`, using a manually prepared reference (McLaren et al., 2010). I generated a phylogenetic tree from this alignment using `FastTree` under the GTR+gamma model (M. N. Price et al., 2009). The carriage duration was mapped on to this phylogeny using `phytools` (Revell, 2013). I then filtered the sites in the alignment to remove any where the major allele was an 'N', any sites with a minor allele frequency lower than 1%, and any sites where over 5% of calls were missing. This left 115 210 sites for association testing and narrow-sense heritability estimation. I also used the 68M non-redundant k-mers with lengths 9-100 from the de novo assemblies of the genomes counted in section 2.2. I filtered out low frequency variants by removing any k-mers with a minor allele frequency below 2%, leaving 17M for association testing.

3.3 Overall heritability of carriage duration is high

To recap section 1.3.2, the variation in carriage duration σ^2_P is partly caused by variance in pneumococcal genetics, and variance in other potentially unknown factors such as host age and host genetics. It is common to write this sum as two components: genetic effects σ^2_G and environmental effects σ^2_E. The proportion of the overall variation which can be explained by the genetics of the bacterium is known as the broad-sense heritability

$H^2 = \frac{\sigma^2_G}{\sigma^2_G+\sigma^2_E}$. Variants which are directly associated with carriage duration independently of other variants (non-epistatic effects) contribute to the narrow-sense heritability h^2, which is smaller than the overall broad-sense heritability (Visscher et al., 2008).

H^2 can be estimated by linear regression on the phenotype of donor-recipient pairs which nearly share their genetics (Fraser et al., 2014). However in this dataset previous work was only able to confidently identify five transmission events, which was not enough to apply this method. Alternatively, analysis of variance of the phenotype between pathogens with similar genetics can be used to estimate heritability (T. J. C. Anderson et al., 2010). By applying this to phylogenetically similar bacteria (fig. 3.4), I estimated broad sense heritability H^2 with the ANOVA-CPP method in the `patherit` R package (Mitov & Stadler, 2016), using a patristic distance cutoff of 0.04 (fig. A.4). This estimated that $H^2 = 0.634$ (95% CI 0.592-0.686), implying that the genetics of *S. pneumoniae* is an important factor in determining carriage duration in this population. If environmental conditions are associated with streptococcal genotype between populations (such as host vaccination status) the heritability estimate may differ.

A lower bound on h^2 can be calculated by fitting a LMM through maximum likelihood to common SNPs (h^2_{SNP}) (S. H. Lee et al., 2011; Manolio et al., 2009). I used the 'GCTA' model implemented in `warped-lmm` (Fusi et al., 2014) to estimate h^2_{SNP} for carriage duration data, using the filtered SNPs and including child age and previous carriage as covariates. This yielded an estimate of 0.445, consistent with the estimate for H^2. I also estimated h^2_{SNP} using `LDAK` (Speed et al., 2012) with default settings, which gave an estimate of 0.437 ($<$1% difference from the `warped-lmm` estimate).

3.4 Lineage effects on carriage duration

After calculating the overall heritability, I wished to determine the amount that the specific variation in the pathogen genome contributes to changing carriage duration. However the strong LD present across the entire genome of *S. pneumoniae*, makes it difficult to pinpoint variants associated with carriage duration and not just present in the background of longer or shorter carried lineages (P. E. Chen & Shapiro, 2015). Serotype and antibiogram are correlated with the overall genome sequence (Brueggemann et al., 2003; Chewapreecha, Harris et al., 2014; Enright & Spratt, 1998), so if these factors are associated with carriage duration, large sets of variants which define long-carried and short-carried lineages will be correlated with carriage duration in a naive association test (P. E. Chen & Shapiro, 2015; T. D. Read & Massey, 2014).

I use the distinction between variants which evolve convergently and affect a phenotype independently of lineage – termed locus effects – to those which are collinear with a genotype which is associated with the phenotype, termed lineage effects (Earle et al., 2016).

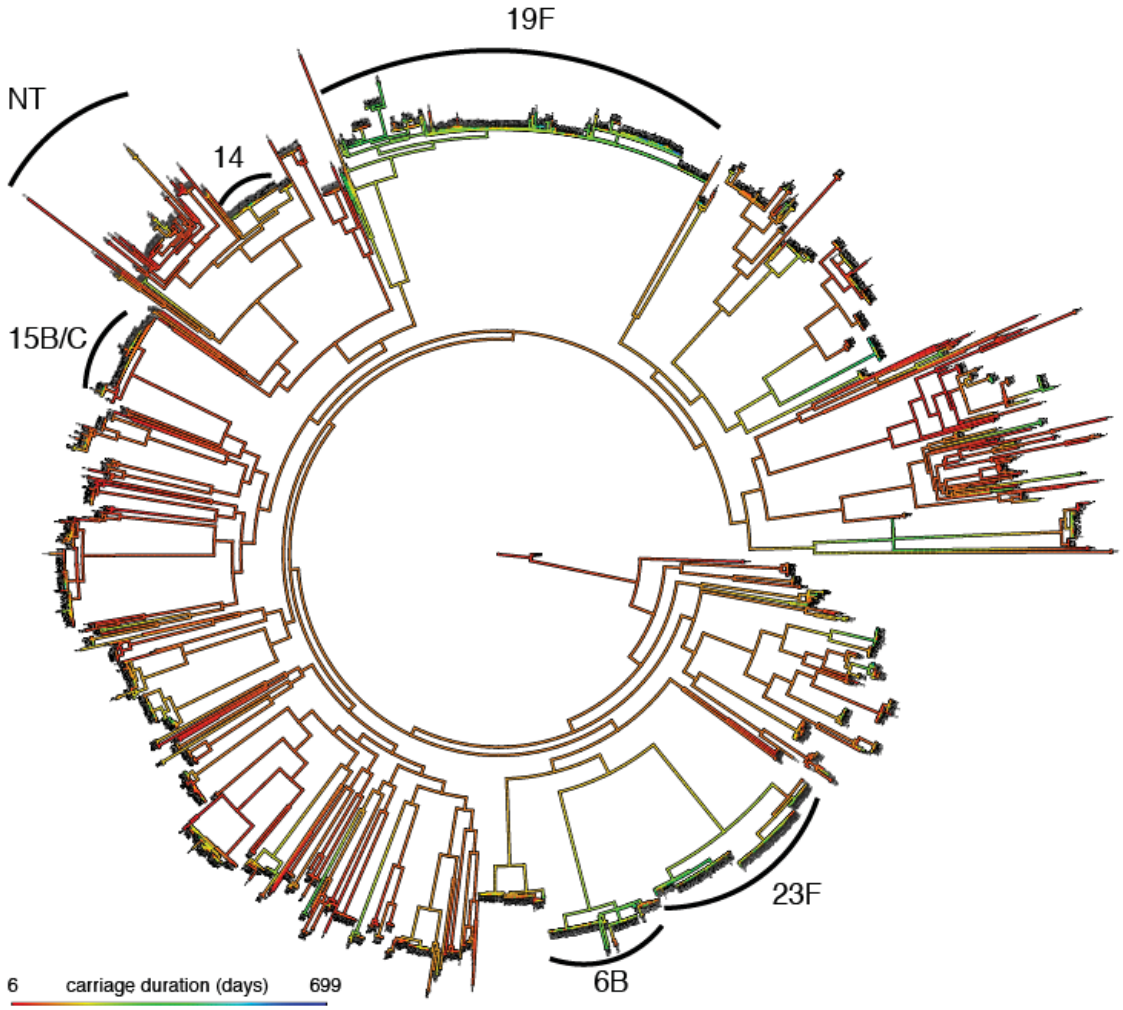


Figure 3.4: Mapping of carriage duration onto phylogeny. Using the carriage duration as a continuous trait, the ancestral state at every node of the rooted phylogeny was reconstructed. Red branches are carriage for a short time, blue for a long time. Clusters identified in previous analysis have been labelled.

Locus effects may be associated with a change in carriage duration due to convergent evolution (which may occur through recombination between lineages). In such regions, the causal loci and corresponding phenotypic effects are easier to identify (Power et al., 2016). The fixed effect model of SEER (chapter 2) or LMMs can be used to find these variants which are associated with a bacterial phenotype independent of lineage; discovery of homoplasic and polygenic variation associated with the phenotype across the entire tree is well powered (Earle et al., 2016).

While the high heritability suggests many pathogen variants do affect carriage duration, it does not give information on how many of these will be locus or lineage effects. I mapped carriage duration onto the phylogeny, reconstructing the ancestral state at each node. Consistent with the high heritability of carriage duration I found that carriage length was clearly stratified by lineage (fig. 3.4): I calculated Pagel's lambda as 0.56 ($p < 10^{-10}$) (Pagel, 1997). $\lambda = 0$ corresponds to a star-like tree, whereas $\lambda = 1$ is Brownian-motion evolution of the trait. I also modelled the evolution of carriage duration along the tree using

an Ornstein-Uhlenbeck model as implemented in `patherit`, and compared the likelihood of the full fit to that with no genetic effect on the trait ($\sigma^2_G = 0$) using a LRT with one degree of freedom. This also suggested that lineage genetics were significantly correlated with the trait (LRT = 952; $p < 10^{-10}$)

3.4.1 Serotype and drug resistance explain part of the narrow-sense heritability

I first tested for the association of serotype with carriage duration using lasso regression and with a LMM. Serotype is correlated with sequence type (Croucher, Harris, Fraser et al., 2011) and has previously been associated with differences in carriage duration (Abdullahi et al., 2012a; P. Turner et al., 2012). I also included resistance to six antibiotics, the causal element to some of which are known to be associated with specific lineages (section 2.6.2) . These are therefore possible lineage effects which would be unlikely to be found associated under a model which adjusts for population structure.

Not all serotypes and resistances may have an effect on carriage duration, or there may not be enough carriage episodes observed to reach significance. As including extra predictors in a linear regression always increases the variance explained, I first performed variable selection using lasso regression (Efron et al., 2004) to obtain a more reliable estimate of the amount of variation explained. Where a resistance and serotype are correlated and both associated with a change in carriage duration, this will produce a robust selection of the predictors (Hebiri & Lederer, 2012).

I encoded all 56 observed serotypes (including NT) and phenotypic resistance to the six antibiotics (chloramphenicol, β-lactams, clindamycin, erythromycin, trimethoprim and tetracycline) as dummy variables. I used serotype 6A/C as the reference level, as this had a mean carriage duration close to the grand mean in previous analysis. Orthogonal polynomial coding was used for the latter four antibiotics, where resistance could be intermediate or full. I then regressed this design matrix $\mathbf{X}$ against the transformed carriage duration $\hat{\mathbf{y}}$. I removed three observations with low carriage lengths due to a delayed initial swab, and seven observations with leverages of one (fig. A.3).

I performed variable selection using lasso regression (Efron et al., 2004), implemented in the R package `glmnet` (Friedman et al., 2010). I used leave-one-out cross-validation to choose a value for the ℓ_1 penalty; the value one standard error above the minimum cross-validated error (Tibshirani et al., 2001) was selected ($\lambda = 0.033$; fig. A.5). The 20 predictors with non-zero coefficients in the model at this value of λ were used in a linear regression to calculate the multiple R^2, which corresponds to the proportion of variance explained by these predictors.

I also estimated the variance components from serotype and resistance using genomic partitioning (J. Yang, Manolio et al., 2011), as implemented in `LDAK`. This estimates h^2

from a subset of the overall genetic loci, allowing for the heritability associated with a particular region of the genome to be tested. I used SNPs in the capsule locus to calculate a kinship matrix approximating the contribution from serotype variation. For antibiotic resistance I used SNPs in the *pbp* genes, *dyr* gene and ICE transposon to calculate a kinship matrix. Restricted maximum likelihood was used to estimate the variance explained by each of these components.

The selected predictors and their effect on carriage duration are shown in table 3.2. The total variance explained by these lineage factors was 0.19, 0.178 for serotype alone and 0.092 for resistance alone. When I used genomic partitioning of variance components these were instead estimated to be 0.253, 0.135 and 0.113, respectively. I applied the covariance test (Lockhart et al., 2014) to determine which lineage effects were significantly associated with carriage duration and found that 19F, erythromycin resistance, 23F, 6B caused significant ($\alpha < 0.05$) increase in carriage duration and being non-typable caused a significant decrease.

Factor	Effect on carriage duration (days)
Mean (intercept)	**59.5**
Erythromycin resistance	**+7.5**
Tetracycline resistance	+3.0
Trimethoprim resistance	+2.9
Clindamycin resistance	+1.8
Penicillin intermediate resistance	+1.3
Serotype 19F	**+46.9**
Serotype 23F	**+21.0**
Serotype 6B	**+16.2**
Serotype 14	+7.2
Serotype 21	+1.6
Serotype 19B	-0.1
Serotype 18C	-1.9
Serotype 29	-4.3
Serotype 3	-4.5
Serotype 4	-7.2
Serotype 24F	-8.5
Non-typable	**-12.3**
Serotype 5	-18.6

Table 3.2: Coefficients from lasso regression model of carriage duration. The mean (intercept) corresponds to a sensitive 6A/C carriage episode, and different serotypes and resistances are perturbations about this mean. Positive effects are expected to have a greater magnitude, due to the positive skew of carriage duration. Rows in bold were significant predictors in the covariance test.

3.4.2 Independent effects of serotype and genetic background

Previous studies have used isogenic strains to look for effects of serotype of colonisation and carriage duration independent of genetic background. Resistance to killing (Weinberger et al., 2009), growth phenotype (Hathaway et al., 2012) and resistance to complement (Melin et al., 2010) have all been shown to affect carriage through serotype rather than genetic background. Conversely, some bacterial genetic variation has been shown to be able to affect colonisation independent of serotype (Nadeem Khan et al., 2014).

I therefore wished to test whether the detected effect of serotype and resistance on carriage duration was entirely mediated through their covariance with lineage, or whether they are independently associated with carriage duration. I first looked for differences in duration over three recent capsule gain/loss events; if there is an effect of serotype independent of genetic background, these would be predicted have the largest difference

between serotypes while controlling for the relatedness of isolates. Capsule switch events had been previously identified by first reconstructing of the ancestral state of the serotype at each node through maximum parsimony (Chewapreecha, Harris et al., 2014). For each node involving loss or gain of the capsule, those with at least one child being a tip were selected to find recent switches (all were capsule gain). The carriage duration of all unencapsulated children (in the phylogenetic sense) of the identified node were used as the null distribution to calculate an empirical p-value for the switched isolate. P-values were combined using Fisher's method (Rosenthal, 1978). No significant difference in duration was seen between isolates with or without capsule within the same lineage (p = 0.39; fig. 3.5).

However, as these events were limited in number, assumed genetic independence within the clade and occurred only in part of the population, I also performed the same regression as above while also including lineage (defined by discrete population clusters) as a predictor. This therefore allows serotypes which appear in different population clusters to distinguish whether lineage or serotype had a greater effect on carriage duration. The covariance test found that 19F, erythromycin resistance and being non-typable had significant effects on the model (in that order). As these terms enter the model before any lineage specific effect, this suggested these serotypes and resistances are associated with variation in carriage duration independent of background genotype

This lasso-based analysis may be vulnerable to confounding from unmeasured variables which may be associated with the explanatory variables (serotype and resistance). To fully account for the effect of the bacterial genome rather than relying on discrete clusters as covariates in the regression, I then performed regression of these lineage effects under a LMM where the relatedness between strains was instead included as a random effect. The predictors had the same order of significance, but only serotype 19F reached genome-wide significance ($p = 3.8 \times 10^{-7}$).

Together, this suggested that the main lineage effect on carriage duration is the serotype, but only some serotypes (19F) have an association independent of genetic background. I also found that erythromycin resistance may be significantly associated with an increased carriage duration. While being a relatively uncommon treatment in this setting (3% of treatments captured), I did not find that other antibiotics were associated. This may be because erythromycin resistance would be expected to cause an almost four order magnitude increase in minimum inhibitory concentration (MIC), whereas other resistance acquisitions have a much smaller effect.

3.4.3 Average carriage duration by serotype

Additionally, I calculated the mean sojourn times (average length of time children are expected to remain in the carrying state of the model with the given serotype) and mean

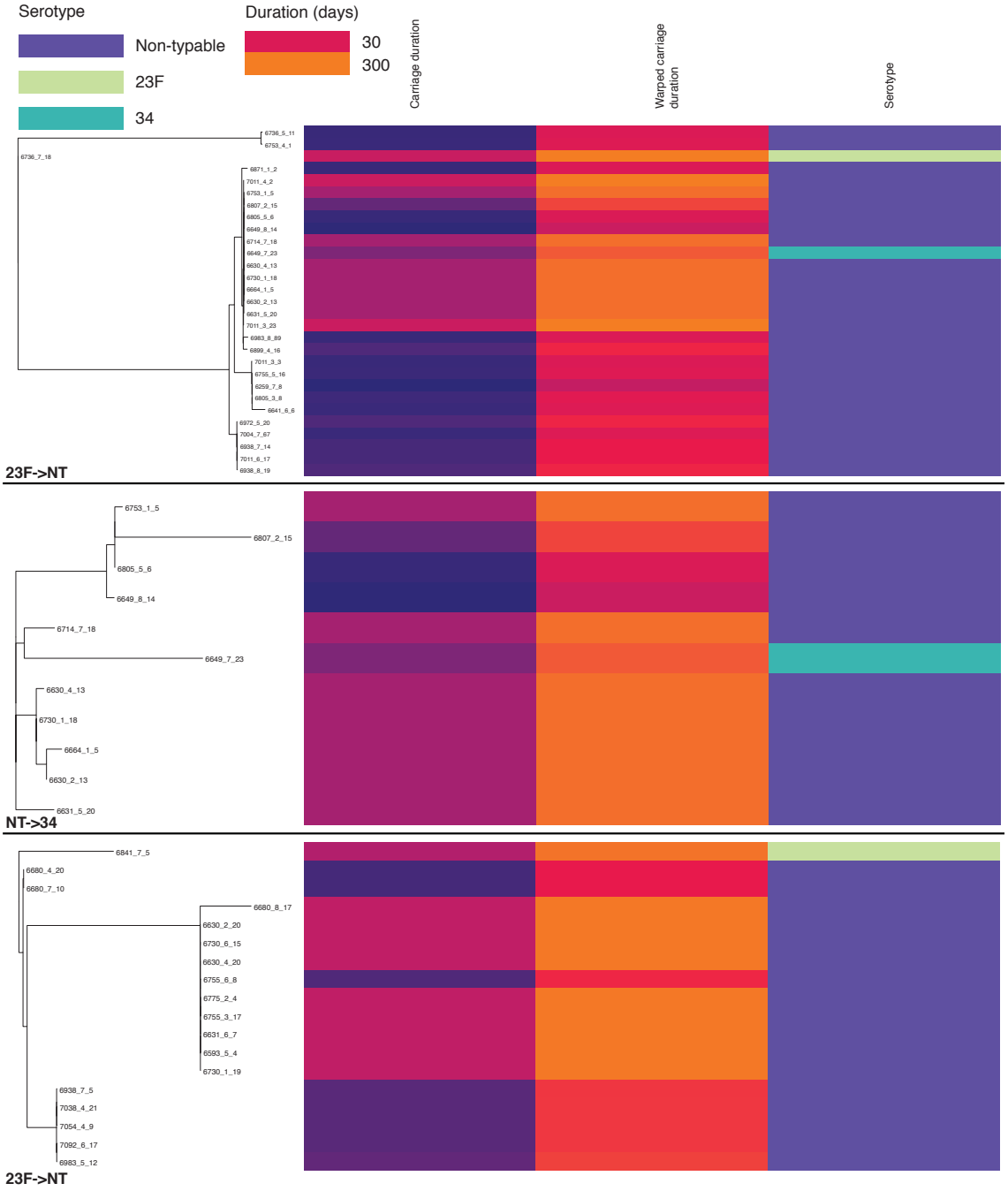


Figure 3.5: Change in carriage duration associated with capsule switching events. For each of the three events analysed the subtree containing the switch is shown on the left. For each isolate within the subtree, carriage duration (on a roughly exponential scale), warped carriage duration (on a roughly linear scale) and serotype are shown as coloured bars aligned with the tip.

number of carriage episodes from the fit to the HMM for commonly carried serotypes (table 3.3), which gave results similar to the regression performed above. These estimates are comparable to the previous analysis on a subset of these samples. The majority of carriage episodes were due to five of the seven paediatric serotypes (Shapiro & Austrian, 1994), or non-typeable isolates. The results show 19F, 23F and 14 were carried the longest, 6A/C and 6B for intermediate lengths, and NT the shortest.

Serotype	Sojourn time (days)	Expected number of infections
19F	292*	0.85
23F	112	0.83
6A/C	76.4	0.88
6B	114	0.75
14	137*	0.58
NT	40.6	2.05

Table 3.3: Mean length of carriage, and expected number of carriage episodes within the first two years of life. Only serotypes with enough data for the HMM fit to converge are shown. Starred observations have a standard error which is larger than the estimated value, indicating low confidence in the estimate.

The overall picture of the first two years of infant carriage is one containing one or two long (over 90 day) carriage episodes of a common serotype (6A/C, 6B, 14, 19F, 23F) and around two short (under a month) carriage episodes of non-typable *S. pneumoniae*. Colonisation by other serotypes seem to cause slightly shorter carriage episodes, though the relative rarity of these events naturally limits the confidence in this inference. That some serotypes are rarer and carried for shorter time periods may be evidence of competitive exclusion (Hardin, 1960; Trzciński et al., 2015), as fitter serotypes quickly replace less fit serotypes thus leading to reduced carriage duration. The calculated mean carriage duration of NT pneumococci is similar to the minimum resolution I was able to measure by the study design, which suggests carriage episodes may actually be shorter than one month. Unfortunately the only existing study with higher resolution did not check for colonisation by NT pneumococci (Abdullahi et al., 2012a).

These estimates are similar to previous longitudinal studies in different populations (P. C. Hill et al., 2010; Högberg et al., 2007; Melegaro et al., 2007), though against the Kilifi study these estimates are systematically larger. This may be due to the lower resolution swabbing we performed, or may be because the previous study was unable to resolve multiple carriage (11% of positive swabs). While the heritability estimates are specific to this population due to differences in host, vaccine deployment and transmission dynamics, the similarity of the estimates of serotype effect to those from different study populations suggests our results may be somewhat generalisable.

3.5 Additional loci identified by genome-wide association

To search for locus effects I used the linear mixed model implemented in `fast-lmm` (Lippert et al., 2011) to associate genetic elements with carriage duration, independent of overall lineage effects. I used the warped phenotype as the response, the kinship matrix (calculated from SNPs) as random effects, and variant presence, child age and previous carriage as fixed effects. For SNPs I used a Bonferroni correction with $\alpha < 0.05$ and an N of 92 487 phylogenetically independent sites to derive a genome-wide significance cutoff of $p < 5.4 \times 10^{-7}$, and a suggestive significance cutoff (Lander & Kruglyak, 1995; Stranger et al., 2011) of $p = 1.1 \times 10^{-4}$. I tested pairwise LD between the significant SNPs by calculating the R^2 between them. I removed those with $R^2 > 0.2$, assuming these represented the same underlying signal, to define the significant loci. For k-mers I counted 5 254 876 phylogenetically independent sites, giving a genome wide significance cutoff of 9.5×10^{-9}. I used `blastn` with default settings to map the significant k-mers to seven reference genomes (ATCC 700669, INV104B, OXC141, SPNA45, Taiwan19F, TIGR4 and NT_110_58), and the possible Tn*916* sequences (Croucher, Harris, Fraser et al., 2011).

The results for SNPs are shown in fig. 3.6 and table 3.4, with 14 loci reaching suggestive significance and two reaching genome-wide significance (top hit $\beta = 0.17$; $p = 2.1 \times 10^{-7}$; MAF = 1%). I also found that 424 k-mers reached genome-wide significance (top hit $\beta = 0.11$; $p = 2.1 \times 10^{-12}$; MAF = 2%), which I filtered to 321 k-mers over 20 bases long to remove low specificity sequences (fig. A.7). To determine their function, I mapped these k-mers to the coordinates of reference sequences.

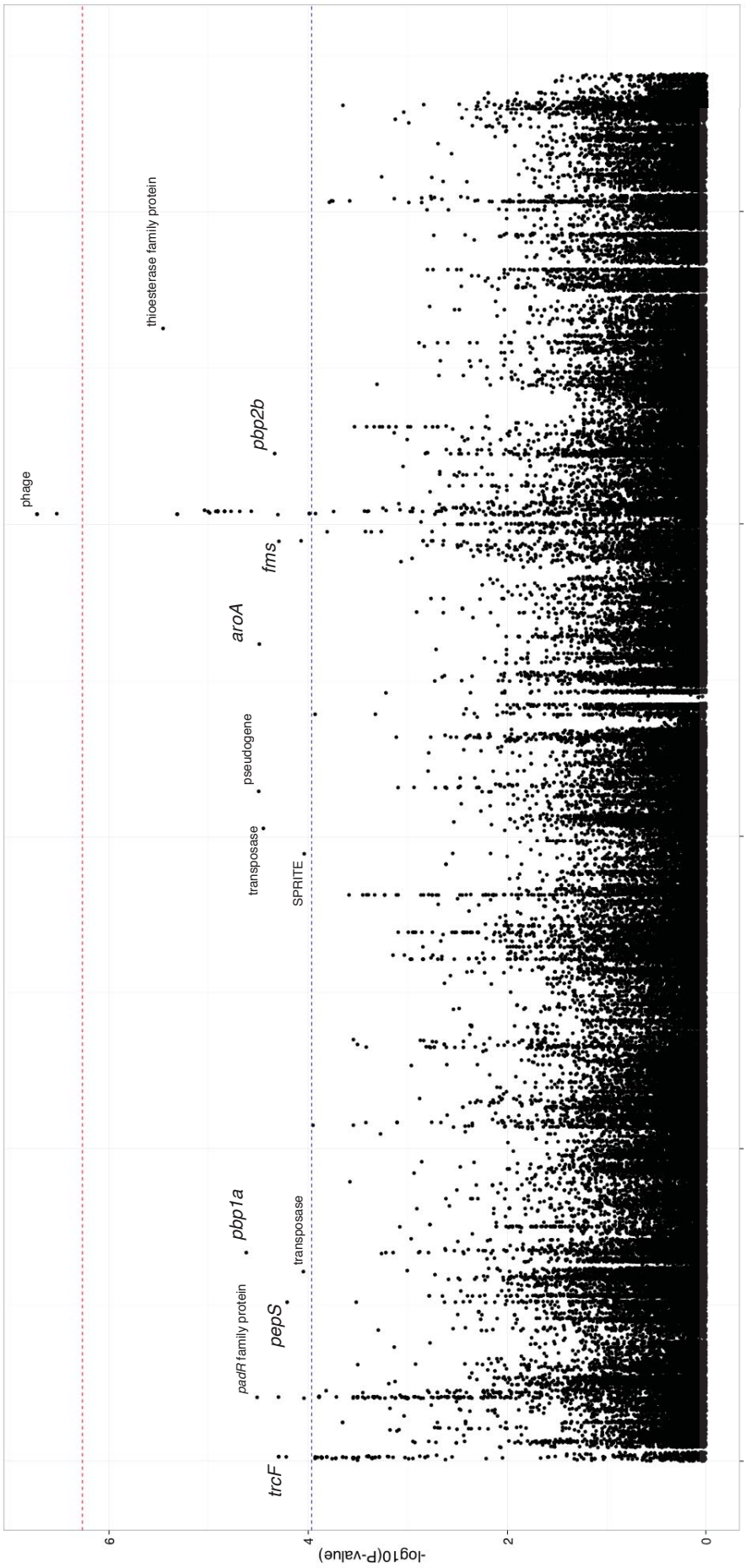


Figure 3.6: Manhattan plot of SNPs associated with carriage duration. The significance of each SNP's association with carriage duration against its position in the ATCC 700669 genome is shown. The red line denotes genome-wide significance ($\alpha < 0.05$ Bonferroni corrected with 92 487 unique tests), and the blue line suggestive significance (2.3 orders of magnitude below significant, following convention). Loci reaching suggestive significance are labelled with their nearest annotation, as in table 3.4.

Co-ordinate	Nearest annotation	Effect size	P-value	Significance level
6753	*trcF*	-0.12	6.2×10^{-5}	Suggestive
254312	*pepS*	-0.11	6.4×10^{-5}	Suggestive
303239	IS630-Spn1 transposase	0.078	9.2×10^{-5}	Suggestive
333632	*pbp1a*	0.079	2.5×10^{-5}	Suggestive
971849	SPRITE repeat region	0.078	9.4×10^{-5}	Suggestive
1013978	IS630-Spn1 transposase	0.11	3.7×10^{-5}	Suggestive
1073185	FM211187.3435 (pseudogene)	0.086	3.3×10^{-5}	Suggestive
1308604	*aroA*	-0.27	3.8×10^{-5}	Suggestive
1472933	Upstream of *fms*	-0.23	5.3×10^{-5}	Suggestive
1473700	putative glutathione S-transferase	-0.16	8.8×10^{-5}	Suggestive
1515497	hypothetical phage protein	-0.099	5.2×10^{-5}	Suggestive
1516293	putative phage Holliday junction resolvase	-0.10	5.1×10^{-6}	Suggestive
1516350	putative phage Holliday junction resolvase	-0.12	2.1×10^{-7}	Genome-wide significant
1517063	phage protein	-0.11	3.3×10^{-7}	Genome-wide significant
1613197	*pbp2b*	-0.21	4.8×10^{-5}	Suggestive
1813192	thioesterase superfamily protein	-0.12	4.8×10^{-6}	Suggestive

Table 3.4: SNP locus effects at genome-wide and suggestive significance. Co-ordinates are with respect to the ATCC 700669 reference genome, and are for the lead SNP in each locus after LD-pruning. Effect sizes are for the warped phenotype.

3.5.1 Prophage sequences associated with reduced carriage duration

The only genome-wide significant SNP hits are synonymous changes in the replication module of the prophage in the ATCC 700669 genome (MAF = 1%), a highly variable component of the pneumococcal genome (Croucher, Coupland et al., 2014) (fig. 3.7). The LD structure suggested there were two separate significant signals found in this region. I therefore performed another GWAS conditioning on the top hit (using it as a fixed effect in the regression at other sites, and removing it from kinship estimation) to test if there was a second independent signal, but found that the second hit in this region was no longer significant (position 1526024; $p = 2.2 \times 10^{-4}$). The current data is therefore consistent

with only a single significant hit to prophage.

The most significant k-mer hits were also located in phage sequence (MAF 2%) and were associated with a reduced duration of carriage. As these mobile genetic elements are less weakly population stratified than other regions of the genome, they are easier to find as locus effects. The LD in this region is less than in the rest of the genome, as prophage sequence is highly variable within *S. pneumoniae* lineages (Croucher, Coupland et al., 2014). Multiple independent phage variants may therefore affect carriage duration, which will increase their significance using a LMM. Indeed, the significant results from the LMM (top SNP $p = 2.1 \times 10^{-7}$; top k-mer $p = 2.1 \times 10^{-12}$) are not significant (top SNP $p = 5.1 \times 10^{-6}$; top k-mer $p = 5.7 \times 10^{-8}$) under a model of association using a linear regression with the first 30 principal components as fixed effects to control for population structure rather than random effects, and are strongly associated with the population structure components of the model (highest association $p = 5.2 \times 10^{-75}$ with principal component 2).

I first postulated that presence of any phage in the genome may cause a reduction in carriage duration. I identified the presence of phage by performing a `blastn` of the de novo assemblies against a reference database of phage sequence (Croucher et al., 2016). If the length of the top hit was over 5000 I defined the isolate as having phage present (fig. A.6). I then used the presence of phage as a trait under the same linear mixed model, however I found no evidence of association when correcting for population structure (p = 0.35). These results are therefore evidence that infection with a specific phage sequence is associated with a slight decrease in carriage duration. A similar result has previously been found in a genome-wide screen in *N. meningitidis*, where a specific phage sequence was found to affect the virulence and epidemiology of strains (Bille et al., 2005; Bille et al., 2008). Additionally, previous in vivo tests have shown phage elements to cause a fitness decrease of *S. pneumoniae* during carriage (DeBardeleben et al., 2014).

The genetic polymorphisms in the prophage associated with changes in carriage duration, found in 2% of viral sequences, were within coding sequences inside the phage replication module (Romero et al., 2009). It is unlikely the specific variants of these proteins cause a significant difference in phenotype, because they are only highly expressed after the prophage is activated, and cell lysis usually happens shortly afterwards. One explanation for these results is that some subpopulations of prophage do not cause a significant decrease in their host bacterium's carriage duration, which could be due to beneficial 'cargo' genes. However previous surveys of pneumococcal prophage have found little evidence of these elements carrying such sequences (Croucher, Coupland et al., 2014; Romero et al., 2009). One phage protein that has been found to alter the bacterial phenotype is PblB, a phage structural protein that can also mediate bacterial adhesion to platelets (Loeffler & Fischetti, 2006). However, *pblB* is within the morphology module (Romero et al., 2009) and as an adhesin might if anything be expected to increase carriage

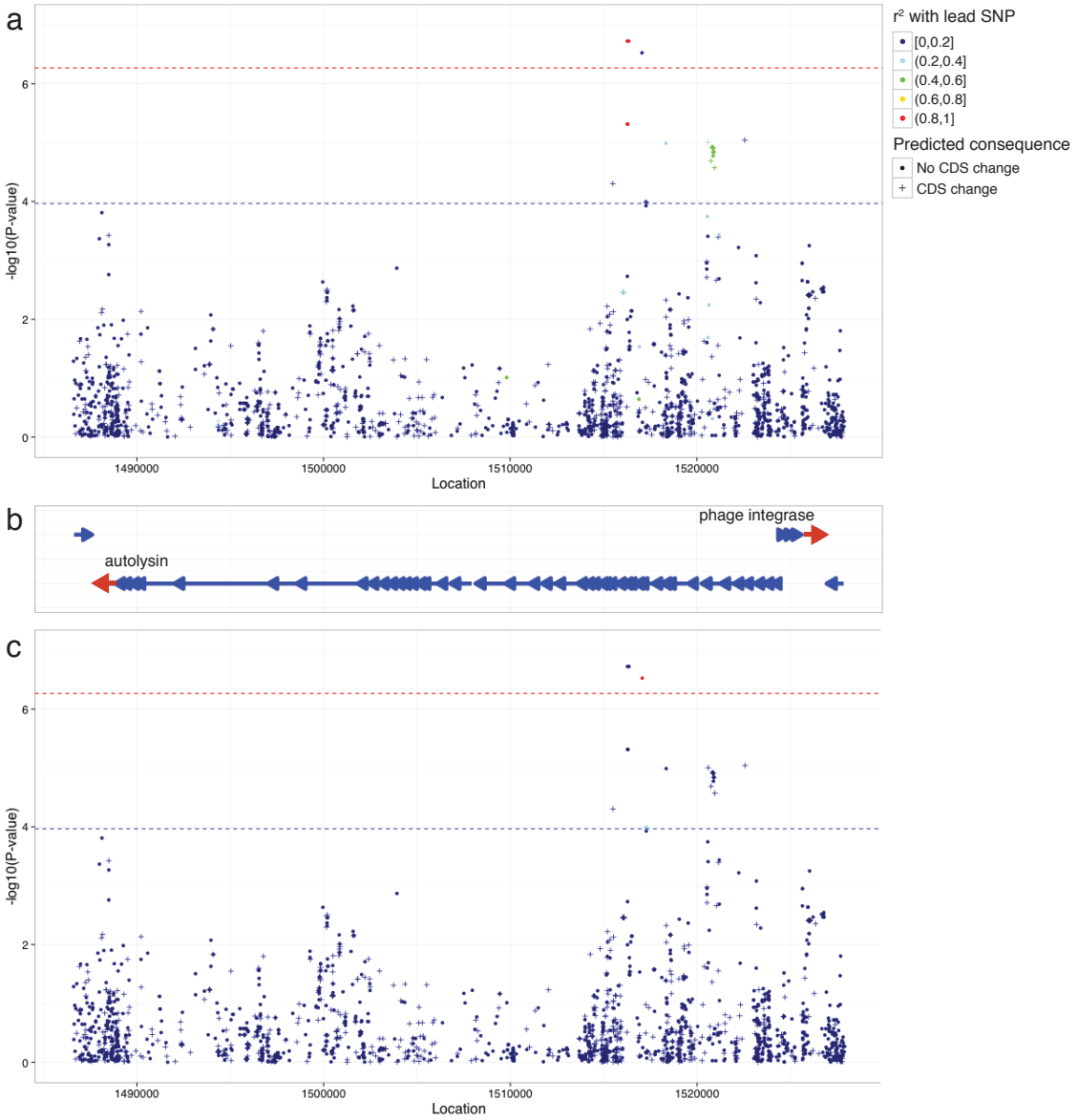


Figure 3.7: Manhattan plots of phage-associated SNPs associated with carriage duration. As in fig. 3.6, but enlarging the phage region found to be significant. SNPs are coloured by their LD with the lead SNP (the highest P-value in the region plotted), and are crosses if they are predicted to cause a change in coding sequence. Panel **a)** shows LD in relation to the lead SNP at position 1516350. Panel **b)** plots genes in the region, with the start and end of the phage genes labelled. Panel **c)** shows LD in relation to the second SNP signal at position 1517063.

duration, and was not detected in the association analysis. Hence the detected association is unlikely to represent expression of viral machinery or cargo genes in the host cell while the prophage is dormant.

Alternatively, the association with only a subset of prophage may have been a consequence of the study sampling design. Using a monthly swabbing approach, it was only possible to infer changes in the carriage duration of genotypes that colonise hosts for relative long periods. Therefore any prophage variant that enhances a virus' ability to infect long carriage duration pneumococci may have an increased association with the variation in the observed phenotype. As phage commonly exhibit high levels of strain specificity (Duplessis & Moineau, 2001), this is a plausible mechanism, although the role of the replication module in such host preference is unclear.

An additional mechanism by which prophage can affect host phenotype is by inserting into, and thereby disrupting, functional genes. Pneumococcal prophage frequently insert into *comYC*, thereby preventing the host cell undergoing transformation (Croucher, Harris, Fraser et al., 2011; Croucher, Hanage et al., 2014). Using previous categorisation of the *comYC* gene in this collection into intact versus interrupted or missing (Croucher et al., 2016), I found that having an intact *comYC* gene (23% of isolates) was significantly associated with an increased carriage duration ($\beta = 0.29$; $p = 1.4 \times 10^{-44}$). The effect size is similar to the associated phage k-mers, but has at a higher allele frequency (hence the increased significance of the result). An interpretation consistent with these findings would be that the effect of phage k-mers is actually through interrupting *comYC*. The k-mers themselves were spread out to lower frequencies due to their sequence variability, and none of the references I used allowed mapping to find the *comYC* interruption directly.

3.5.2 Other loci associated with altered carriage duration

Signals at the suggestive level included *pbp1a* and *pbp2b*, which suggested as above that penicillin resistance may slightly increase carriage duration, but there are not enough samples in this analysis to confirm or refute this. Other signals near genes at a suggestive level included SNPs in *trcF* (transcription coupled DNA repair), *padR* (repressor of phenolic acid stress response), *pepS* (aminopeptidase), *aroA* (aromatic amino acid synthesis), *fms* (peptide deformylase) and a thioesterase superfamily protein. K-mers from erythromycin resistance genes (*ermB*, *mel*, *mef*) were expected to reach significance from the above analysis, but did not: however I showed in section 2.6.2 that the power to detect these elements in a larger sample set taken from the same population is limited due to the multiple resistance mechanisms and stratification of resistance with lineage.

The test statistic from `fast-lmm` roughly followed the null-hypothesis, with the exception of the significant phage k-mers (fig. A.8). However there is limited power to detect effects associated with both the lineage and phenotype. This effect has been previously

noted, and while LMMs have improved power for detecting locus specific effects they lose power when detecting associated variants which segregate with background genotype (Earle et al., 2016). To search for candidate regions which may be independently associated with both a lineage and increased carriage duration, I ran an association test with SEER (chapter 2) using a set number of fixed effects as the population structure correction. In this case I used the patristic distances from the phylogenetic tree as the kinship matrix, which I then projected into 30 dimensions using metric multidimensional scaling to obtain covariates. This may be expected to have higher power than an LMM for true associated variants on ancestral branches as some association with population structure is permitted, but will also increase the number of false positives (variants co-occurring on these branches which do not directly affect the carriage duration themselves). To reduce the number of false positives I used a strict threshold of $p < 10^{-14}$. I separately tested SNPs for their association with those principal components which were themselves significantly associated with carriage duration, and therefore may be driving the lineage associations using the model of Earle et al. (2016).

The most highly associated SNPs were in all three *pbp* regions associated with β-lactam resistance, the capsule locus, *recA* (DNA repair and homologous recombination), *bgaA* (beta-galactosidase), *phoH*-like protein (phosphate starvation-inducible protein), *ftsZ* (cell division protein) and *groEL* (chaperonin). As 19F, the serotype most associated with carriage duration, is predominantly the β-lactam resistant PMEN14 lineage the *pbp* association may be driven through strong LD between with this serotype. Figure A.9 shows the analysis of SNPs which may be driving significant lineage associations – this also suggested *dnaB* (DNA replication) may be associated with altered carriage duration. Associated k-mers were also found in *phtD* (host cell surface adhesion), *mraY* (cell wall biosynthesis), *tlyA* (rRNA methylase), *zinT* (zinc recruitment), *adcA* (zinc recruitment) and *recJ* (DNA repair). Additionally I found k-mers in the bacteriocin *blpZ* and immunity protein *pncM* (Bogaardt et al., 2015) to be associated with variability in carriage duration. This could be evidence that intra-strain competition occurs within host via this mechanism, consistent with previous in vitro mouse models (Dawid et al., 2007).

It is not possible to determine whether variation in these genes is associated with a change in carriage duration or if the variation is present in longer carried, generally more prevalent lineages. For example, β-lactam resistance may appear associated as the long carried lineages 19F (dominated by PMEN14, as noted above) and 23F are more frequently resistant, or it may genuinely provide an advantage in the nasopharynx that extends carriage duration independent of other factors. Future studies of carriage duration, or further experimental evidence will be needed to determine which is the case for these regions.

Antigenic variation in known regions (of *pspA*, *pspC*, *zmpA* or *zmpB*) may be expected to cause a change in carriage duration (Lipsitch & O'Hagan, 2007), however I did not find

any of these to be associated with a change in carriage duration. This was likely due to stratification of variation in these regions with lineage, but may also be caused by a larger diversity of k-mers in the region reducing power to detect an association.

3.6 Child age independently affects variance in carriage duration

Finally, I wished to determine the importance of two environmental factors which are known to contribute to variance in this phenotype: child age and whether the carriage episode is the first the child has been exposed to (Abdullahi et al., 2012a, 2012b; P. Turner et al., 2012). These have been applied throughout the analysis as covariates, both in the estimation of carriage episodes and in associating genetic variation with change in carriage duration.

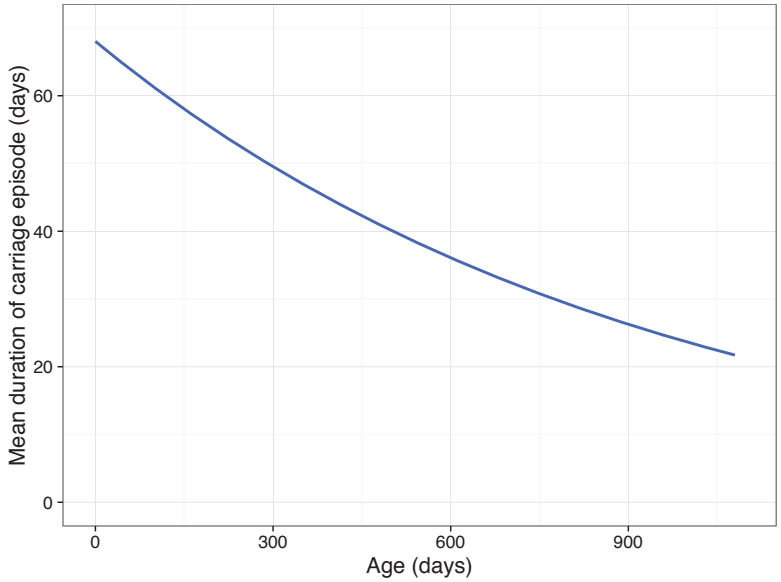


Figure 3.8: Predicted mean carriage duration as a function of child age. Fit is an exponential decay over the first two years of life, using the decay rate inferred from a linear regression of log(carriage duration).

I applied linear regression to these factors while using the first 30 PCs to correct for the effect of the bacterial genome, which showed they were both significantly associated with carriage duration as expected (age $p = 3.9 \times 10^{-7}$; previous carriage $p = 2.5 \times 10^{-8}$). Using the linear mixed model to control for bacterial genotype both factors were again significant (LRT = 26.4; $p = 1.8 \times 10^{-6}$). Together, they explained 0.046 of variation in carriage duration. As found previously, increasing child age contributes to a decrease in the duration of carriage episodes. From a mean of 68 days long, I calculated a drop of 19 days after a year, and 32 days after two years. Extrapolating, this causes carriage episodes longer than two days to cease by age 11 (fig. 3.8). Previous carriage of any serotype was

estimated to cause an increase in the duration of future carriage episodes, though previous studies have found no overall effect (Weinberger et al., 2008). It has previously been shown that prior exposure to non-typables in this cohort make colonisation by another non-typable occur later, and for a shorter time (P. Turner et al., 2012). The positive effect observed in this analysis is therefore likely to be an artefact due to subsequent carriage episodes being more likely to be due to typable pneumococci.

Additional environmental factors that explain some of the remainder of the variance may include the variation of the host immune response and interaction with other infections or co-colonisation. In particular, co-infection with influenza A was not recorded but is known to affect population dynamics within the nasopharynx (Kono et al., 2016). Fundamentally, imprecise inference of the carriage duration will limit the ability to fully explain its variance here.

3.7 Conclusions

Other than serotype, the genetic determinants of pneumococcal carriage duration were previously unknown. By developing models for longitudinal swab data and combining the results with whole genome sequence data I have quantified and mapped the genetic contribution to the carriage duration of *S. pneumoniae*. I found that despite a range of other factors such as host age which are known to cause carriage duration to differ, sequence variation of the pneumococcal genome explains most of this variability (63%). Common serotypes and resistance to erythromycin caused some of this effect (19% total), as does the presence or absence of particular prophage sequence in the genome. Table 3.5 summarises the sources I found to be significantly associated with variation in carriage duration.

Source	of which is	Total variance explained	Proportion of total heritability explained
Total heritability (H^2)		0.634 (CPP)	1.00
	Common SNP heritability (h^2_{SNP})	0.438 (LMM)	0.691
	Serotype and resistance	0.190 (R^2)	0.300 (R^2)
		0.253 (LMM)	0.399 (LMM)
	Serotype only	0.178 (R^2)	0.281 (R^2)
		0.135 (LMM)	0.213 (LMM)
	Resistance only	0.092 (R^2)	0.145 (R^2)
		0.113 (LMM)	0.178 (LMM)
	Phage k-mers	0.067 (LMM)	0.106
	Intact *comYC*	0.127 (LMM)	0.201
Measured environmental effects	Age and previous carriage	0.046 (R^2)	-

Table 3.5: Summary of variance of carriage duration explained by genetic and environmental factors. H^2 encompasses all rows, other than the measured environmental effects. For each variant component the method used to estimate it is reported: CPP - closest phylogenetic pairs; LMM - variance component using a linear mixed model with pathogen genotype as random effects; R^2 - linear regression using lasso to select predictors.

I have provided a quantitative estimate of how closely transmission pairs share their carriage duration, and show evidence for differences both between and within serotypes. The implication of phage as having a significant effect on carriage duration has interesting corollaries on pneumococcal genome diversification through frequent infection and loss of prophage, even during carriage episodes in this dataset.

Investigating a mechanism for the prophage association, I found that having an intact *comYC* gene, which is frequently interrupted by prophage causing loss of function of the competence system, was associated with increased carriage duration. While the competence system is observed to remain intact over the evolutionary history of the species, these disruptive mutations spread irreversibly through the population as competent bacteria can acquire the mutation, and non-competent bacteria can no longer reverse it through recombination (Croucher, Hanage et al., 2014). Selection must therefore maintain

the function at this locus over short timescales, and an increased carriage duration may be evidence of this. I therefore hypothesise that the associated prophage sequences may haved affected carriage duration through disruption of the competence system.

The results presented here have important implications for the modelling of pneumococcal transmission and their response to perturbation of the population by vaccine. Importantly, the analysis of heritability shows that variants other than serotype affect carriage duration, consistent with recent theoretical work (Lehtinen et al., 2017). Here I have shown that these alleles do exist in a natural population, and also identified candidates for the loci which fulfil this role. Together these studies suggest that variants exist in the pneumococcal genome which alter carriage duration, which in turn is linked to antibiotic resistance.

I was not able to fully explain the basis for heritability of carriage duration for a number of reasons. The close association of the phenotype with lineage limited our power to fine-map lineage associated variants other than capsule type which may affect carriage duration. Meta-analysis with more large studies with higher resolution may help to resolve these issues. Additional environmental factors that explain some of the remainder of the variance may include the variation of the host immune response and interaction with other infections or co-colonisation. In particular, co-infection with influenza A was not recorded but is known to affect population dynamics within the nasopharynx (Kono et al., 2016).

This is a phenotype which would have been difficult to assay by traditional methods such as in an animal model due to the cohort size needed and the length of time experiments would need to be run for. By using GWAS I have been able to quantitatively investigate a complex phenotype in a natural population. This chapter has also advanced the application of GWAS methods applied to bacteria started in chapter 2 by application to a more difficult to define phenotype, introducing heritability and genomic partitioning, and testing specifically for locus effects. I have also implicitly compared fixed effect and random effect models to control for population structure. In the next chapter I will continue using these approaches to identify pneumococcal genetic variation associated with bacterial meningitis, while developing a more thorough catalogue of variation within the pneumococcal genome.

Chapter 4

Bacterial genetics contributing to invasive pneumococcal disease

Declaration of contributions

Stephen Bentley, Diederik van de Beek and Julian Parkhill supervised this work. Diederik and Arie van der Ende's groups designed and ran the MeninGene study on which this work is based, performed sample collection and DNA extraction. Philip Kremer performed the reciprocal best hits analysis. Ana Manso performed the long-range PCR to determine *ivr* allele type directly from clinical samples. I performed all other analyses.

Publication

The following has been published as:
Lees, J. A., Kremer, P. H. C., Manso, A. S., Croucher, N. J., Ferwerda, B., Serón, M. V., Oggioni M. R., Parkhill J., Brouwer M. C., van der Ende A., van de Beek D., Bentley, S. D. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microbial Genomics*. 2017; 3. doi:10.1099/mgen.0.000103
Lees, J. A., Brouwer, M., van der Ende, A., Parkhill, J., van de Beek, D., Bentley, S. D. Within-Host Sampling of a Natural Population Shows Signs of Selection on Pde1 during Bacterial Meningitis. *Infection and Immunity*. 2017; 85(3). doi:10.1128/IAI.01061-16

4.1 Introduction

This chapter deals with the contribution of variation in the pathogen genome to bacterial meningitis. Using a GWAS framework, I first wished to test whether any variation in the pneumococcal genome is associated with susceptibility to meningitis or with a poor outcome of the infection. To study this, I used isolates collected as part of the MeninGene study (section 1.1.4). This part of the study consists of 3 089 pathogen isolates from the culture-proven cases of bacterial meningitis from the Netherlands. DNA was extracted from CSF and blood cultures and sequenced with 100bp paired end reads on the Illumina HiSeq platform. Of these sequences, 1 984 were *S. pneumoniae*. 751 carriage genomes from Dutch adults and children were also sequenced to use as controls for susceptibility analysis.

I first catalogued all forms of variation to use as the loci to test in a GWAS (section 4.3). While k-mers cover most of this variation, I also included tests of SNPs and genes due to their more straightforward interpretation. Some forms of variation such as inverting repeats, CNVs, recombinogenic antigens cannot be captured by these methods, so I developed new techniques to call variants at these loci. While this covered all forms of common variation detectable by short reads in the pneumococcal genome, rare variants may also play a role in disease pathogenesis. I annotated the predicted effect of rare coding variants to choose which to use in burden tests.

Using the SNP variation to tag other forms of variation in the genome, I was able to estimate the heritability of each of these traits (the proportion of variance in the phenotype is explained by variation within the genome). Finding evidence for pneumococcal genetic variants contributing to invasiveness other than serotype, I then used the methods presented in chapters 2 and 3 to test whether any of the specific variants that I called were associated with susceptibility to or severity of meningitis.

Section 4.5 concerns pathogen variation that occurs over the course of a single infection. Croucher, Mitchell et al. (2013) have previously shown that in a single patient bacteria appeared to adapt to the distinct conditions of blood and CSF. These are very different niches from that of nasopharyngeal carriage where this variation is well documented (Cremers et al., 2014), not least because the bacteria are exposed to different immune pressures (Habets et al., 2012) and have less time over which to accumulate mutations.

It is possible that bacteria inhabiting the nasopharynx are already well adapted for CSF invasion. However, genetic variants that enable invasion of the CSF are not expected to be under positive selection, since invasion is an evolutionary dead end for the bacterium. Studies of carriage alone will therefore be unable to detect selection during invasion. Current knowledge on within-host variation during invasive disease is mostly focused at the serotype and MLST level, and lacks the resolution and sample size to be able to address this question (Brueggemann et al., 2003; del Amo et al., 2015; D. A. Robinson et al., 2001).

Though the only whole genome based study suggests there is no difference between blood and CSF populations (at the gene level) in *S. pneumoniae* (Kulohoma et al., 2015), larger sample sizes are needed to better answer this question.

I therefore wished to expand the analysis of Croucher, Mitchell et al. (2013) by including more cases of disease, and used 938 pairs of genomes from the blood and the CSF of the same patient, and 54 pairs from the nasopharynx and CSF of the same patient sequenced as part of the MeninGene study described above. This sample set included both *N. meningitidis* isolates and *S. pneumoniae* isolates, each of which was analysed separately. As isolate pairs are matched they are closely related; the issue of population structure affecting bacterial GWAS is no longer a problem. Variants between pairs can be grouped by functional effect and tested for association with a niche straightforwardly.

4.2 Quality control and processing

In this section I discuss initial QC of isolates in the collection, and evaluations of both assembly and variant calling software to be used throughout the chapter.

Using a single *S. pneumoniae* isolate, I compared the quality of three assembly methods that have previously been shown to perform well on bacterial genomes (Magoc et al., 2013): Velvet (Zerbino & Birney, 2008), SPAdes (Bankevich et al., 2012) and SOAPdenovo2 & MaSuRCA (Zimin et al., 2013). Statistics from this comparison are shown in table 4.1. I decided that the SPAdes pipeline provided good quality assemblies while being easy to run, so assembled all isolates in the collection with v3.5 of the software using default settings. Additionally I ran velvet on all samples, which when k-mer length is optimised and scaffolds are improved, gave similar results to SPAdes. I corrected the resulting velvet assemblies with SSpace and GapFiller (Page et al., 2016). The assembly result used for each purpose will be stated throughout the rest of the thesis.

	Velvet	SPAdes	SOAPdenovo2 & MaSuRCA
# contigs	48	**7**	**7**
Total length	2 096 048	**2 205 585**	2 139 022
N50	77 648	429 779	**481 453**
# genes	2 073	**2 208**	2 166
CPU time	6 h	7.2 h	5.5 h
Maximum memory	3.7 GB	7.0 GB	4.3 GB
Disk space	0.1 GB	0.6 GB	4.2 GB

Table 4.1: Assembly and annotation of *S. pneumoniae* isolate 11822_8_30. N50 is the median contig length. For each performance metric the best scoring method is in bold.

I then analysed the quality of the SPAdes assemblies using quast (Gurevich et al., 2013) and kraken (Wood & Salzberg, 2014). I performed this analysis at the sample level, rather than at the contig level. As the primary aim is a GWAS I desired complete and comparable assemblies, so the number of included samples at each variant is the same. I found two assemblies which were predominantly another species, and discarded them. I also discarded five sequence runs with low yield, 17 with total lengths over 2.5Mb, two with total lengths under 1.8Mb and one with a GC content of 31.4%. This left 1 144 CSF isolates and 674 pairs of blood and CSF isolates for downstream analysis. For the carriage samples I removed 29 isolates contaminated with another species (determined by kraken, and the position on a preliminary core gene alignment phylogeny), and 8 isolates which showed evidence of being mixed samples (number of heterozygous SNPs in preliminary mapping was greater than two standard deviations above the mean). This left 693 carriage isolates for downstream analysis.

To compare variant calling methods I produced a set of true variant calls for 30 samples. I did this by simulating evolution of *S. pneumoniae* genomes along the branch of the tree between *S. pneumoniae* R6 (Hoskins et al., 2001) and the common ancestor with *S. mitis* B6 (Denapaite et al., 2010). The rates in the GTR matrix and insertion/deletion frequency distributions were estimated as in section 2.3.1. I created an average of 10 000 mutations with these rates, and Illumina paired end read data at 200x coverage simulated using pIRS (Hu et al., 2012).

Method	True positives	False negatives	False positives
bcftools	24922	900	244
freebayes	22253	3569	1465
GATK	25024	798	191

Table 4.2: Performance of variant calling algorithms on simulated data. True positives are SNPs or INDELs correctly called; false negatives are variant sites which were missed by the caller; false positives are sites without variation but called as a variant.

I mapped the reads with bwa-mem (H. Li, 2013), followed by samtools fixmate, sort and markdup. I then called variants using bcftools, freebayes (Garrison & Marth, 2012) and GATK (Van der Auwera et al., 2002). The results are shown in table 4.2. freebayes performed poorly due to its use on multiple nucleotide polymorphism (MNP)s, which were difficult to compare to the simulations. GATK performed the best on all measures, and in particular achieved much better power at calling indels. I used it for calling SNPs and indels throughout, unless otherwise stated.

4.3 Catalogue of all pneumococcal variation

In this section I detail how I catalogued population level variation in the pneumococcal genome. These variants are then used throughout the rest of this chapter as the predictor variable in GWAS with various phenotypes of interest, analysis of within-host variation and in chapter 5 as the phenotype in a genome to genome analysis. As discussed in section 2.2, variation in bacterial genomes is not well represented by short changes compared to a linear reference due to extensive variation of the accessory genome (Donati et al., 2010; McInerney et al., 2017), mosaic alleles created by recombination (Hanage et al., 2009), structural variation (Croucher, Coupland et al., 2014; Manso et al., 2014) and copy number variation (Howden et al., 2015). I used different techniques to determine the variation present in each sample from each of these sources to ensure maximum discovery power of the GWAS performed.

While short variants (i.e. SNPs and small indels) with respect to a single linear reference only partially covers the variation present in the pneumococcal population, it is still a useful dataset to produce. A genome alignment produced this way can be used to generate the phylogenetic relationship between all samples from the population and create discrete related clusters. Both of these are useful for QC, heritability analysis and evaluating population structure. Additionally, the effect of these variants on protein function can be straightforwardly predicted, making conclusions drawn from them more easily interpreted and also of use in indirect tests of association section 4.4.2.

I produced a whole genome alignment in two ways. Firstly I mapped reads to the ATCC 700669 reference using bwa mem with default settings

```
bwa mem reference.fa forward_reads.fastq reverse_reads.
    fastq | samtools fixmate -O bam - > output.bam
```

and finally marked duplicate reads in these binary sequence alignment/map (BAM) files using Picard. I then called variants from each of these BAM files separately using samtools mpileup and bcftools call, and as a population using GATK HaplotypeCaller. I then applied hard quality filters to each of these call sets to create initial calls. To select variants based on a correctly scaled sensitivity and specificity I used GATK VariantRecalibrator to scale the variant quality scores. This tool requires known true positive calls as a prior – I used the intersection of hard filtered variants from GATK and bcftools with 90% confidence (Q10), and filtered variants from the Maela and Massachusetts studies with 68% confidence (Q5) as recommended. After recalibration, I applied 99.9% power as the cut-off for variants to maximise sensitivity at this stage. Finally, I annotated the predicted consequence of all passing variants with variant effect predictor (VEP) (McLaren et al., 2010).

I also produced a core-genome alignment using roary (Page et al., 2015) with a 95% blast ID cut-off. Roary efficiently performs all by all alignment using every annotated

protein in the dataset. Those matches with over 95% ID are assumed to be orthologs and are clustered and undergo multiple sequence alignment. Using a single cut-off will mean that some genes with orthologous function but without sequence homology (for example different alleles of a gene) will not be clustered together, and that some genes without orthologous function but with sequence homology will be incorrectly clustered. We chose the cut-off of 95% based on having the best balanced accuracy of these two error classes when using reciprocal best BLAST hits to define true orthologs (Ward & Moreno-Hagelsieb, 2014). As well as core genes (present in at least 99% of samples) roary also clusters accessory genes into COGs, which I later used as a variant in association. In this case the annotated function helped determine whether the cluster is showing presence or absence of gene or groups of different alleles of a gene that is being tested for association.

I counted k-mers using fsm-lite (section 2.2), which required 75Gb RAM and 14hrs CPU time to count all informative k-mers with a minor allele count (MAC) of ten or more. In this sample set there were 11.7M informative k-mers with 2.6M unique patterns. I called CNVs from the BAM files produced above using cn.mops (Klambauer et al., 2012) which fitted the coverage of mapped reads in 1kb windows with a mixture of Poisson distributions, and determined the most likely integer coverage value for each sample in each window. I extracted the inferred copy number from those windows which had support for a CNV from more than one sample.

4.3.1 Allelic variation of three pneumococcal antigens

I wished to determine whether sequence variation of pneumococcal antigens is associated with virulence and disease outcome. As well as being plausible GWAS hits, these antigens vary rapidly (Croucher, Vernikos et al., 2011), meaning sequence variation is not population stratified, which increases discovery power. Conversely, while the k-mer approach (section 2.2) either directly assays or indirectly tags most variation in the population, variation of these antigens may not be captured by this method. For example, *pspC* can be difficult to assemble due to repeats and copy number variation (Iannelli et al., 2002), and therefore k-mers from the gene sequence will not appear in the assembly, and not be counted or tested. In *pspA* and *zmpA*, mapping of k-mers may not be specific to the allele sequence due to sequence homology with orthologous and paralogous genes (Hollingshead et al., 2000; Bek-Thomsen et al., 2012).

Here I consider *pspC*/*cbpA*, *pspA* and *zmpA*, which have all been shown to have interactions with the host immune system (Croucher et al., 2017), but have variability that may not be assessed by the methods discussed above. I needed to develop a way first to classify possible alleles, then determine the allele of each sample from short read sequence data. For the latter issue, de novo assembly (followed by a BLAST with a set of reference alleles) is unreliable for completely reconstructing the gene sequences, but

usually contained some information about the allele present. Alternatively mapping the sequence reads to a set of reference alleles is less affected by repeat sequences and may be more accurately used to find the allele of genes (Inouye et al., 2014), but determining the closest match is non-trivial. I decided to use a method which combines summary statistics from both of these techniques to determine the allele type. This has previously shown to be advantageous for antibiotic resistance typing from genomic data; Hunt et al. (2017) designed a method using combination of assembly and mapping which had improved type I and type II error rate over either technique alone.

I defer discussion of the variability and construction of a reference panel specific to each of these alleles until the sections below, and first discuss the typing method I applied to determine the allele of all three antigens given such a reference panel. I first generated statistics from the assemblies of all samples by running blastp between the annotated genes in both the velvet and SPAdes assemblies and the reference panel. From this, I extracted the % ID, number of mismatches, number of gaps, E-value and bitscore between the two assemblies of sample and every possible reference. For mapping I used srst2 (Inouye et al., 2014) in a mode which maps reads to all reference sequences, and reports information about coverage over every possible allele. I used the coverage, number of SNP mismatches, number of indel mismatches and number of truncated bases.

This led to a data frame with 16 predictors for every reference sequence, per sample (for example *pspC* had 48 references, so there were 768 predictors). When a match was not reported by blastp or srst2 I filled in value with the minimum reported value of the predictor (or maximum for the number of mismatch fields), and removed predictors without variation.

To produce labelled training data I performed the same process on the reference panel itself, for each sequence using blastp against all the reference sequences and srst2 with simulated reads (these were error-free 100bp reads with 200x coverage and 350bp insert size with 80bp standard deviation (s.d.)). In all cases, on the test data simple variance analysis showed these statistics could be used to predict classification of alleles successfully (fig. A.13). I fitted a classifier to this training data (see section 4.3.1 for details), then finally used the trained model to predict the allele for all samples. The results are shown in fig. 4.1. As expected, all the antigens show some, but not total, concordance with background genotype. I used the above process for typing all antigens; I now discuss the specifics of constructing the reference panel for each antigen.

pspC/*cbpA* allele

The *pspC* gene, also known as *cbpA*, *hic*, *spsA* or *pbcA*, is paralogous to *pspA* and is known to have a number of immunogenic functions. These include binding host proteins C3, CFH and IgA, all of which are involved in the immune response to pneumococcal colonisation

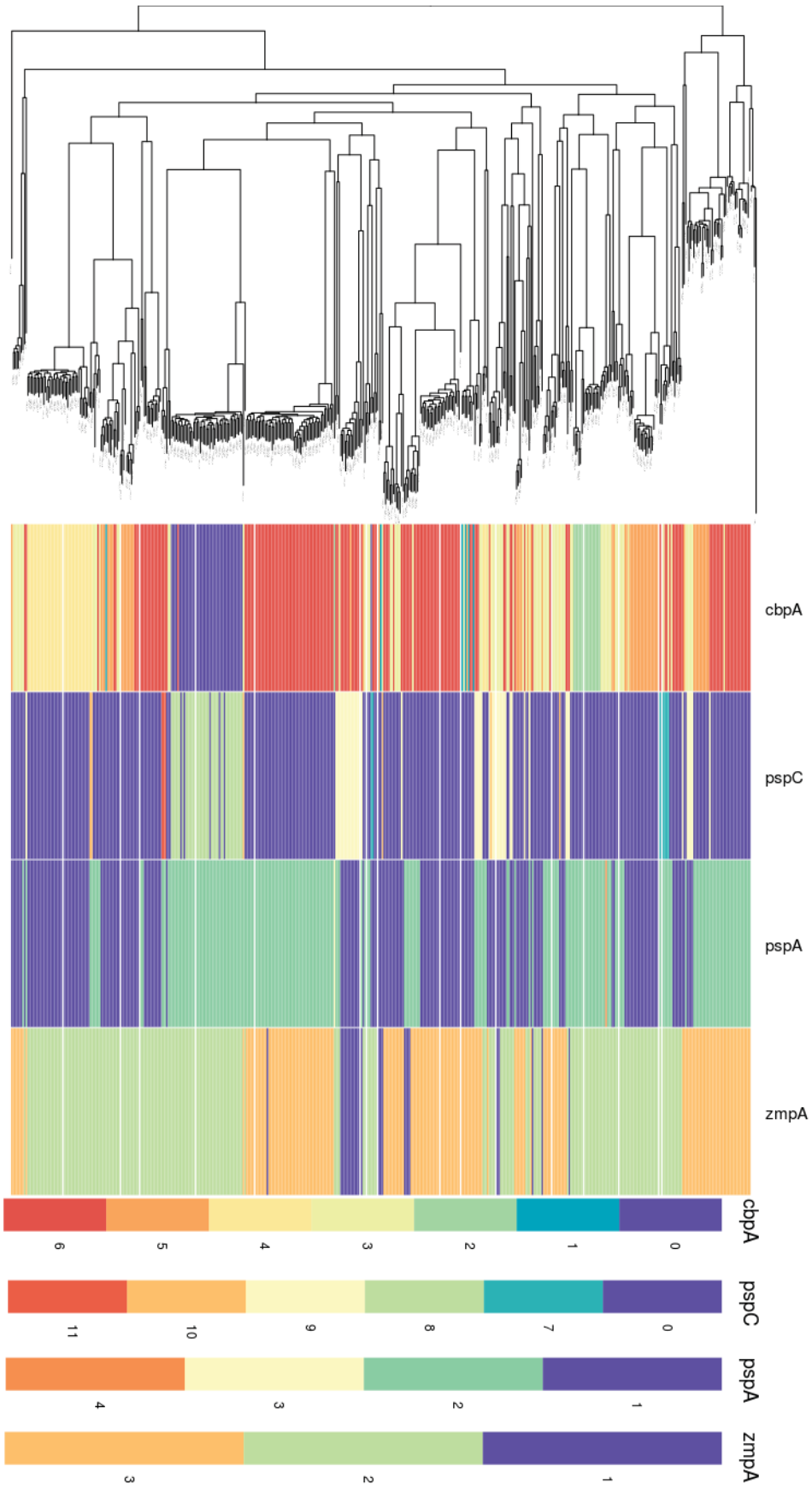


Figure 4.1: The inferred allele of pneumococcal antigens *zmpA*, *pspA* and *pspC*. Left: phylogenetic tree of CSF isolates. Right: tips coloured by the inferred allele for three antigens, and key. The first two columns are alleles 1–6 and 7–11 of pspC, which may have two copies present (Iannelli et al., 2002).

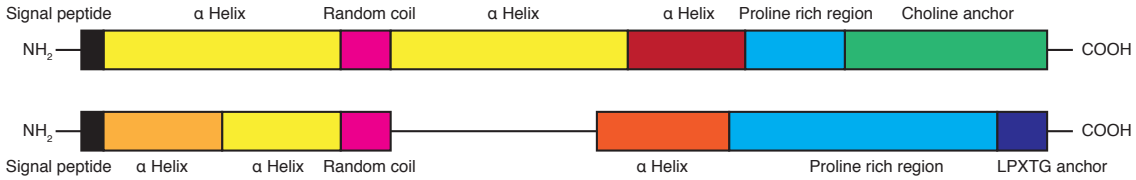


Figure 4.2: Pictographic alignment of the two forms of PspC, as in Iannelli et al. (2002). The top shows cbpA-5; all alleles *cbpA* 1-6 have a choline anchor, and otherwise vary in their α helix content. The bottom shows pspC-7; all alleles *pspC* 7-11 have an LPXTG anchor instead of a choline anchor.

(Brooks-Walter et al., 1999). The locus encoding PspC varies extensively, and two main forms exist (fig. 4.2) which are distinguished by having a choline anchor (alleles 1-6) or a LPXTG anchor (alleles 7-11) (Iannelli et al., 2002). Each genome may encode neither, one or both of these forms and they are normally found in tandem.

I used the existing classification of 11 alleles described by Iannelli et al. (2002), and the 48 sequences reported by these authors (fig. A.10). To allow for the fact that each of the two forms may be present or absent I trained two classifiers. The first, referred to as the *cbpA* allele, used alleles 1–6 and treated 7–11 as missing. The second, referred to as the *pspC* allele, used alleles 7–11 and treated 1–6 as missing. Though there was correlation between the two allele types (for example 4 and 10 were more likely to co-occur) I trained the two classifiers independently. I first checked whether the reference data could distinguish between the labels using PCA, and then predicted two different alleles for each sample.

I tried four different 'out of the box' classifiers: support vector machine (SVM) with a linear kernel, weighted k-nearest neighbours, random forests and DAPC (Jombart et al., 2010). I inspected the statistics and annotations to manually assign the allele pair for 25 genomes from across the tree, then using these truth values and compared the classification accuracy of each method. Table 4.3 shows that the SVM performed best; I used it for all four classifiers. Inspection of the feature importance showed the blastp bitscore, E-value, and number of mismatches as well as the srst2 number of truncated bases and number of mismatches were the most informative predictors.

Method	Balanced accuracy
SVM	0.86
kknn	0.73
Random forest	0.50
DAPC	0.14

Table 4.3: Comparison of classifiers of antigen alleles. The balanced accuracy is given by the average of $\frac{1}{2}(\text{sensitivity} + \text{specificity})$ for all alleles.

pspA and *zmpA* alleles

PspA is a well studied pneumococcal antigen (Crain et al., 1990) which binds C3 (Tu et al., 1999) and lactoferrin (Shaper et al., 2004). Its locus is involved in both ancestral and recent recombination events which has created variation at the locus (Hollingshead et al., 2000; Croucher, Harris, Fraser et al., 2011). ZmpA, also known as Iga, is a zinc metalloprotease which cleaves IgA molecules (Wani et al., 1996). Similarly to *pspA*, the sequence is variable within the population and is under diversifying selection (Bek-Thomsen et al., 2012).

Croucher et al. (2017) have manually created clusters of sequences for both of these antigens using 616 carriage genomes (Croucher, Finkelstein et al., 2013). Sequences were combined into the same allele if their translated sequence was identical, giving 39 possible sequences for *pspA* and 18 possible sequences for *zmpA*. I used these sequences as the reference panel for each allele.

Unlike *pspC* where sequences had been further clustered based on functional domains by Iannelli et al. (2002), this reference panel contained very similar sequences with different allele labels. Using this directly for GWAS would lead to low power as the number of sequences with each allele would be very small, and the classification would also likely be poor due to the relative paucity of reference data for each allele. To avoid this I used the phylogentic relationship between sequences to clustered similar sequences into allele groups before training each classifier.

For both antigens I aligned the reference panel of amino acid sequences using MUSCLE (Edgar, 2004), and built a phylogeny with RAxML with a CAT+GAMMA model. To test the robustness of these phylogenies I ran 100 maximum-likelihood bootstrap replicates, and 10^6 mrbayes Markov-chain Monte Carlo (MCMC) iterations (discarding the first 25% as burn-in, sampling every 10^3 steps) to generate a sample of 750 trees from the posterior distribution. I compared the topology of these trees using treescape (Kendall & Colijn, 2015), and found the placement of ancestral branches of the topology were poorly resolved, though placement of sequences in main clades was well supported. I therefore took a cut through the deep branches of the two phylogenies, defining four alleles for *pspA* (fig. A.11) and three alleles for *zmpA* (fig. A.12). This phylogeny and classification is similar to three families previously defined for *pspA*, and three families previously seen for *zmpA*. Using these alleles I then fitted classifiers to the reference panels as in section 4.3.1, and predicted the allele for all samples in the study.

4.3.2 Phase variable type I R-M system allele (*ivr*)

Croucher, Coupland et al. (2014), J. Li et al. (2016), Manso et al. (2014) have highlighted a potential role in virulence for the *ivr* locus, a type I restriction-modification system with a phase-variable specificity gene allele of *hsdS* in the host specificity domain (fig. 4.3).

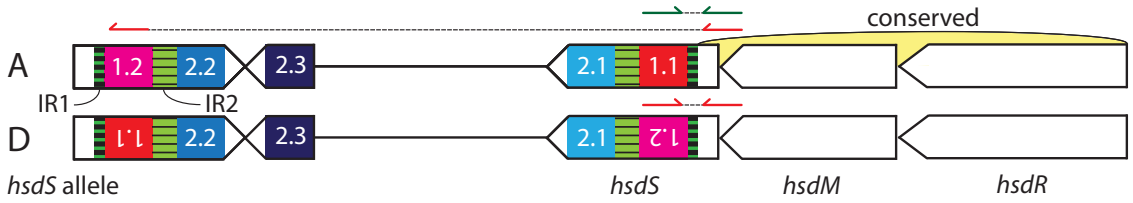


Figure 4.3: The structure of the *ivr* type I restriction-modification locus. The restriction (*hsdR*) and methylation (*hsdM*) subunits, and the 5' end of the specificity subunit (*hsdS*) are generally conserved. Inverted repeats IR1 (85bp) and IR2 (333bp) facilitate switching of downstream incomplete *hsdS* elements into the transcribed region. Top: The green read pair has the expected insert size, and suggests allele A (1.1, 2.1) is present. The red read pair is in the wrong orientation and has an anomalously large insert size. Bottom: The red read pair is consistent with the displayed inversion, suggesting allele D (1.2, 2.1) is present.

There are six possible different alleles A-F for *hsdS*, each corresponding to a different level of capsule expression. Some of these alleles are more successful in a murine model of invasion, whereas others are more successful in carriage.

Due to the high variation rate and structural rearrangement mediating the change the allele cannot reliably be determined using assembly and/or standard mapping of short read data. Instead, I extracted mates of reads mapping to the reverse strand of the conserved 5' region for each sample, and mapped with BLAT (Kent, 2002) to the possible alleles in position 1. This forms a vector r_i of length two for each sample i, with the number of reads mapped to 1.1 and 1.2. Similarly, to determine the 3' allele (position 2), I extracted pairs of reads mapping to each of the reverse strand of allele 1.1 and the forward strand of allele 1.2 and mapped to the three possible alleles in position 2. This forms a vector q_i of length six for each sample i, with the number of reads mapped to each allele A-F.

I performed this on all samples in the collection and found 677 of 693 carriage samples and 1 052 of 1 144 invasive CSF samples had at least one read mapping to an allele of the *ivr* locus *hsdS* gene. In the invasive samples, this corresponded to 621 CSF blood sample pairs. Those without any reads mapping had either a deletion of one component of the locus, or a large insertion mediated by the *ivr* recombinase.

4.4 GWAS of bacterial variants associated with meningitis

While it is well known that pneumococcal serotype contributes to invasive propensity (Hausdorff et al., 2000; Brueggemann et al., 2003), it is of great interest in the field of pneumococcal biology whether variation in other regions of the genome can independently affect invasiveness. Many virulence factors are known to be involved in and essential for pneumococcal colonisation and disease (Kadioglu et al., 2008), but whether natural variation in these regions affects clinical cases of disease has yet to be assessed. Indeed, the overall role of pneumococcal variation in invasive disease is as yet unknown, and therefore the proportion of variation in invasiveness which can be ascribed to the capsule and the proportion due to other factors cannot be determined. Additionally, the lack of

large cohorts combining detailed clinical metadata with bacterial data means that little is known about about the effect of pneumococcal variation on disease outcome. Previous studies with small sample sizes have suggested a role for platelet binding (Tunjungputri et al., 2017) and arginine synthesis (Piet et al., 2014), with additional evidence from *in vitro* observations.

I first performed a heritability analysis to quantify the amount of variation due to the pneumococcal genome for each phenotype. As well as using the methods described in section 3.3 I also applied a phylogenetic mixed model assuming an Ornstein-Uhlenbeck (OU) process of trait evolution as implemented in the patherit package (Mitov & Stadler, 2016), which has previously been shown to be less biased than other techniques for estimating the heritability of pathogen traits (Blanquart et al., 2017). I performed 200 000 MCMC iterations, discarding the first half as burn-in and thinning the chain to every hundredth value. LDAK performs heritability estimation of this binary trait on the liability scale (Lynch & Walsh, 1998). I peformed this analysis within genomes collected from meningitis, stratified using GOS to define clinical outcome, and between genomes from carriage and genomes from meningitis (referred to as 'invasiveness').

Trait	Method		
	LDAK	OU	closest phylogenetic-pairs (CPP)
Invasiveness	0.983 ± 0.003	0.9936 (0.9928-0.9943)	0.995 (0.991-0.998)
Unfavourable outcome	0.006*	did not converge	0.05 (-0.04-0.16)
Death	0.0001*	0.02 (-0.07-0.11)	0.07 (-0.03-0.17)

Table 4.4: Estimated heritability of pneumococcal invasiveness and outcome due to variation of the pathogen genome. Values shown in brackets are the 95% CIs, where provided by the method, for LDAK the standard error is shown, unless the LRT p-value was > 0.05 so there is no support for a non-zero heritability (shown by an asterisk).

Table 4.4 shows the predicted heritability from each method. There is evidence that invasive propensity is highly heritable, but that disease outcome is not determined by natural variation of pathogen genetics. The latter is not surprising as invasive disease as an evolutionary dead end for the pathogen, adaptations affecting virulence over the short course of infection are unlikely to be selected for. The dependence on invasiveness is well known to depend on pneumococcal genetics, but not the degree. The high heritability estimated here, supported by three different techniques, suggests that in this population some bacteria are able to invade while others are not, with almost certainty depending on the genetic background. This is consistent with some serotypes not being found in invasive disease (Hausdorff et al., 2000), and their wide genetic separation from invasive serotypes. The complete heritability is likely an overestimate due to the binary nature of the trait, but does show that pathogen genetics are important in invasiveness and not likely to be important in severity.

I then wished to quantify the amount of this heritability which was due to serotype, which is the current focus of pneumococcal vaccination and the most well known invasiveness determinant, versus other factors. As in section 3.4.1 I used leave-one-out cross validation with lasso logistic regression to select the 36 serotypes (of 63 observed) which were informative of invasiveness. I then assessed the variance in invasiveness explained by these serotypes using Nagelkerke's pseudo R^2 from logistic regression (International Schizophrenia Consortium et al., 2009; Hosmer et al., 2013), which was 0.45. Caution should be used in directly interpreting this R^2 as variance explained, but it does show the model fit from serotype alone is not as good as using the pneumococcal kinships, suggesting there are factors other than serotype which affect invasiveness. I also checked whether invasiveness is well predicted by capsule charge, as has been previously suggested by Y. Li, Weinberger et al. (2013). Using the previously measured zeta potentials, and using the serogroup average when a serotype charge was not available, I performed the same logistic regression using charge as the predictor rather than serotype. Charge significantly affected invasiveness but was not as informative as the specific serotype ($p < 10^{-10}$; Nagelkerke's $R^2 = 0.08$), suggesting a role for finer structure of the capsule structure (Bentley et al., 2006).

In the rest of this section, using the variation defined for all samples as in section 4.3 and the GWAS methods developed in chapters 2 and 3, I tried to find the pneumococcal variants other than serotype which affect invasivness. Even though there is no evidence from the above heritability analysis that variation in the pneumococcal genome contributes to disease outcome I ran the same analysis on these phenotypes anyway – it may be that the common/core variation used to produce these estimates fails to tag variation in the accessory genome or phase variable regions which may contribute to outcome. In this case a lack of association will also provide further support for zero heritability due to the bacterial genome.

In the first section I consider association of common variants in the pan-genome (all of those described in section 4.3) with the phenotypes predominantly using the techniques already described. I then go on to asses the role of rare variation firstly using tests of selection, and more directly using an association combining variants with the same predicted effects. Finally I developed a model to test whether any particular *ivr* allele, or the amount of variation of the allele is associated with any of the phenotypes.

4.4.1 Role of common variation

Using the variants catalogued above, with previously described filtering thresholds, I performed a GWAS between the isolates from invasive disease and asymptomatic carriage, as well as unfavourable outcomes and/or death within the invasive isolates. I used SEER with the first ten MDS components to correct for population structure, as well as FaST-

LMM (Lippert et al., 2011) using the kinship matrix estimated from SNPs and INDELs as random effects.

Figure 4.4 shows the Q-Q plots of the resulting p-values from these methods on SNPs and k-mers with invasive versus carriage isolates. In both cases the test statistic from SEER is clearly highly overinflated for this population and phenotype, meaning a high significance threshold would be needed to remove population structure confounded associations. I have shown that invasiveness is highly heritable, so population structure being highly confounding is unsurprising. Increasing the number of fixed effect population structure covariates may help alleviate this issue, but as the LMM test statistic is better controlled, and as it was a successful method in chapter 3, I have used it for all associations of common variants with the three phenotypes. For significance thresholds I used the unique number of patterns as the number of tests in a Bonferroni correction, giving $p < 8.2 \times 10^{-7}$ for SNPs and $p < 1.9 \times 10^{-8}$ for k-mers. However, inspection of the Q-Q plots shows that for k-mers the LMM is still overinflated, so I have instead taken $p < 1 \times 10^{-16}$ to describe the top hits.

From all three of SNPs, COGs and k-mers by far the most highly associated variants are transposons. These mobile elements of DNA can insert into different places in the bacterial host genome through inverted repeat sequences, and coevolve with the bacterial population (Kleckner, 1981; Levin & Moran, 2011). In some cases transposons can carry cargo genes, such as antibiotic resistance conferring mechanisms, which increase host fitness (Croucher, Harris, Fraser et al., 2011). However, the transposons here appear to be simple elements lacking such cargo, and are therefore unlikely to explain a difference between carriage and invasive isolates directly. Most likely these transposons are present in some genetic backgrounds and not others, and are therefore a population structure confounded result. Their variability in position in the genome and specific sequence may mean they are less well controlled for against genetic background. Due to the lack of plausible functional link with the phenotype I do not consider them further here.

Other hits are shown in table 4.5, ordered by the variant type discovered. In some cases COGs were incorrectly clustered and actually represent two alleles of the a gene orthologs. For three of these alleles I found a positive association with invasive isolates from one allele, and a negative association from the alternative allele. To annotate the genes here I used the best blastp match to the core and accessory genome defined by Croucher, Finkelstein et al. (2013), and if not annotated already I used blastp with the `nt/nr` database to find annotated orthologs, and hmmscan and cd-hit to find functional domains to inform the annotation.

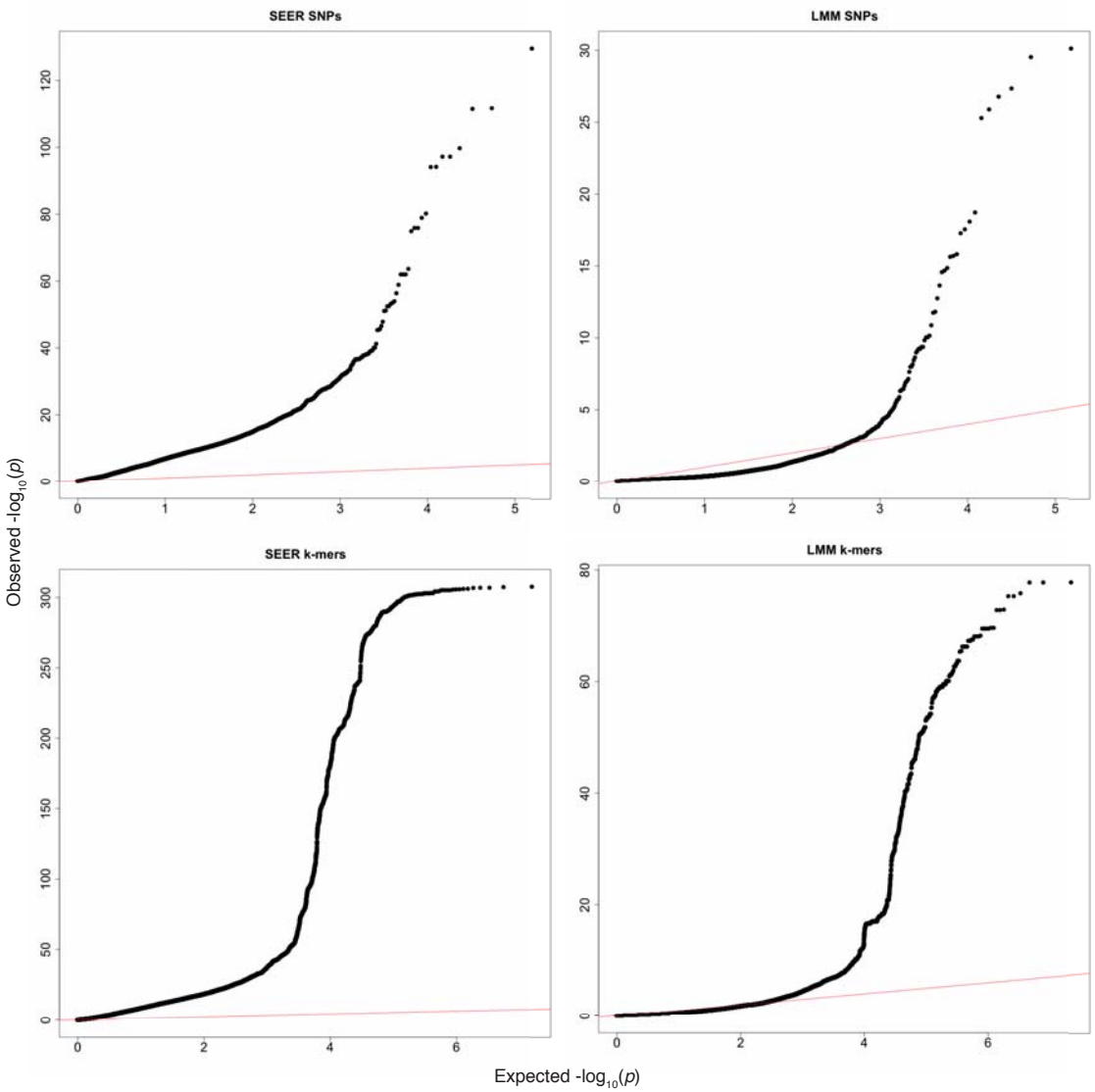


Figure 4.4: Quantile-Quantile plots for invasive *S. pneumoniae* GWAS methods. Red line is for observations following the null-hypothesis of no association, plotted points are observed p-values from each method. Top row: p-values from SNPs and INDELs from mapping; bottom row: p-values from k-mers. Left column: SEER run with the first ten population structure components. Right column: FaST-LMM run on the same input.

Gene ID	Annotation	Core	Method	p-value
FM211187.6011	*tlyC*; Membrane protein (upstream)	Yes	Mapped variants	7.7×10^{-31}
FM211187.977	*pbpX*; penicillin binding protein	Yes	Mapped variants	3.6×10^{-18}
FM211187.313	hypothetical protein (upstream)	Yes	Mapped variants	2×10^{-16}
FM211187.1802	*yhfE*; Aminopeptidase (upstream)	Yes	Mapped variants	1.0×10^{-9}
FM211187.1019	*wzh*; capsule synthesis	No	Mapped variants	3.6×10^{-9}
FM211187.150	*comA*; bacteriocin/competence (upstream)	Yes	Mapped variants	9.9×10^{-9}
FM211187.3083	*pbl3e*/*pldT*; bacteroicin	No	COG absent	4.0×10^{-10}
N/A	transcriptional regulator (pseudogene)	No	COG absent	1.4×10^{-8}
FM211187.3090	bacteriocin precursor	No	COG absent	1.7×10^{-8}
FM211187.6181	FtsX-family transport protein (ABC transporter permease)	No	COG alleles	4.7×10^{-9}
FM211187.6189	C4-dicarboxylate (citrate) ABC transporter	Yes	COG alleles	1.4×10^{-7}
FM211187.5843	23S rRNA (uracil-5-)-methyltransferase RumA2	Yes	COG alleles	5.5×10^{-7}
FM211187.939	galactose-6-phosphate isomerase	No	K-mers	3.0×10^{-60}
N/A	phage-related chromosomal island protein	No	K-mers	3.0×10^{-60}
FM211187.4259	Peptidase U32	Yes	K-mers	1.7×10^{-59}
FM211187.4090	*aroK*; Shikimate kinase	Yes	K-mers	1.7×10^{-59}
FM211187.1923	*yehU*; Sensor kinase	Yes	K-mers	3.1×10^{-59}
FM211187.6369	*patA*; efflux pump (upstream)	Yes	K-mers	2.0×10^{-54}

FM211187.6823	*tauA*; Nitrate/sulfonate/taurine ABC transporter solute-binding protein	Yes	K-mers	1.9×10^{-43}
FM211187.213	Galactose uptake PTS transporter, IIB subunit	Yes	K-mers	2.5×10^{-42}
FM211187.3677	*pyrB*; Aspartate carbamoyltransferase PyrB	Yes	K-mers	4.6×10^{-38}
FM211187.6594	*ulaA*; Pentose PTS transporter IIA	Yes	K-mers	3.7×10^{-25}

Table 4.5: Common variation associated with invasiveness using FaST-LMM. I have annotated the gene the significant locus overlaps, and intergenic variants are annotated with the nearest downstream genes as noted. Gene ID is the name in the ATCC 700669 reference if present; 'core' refers to whether this gene was in the core genome defined by Croucher, Finkelstein et al. (2013); method describes the type of variant that was found to be associated.

The *wzh* gene is involved in capsule synthesis and is part of the gene cassette which determines serotype (Bentley et al., 2006). As shown above and in previous studies, serotype has a large effect on invasiveness and hence this association serves as a positive control. The association of variants in *pbpX* is likely due to mosaic alleles which confer resistance to β-lactams being common in invasive serotypes, similar to what I found in section 3.5.2. The bacteriocins mediate intraspecies competition and determine strain fitness (Dawid et al., 2007), but a specific association with invasiveness independent of strain background has not previously been reported. *comA*, a core gene essential for competence, affects the expression of these bacteriocins so may represent an effect through the same pathway (Kjos et al., 2016).

The adhesin *yhfE* has previously been associated with virulence of *S. pneumoniae* (M. W. Robinson et al., 2013). This adhesin functions as a peptidase, hence the other peptidase may found to be associated also have similar role. Other genes found here previously associated with virulence in animal models include: *ulaA* which utilises ascorbic acid has been found to be upregulated in invasion (Afzal et al., 2015; Mahdi et al., 2015); *pyrB* is involved in cell wall biosynthesis and can affect virulence (Mohedano et al., 2005); *aroK* is involved in biofilm formation (Domenech et al., 2012); both *comA* and *tauA* were found to be essential for growth during meningitis using a genome-wide screen (Molzen, Burghout, Bootsma, Brandt, van der Gaast-de Jongh et al., 2011). For the other identified regions I couldn't find reference to a previous report relating them to a role in invasiveness or virulence of *S. pneumoniae*.

For unfavourable outcome and death, none of the above classes of variant reached genome-wide significance. This is consistent with the low heritability estimated for these phenotypes. No alleles of *pspC*, *pspA* or *zmpA* or any CNVs reached genome-wide significance for any of the phenotypes.

4.4.2 Role of rare variation

The availability of whole genome sequence data for these samples allows the identification of rare variants, here defined as those present in the population with MAF $< 1\%$, which are also plausible as having an effect on the phenotypes of interest. The amount of rare variation compared to common variation present in a population is informative of recent selection and population size changes (Ziheng Yang, 2006). An overall difference may therefore be informative of different selection on regions of the genome depending on the niche. In fig. 4.5a I have plotted the SFS by niche and predicted consequence to look for an overall difference. Across the range of common MAFs in both niches the proportion of synonymous/nonsynonymous/intergenic/LoF mutations is roughly constant and as expected (Ziheng Yang, 2006; Thorpe et al., 2017), though at low frequencies, there is an excess of potentially damaging variants.

Interestingly, there is a clear excess of rare variants in invasive samples compared to carriage samples. To quantify this difference and identify which regions of the genome are responsible for the excess of rare alleles I calculated Tajima's D for each coding sequence in the genome, and looked for differing signs of selection between cases and controls. Tajima (1989) developed the summary statistic D to look for differences between an observed population and an idealised population of a stable size evolving under neutral selection, where mutation frequency is dominated by drift rather than selection. By comparing the number of segregating sites with the average number of differences between pairs of sequences, a statistic D can be calculated. Deviations with $D < 0$ are indicative of selective sweeps and/or recent population expansion, whereas $D > 0$ is indicative of balancing selection and/or recent population contraction. In terms of differences between SFS, a negative D manifests as an excess of rare variants whereas a positive D manifests as a uniform distribution (Bamshad & Wooding, 2003).

For speed, I implemented code in C++ (https://github.com/johnlees/tajima-D) which uses the same optimised strain-wise distance calculation as SEER (section 2.3.2) to calculate the average number of pairwise strain differences $\hat{k}$. Unknown or gap sites are ignored in the calculation, and the codes produces the same value of D on standard test data. The code uses a variant call format (VCF) file as input, so is readily generalisable to other applications. Using this code, I calculated D for all coding sequences in the ATCC 700669 reference separately for carriage and invasive isolates, and the difference in D between niches.

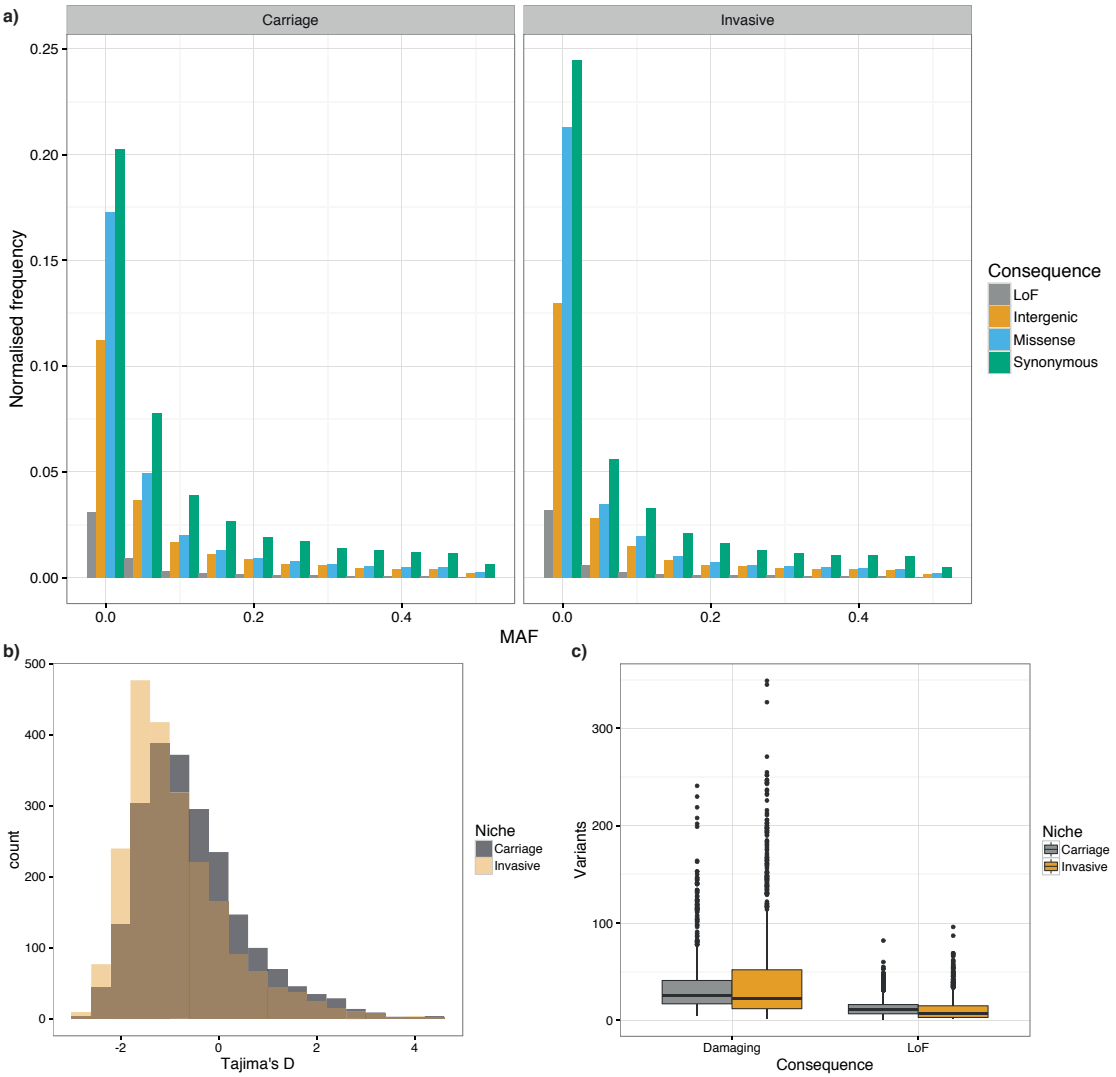


Figure 4.5: Differing burden and frequency of rare variation between invasive and carriage isolates, based on short variation called from mapping against the ATCC 700669 reference genome. LoF are frameshift or nonsense mutations. **a)** The SFS stratified by niche and by predicted consequence. Frequency has been normalised with respect to the number of samples in each population. **b)** Histogram of Tajima's D for all coding sequences in the genome, stratified by niche. **c)** Boxplot of number of rare variants per sample, stratified by niche and predicted consequence. Damaging mutations are LoF mutations and missense mutations predicted damaging by SIFT.

Comparison between D values to test for different selection between niches will only work within the same population, otherwise changing population size may cause an overall difference in D. The assumption that invasive and carriage populations are the same is potentially reasonable, as all invasive isolates must first have been carriage isolates, however the biased selection of case isolates used for GWAS and potential adaptation and population growth after invasion (described futher in section 4.5) may violate this assumption. In GWAS terms, although the calculation of Tajima's D uses rare variation, which is less prone to population structure confounding, common variation is also used which is affected by population structure.

To test for an overall difference I compared the distributions of D by gene in each

phenotype shown in fig. 4.5b. Genes in invasive isolates had a lower average D (difference in medians -0.34; W = 1 996 100, $p < 10^{-10}$) and a more positively skewed D (difference in skewness 0.30; 95% bootstrapped CI 0.17-0.44). This difference in D may be representative of a difference in population dynamics or population structure between niches, or may show genuine differences in selection. To find individual genes which show a difference in selection between niches I then ran 44 000 permutations per gene with randomised phenotype labels to calculate a p-value on the difference in D between niches, to which I applied a Bonferroni correction to adjust for testing of all genes (Winantea et al., 2006). 156 genes had a significantly different D between niches; in table 4.6 I report 18 of these coding sequences which were outside of the 95% central mass of the D distribution for one niche but not the other. Due to potential population structure effects results should therefore be seen as suggestive, and potential for follow-up work.

Gene ID	Annotation	Invasive D	Carriage D	Direction
FM211187.1040	*wzx*; capsule synthesis	-2.53094	-1.79867	Negative in invasive
FM211187.5843	23S rRNA (uracil-5-)-methyltransferase RumA2	-2.4028	-1.63478	Negative in invasive
FM211187.2360	*ezrA*; septation ring formation regulator	-1.1051	-2.17726	Negative in carriage
FM211187.4024	replication initiator protein (on ICE)	-1.55767	-2.16733	Negative in carriage
FM211187.4026	hypothetical, contains FtsK gamma domain (on ICE)	-1.61993	-2.21525	Negative in carriage
FM211187.357	bacteriocin	4.19212	1.30796	Positive in invasive
FM211187.420	*tsaB*; tRNA threonylcarbamoyladenosine biosynthesis protein	3.49345	1.39805	Positive in invasive
FM211187.769	aceytltransferase	2.9055	1.80787	Positive in invasive
FM211187.1019	*wzh*; capsule synthesis	2.76882	1.63677	Positive in invasive
FM211187.1802	*yhfE*; Aminopeptidase	2.28654	1.19784	Positive in invasive

FM211187.1804	bacteroiocin	1.94491	-0.56384	Positive in invasive
FM211187.1806	*dacC*; D-alanyl-D-alanine carboxypeptidase	2.23028	0.722447	Positive in invasive
FM211187.5184	*dnaI*; primosomal protein	2.32212	0.197632	Positive in invasive
FM211187.3651	*tarI*; Ribitol-5-phosphate cytidylyltransferase	-0.146237	2.34171	Positive in carriage
FM211187.3804	*nanB*; neuraminidase	1.6805	3.19937	Positive in carriage
FM211187.5053	membrane protein	0.311619	2.46774	Positive in carriage
FM211187.5358	*secY*; accessory secretion system translocase	0.471641	2.36541	Positive in carriage

Table 4.6: Coding sequences with extreme values of Tajima's *D*, with a difference between carriage and invasive isolates as determined by permutation testing.

A positive *D* statistic implies common variants are being maintained in the population more than expected, suggesting that multiple alleles of the gene are common. The positive estimates of *D* in bacteriocins are consistent with their function, where having a different allele to competing strains is advantageous and increases fitness (Bogaardt et al., 2015; Miller et al., 2017). *nanB* is similarly involved in competition and in virulence (Shakhnovich et al., 2002; Brittan et al., 2012); the difference in *D* I found suggests that this selection may be more important in carriage where more common alleles appear to be maintained. A negative *D* suggests purifying selection acting on a gene. For example, *ezrA* is essential for growth in carriage (van Opijnen et al., 2009; Cleverley et al., 2014), so a negative *D* suggests that changes to the protein are not tolerated in this niche. As *wzx*, *wzh*, *yhfE*, RumA2 and bacteriocins were found to be associated with invasiveness above, this suggests that the difference in *D* I observed is less likely to be due to population stratification and more likely a real sign of selection. Genes found through these approach which may affect cell growth such as *ezrA*, *secY*, *dnaI* and *tarI* may make the population more or less immune stimulating, depending on their direction of effect.

Burden testing of coding sequences

I then wished to consider whether rare variants were associated with any of the three phenotypes. These variants will have occurred on terminal (or close to terminal) branches and therefore population structure is less of an issue than for common variants. Power to detect associations is proportional to MAF and OR, so and at low MAF, there is only power to detect those variants with a large effect size (Liu & Anderson, 2014). However, for rare alleles the statistical tests described so far lack the power to test for an association even for an infinite OR. In human genetics combining sets of variants with the same predicted effect on a more complex biological function (yet simpler than the whole phenotype), for example grouping rare LoF variants in the same gene, then testing the group for association with the phenotype of interest has been the most common approach (B. Li & Leal, 2008; Morris & Zeggini, 2010). This is known as a burden test – in bacterial genomes this technique has successfully found LoF variants associated with antibiotic resistance in *M. tuberculosis* (Desjardins et al., 2016).

In each test I used only variants with MAF $< 1\%$ from the variant calls derived from mapping. Using the annotations from VEP, I defined frameshift and stop gained mutations as LoF – 6 825 variants in total. I also analysed the effect of all predicted missense variants using Provean (Ng & Henikoff, 2003; Choi et al., 2012), and used the default threshold of -2.5 to select variants with a predicted effect on protein function – 26 206 of 50 383 missense variants passed this threshold. I combined these variants with LoF variants to define a damaging class. Figure 4.5c shows the overall burden of damaging rare variants between carriage and invasive samples; in both classes there was higher burden in carriage isolates (median LoF: invasive 7, carriage 11, W = 297 440, $p < 10^{-10}$; median damaging: invasive 22, carriage 26, W = 345 370, $p = 8 \times 10^{-4}$), so results showing a burden in carriage should be interpreted with caution.

I then used `plink/seq` to perform a burden test on all coding regions in the ATCC 700669 reference genome, which looked for an excess of rare damaging alleles in genes, and Bonferroni corrected all resulting p-values. I tested all six possible phenotypes: invasiveness, carriage, favourable outcome, unfavourable outcome, survival, death. For the latter four phenotypes based on clinical outcome no genes showed a significant burden of LoF or damaging variants. Table 4.7 shows the results for carriage and invasive isolates.

Gene ID	Annotation	p-value	Class	Direction
FM211187.1036	*wchV*; capsule synthesis	0.0022	LoF	Carriage
FM211187.1143	membrane protein	0.0022	LoF	Carriage
FM211187.1634	*bglG*; transcription anti-terminator	0.0022	LoF	Carriage
FM211187.3315	*zmpD*; zinc metalloprotease	0.0022	LoF	Carriage
FM211187.4588	*pclA*; collagen-like surface-anchored protein	0.0022	LoF	Carriage
FM211187.4679	platelet binding phage protein	0.0022	LoF	Carriage
FM211187.4714	prophage protein	0.0022	LoF	Carriage
FM211187.4939	membrane protein	0.0022	LoF	Carriage
FM211187.5113	*nanA*; neuraminidase	0.0022	LoF	Carriage
FM211187.5328	uncharacterised repeat protein	0.0022	LoF	Carriage
FM211187.5369	PsrP glycosyltransferase	0.0045	LoF	Carriage
FM211187.6773	*dusB*; tRNA-dihydrouridine synthase	0.0045	LoF	Carriage
FM211187.1025	*wze*; capsule synthesis	0.0067	LoF	Carriage
FM211187.4017	hypothetical protein (on ICE)	0.0067	LoF	Carriage
FM211187.1040	*wzx*; capsule synthesis	0.0089	LoF	Carriage
FM211187.92	cell wall-binding amidase/autolysin (pseudogene)	0.0089	LoF	Carriage
FM211187.6861	*comFC*; competence	0.011	LoF	Carriage
FM211187.6608	*pcpA*; choline binding protein	0.016	LoF	Carriage
FM211187.4717	prophage protein	0.018	LoF	Carriage
FM211187.2642	chlorohydrolase	0.029	LoF	Carriage
FM211187.5374	PsrP glycosyltransferase	0.038	LoF	Carriage
FM211187.1804	bacteriocin	0.039	LoF	Carriage

FM211187.3950	conjugal transfer protein (on ICE)	0.042	LoF	Carriage
FM211187.3204	*ybaB*; DNA-binding protein	0.0089	Damaging	Carriage
FM211187.4311	multidrug transporter	0.050	Damaging	Carriage
FM211187.4424	sortase-sorted surface anchored protein (pseudogene)	0.0067	LoF	Invasive
FM211187.2661	*bceA*; ABC exporter ATPase	0.0045	Damaging	Invasive
FM211187.3585	*smc*; Chromosome partition protein	0.0045	Damaging	Invasive
FM211187.5524	*trpD*; anthranilate phosphoribosyltransferase	0.0045	Damaging	Invasive
FM211187.2550	*fruA*; Fructose PTS ABC transporter	0.027	Damaging	Invasive
FM211187.3460	*ispA*; Farnesyl diphosphate synthase	0.038	Damaging	Invasive
FM211187.2615	*pfkA*; ATP-dependent 6-phosphofructokinase	0.042	Damaging	Invasive

Table 4.7: Burden testing of rare LoF and damaging variants in coding sequences associated with invasive or carriage isolates. P-values are Bonferroni corrected using the total number of genes.

Those regions found with a larger number of LoF variants in carriage than disease represent genes which are advantageous in invasion, and hence include a number of well-known virulence factors. Specifically, capsule related genes, *zmpD* and *nanA* have all been previously described as increasing virulence in animal models (Brueggemann et al., 2003; Bek-Thomsen et al., 2012; Brittan et al., 2012) and have some overlap with associations found through common variant association. The large effect size caused by these LoF mutations is similar to the gene knock-outs used in these experiments.

As well as these well-described virulence factors, I found four more genes which were more likely to be functional in invasive isolates which had been previously described as virulence related in a single or small number of studies. PsrP is an adhesin which has been shown to increase virulence in mice (Obert et al., 2006; Shivshankar et al., 2009), and found here were two genes which affect the protein's function. *pcpA* (Glover et al., 2008; Sánchez-Beato et al., 1998) and *pclA* (Paterson et al., 2008) are choline binding and surface

anchored proteins respectively, both previously associated with virulence. Tunjungputri et al. (2017) have reported association with presence of the phage-derived platelet binding protein PblB with 30-day mortality of meningitis in humans – the platelet binding protein found here may have a similar role in invasiveness (though I did not find it to be associated with severity or mortality).

I could not find previous reports of association with virulence or invasive potential of the other hits in this category. Also, few of the genes found to be essential in a mouse model of meningitis (Molzen, Burghout, Bootsma, Brandt, van der Gaast-de Jongh et al., 2011) were found here, suggesting either that the induced variants do not occur in natural populations, that the mouse does not perfectly model human meningitis or that the sample size here was too low to discover these effects.

Only one gene was found to lose function more frequently in invasive disease, though as it is a pseudogene in the reference this is unlikely to be a real functional effect. For missense variants affecting protein function the direction of effect is less clear, as the variants may be fitness increasing or decreasing. This inconsistent direction may also make the burden test less powerful, and a test which does not rely on this assumption such as the SKAT test may be preferred (Wu et al., 2011; S. Lee et al., 2012). In carriage isolates, including missense variants also found *ybaB* and a multidrug transporter to be significantly altered in carriage but not in invasion. In invasive isolates a few more possible hits were found. *smc* is involved in cell division and growth, but also has epistatic links to much of the rest of the chromosome (Skwark et al., 2017). LoF in *trpD* has previously been associated with attenuated virulence (Hava & Camilli, 2002), and *fruA* as being associated with the switch in virulence between nasopharyngeal colonisation and bloodstream invasion (Trappetti et al., 2017).

4.4.3 Hierarchical Bayesian model for ivr allele prevalence

Manso et al. (2014), J. Li et al. (2016) have reported an association with *ivr* allele and invasive propensity in a murine model; this dataset offers the opportunity to test whether such as an association exists in clinical samples. As the *ivr* varies rapidly and independently from population structure (Croucher, Coupland et al., 2014) a simple association test can be performed for each allele. I first used the mapping approached described in section 4.3.2 to determine the *ivr* allele for each sample. However, as even a single colony contains heterogeneity at this locus, simply taking the allele with the most reads mapping to it in each sample gives a poor estimate of the overall presence of each allele in the invasive and carriage niches. To take into account the mix of alleles present in each sample, and to calculate confidence intervals, I developed a hierarchical Bayesian model for the allele in each niche (fig. 4.6). This simultaneously estimates the proportion of each colony pick with alleles A-F for both individual isolates (π), and summed over all the samples in each

niche (μ). The model is applied this over i samples and c niches (in this case c can be blood, CSF or carriage).

I first modelled the state of the 5' allele (TRD1.j) only. For the two possible alleles 1.1 and 1.2, the number of reads mapping to each allele (a 2-vector r_i) was used as the number of successes in multinomial distribution z_c (c – index for niche). From these I inferred the proportion of each allele in each individual sample π_i, and in each niche overall μ_c. This was done by defining Dirichlet priors expressing the expected proportion of an allele in a given sample π_i to be drawn from a Dirichlet hyperprior representing the proportion of the allele that is found in each niche as a whole μ_c. The κ parameter sets the variance of all the individual sample allele distributions π_{ic} about the tissue average μ_c, with a higher κ corresponding to a smaller variance.

The hyperparameter A_μ, which encodes the total proportion of each allele we expected to see over all samples, was set to the average amount of the allele observed from the long range polymerase chain reaction (PCR) in a subset of 53 paired samples, as described in section 4.5.4.

The observed number of reads mapping to each allele, prior distributions defined above, and structure of the model in fig. 4.6 defines a likelihood which can be used to infer the most likely values of the parameters of interest π and μ. I used `Rjags` to perform MCMC sampling to simulate the posterior distribution of these parameters. I used 3 different starting points (i.e. three chains), and took and discarded 30 000 burn in steps, followed by 45 000 sampling steps. Noticeable auto-correlation was seen between consecutive samples, so only every third step in the chain was kept when sampling from the posterior. I manually inspected plots of each hyperparameter value and mean at each point in the chain, as well as the Gelman and Rubin convergence diagnostic, which showed that the chains had converged over the sampling interval.

To model both the 5' end (TRD 1.1 and 1.2) and the 3' end (TRD 2.1, 2.2 and 2.3) together, so each isolate i is represented by an allele A-F, for each isolate the total number of reads mapping n_i was drawn from the distribution in equation eq. (4.1)

$$n_i \sim \sum_j \pi_{ij} \cdot r_{ij} \tag{4.1}$$

where j is the index of the TRD region, r_{ij} is the number of reads in sample i that had a mate pair downstream from TRD1.j mapping to any TRD2 region, and π_i is the posterior for allele frequency in the sample.

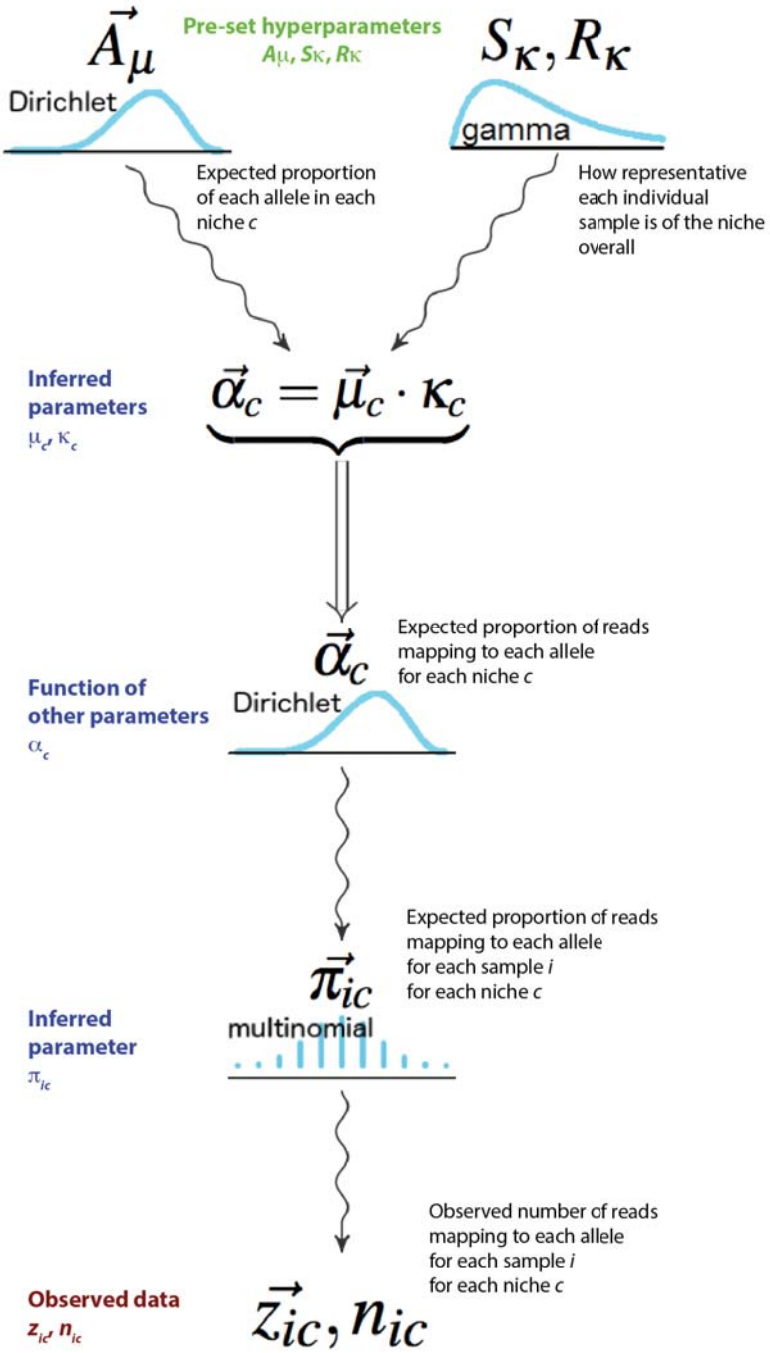


Figure 4.6: Hierarchical model for *ivr* allele. Solid double arrows denote a deterministic relationship; wavy arrows represent a value drawn from a distribution. z is a vector of the number of reads mapping to each allele from a total of N reads mapping to the variable region; i is the sample number; c is an index for tissue type. μ_c, κ are hyperparameters for mean allele prevalence and how closely a sample is representative of a tissue type respectively. A_μ, B_μ are priors for allele prevalence in invasive disease. S_κ, R_κ are the shape and rate parameters for a gamma distribution, which were used to set a broad prior on κ.

The distribution for the number of reads mapping to each allele j, z_{ij} was then defined as in equation eq. (4.2)

$$z_{i,j} \sim \begin{cases} n_i \cdot \dfrac{q_{i,j}}{\|\vec{q}_i\|} \cdot \pi_{i,1.1}, & \text{if } j \in \text{A, B, E} \\ n_i \cdot \dfrac{q_{i,j}}{\|\vec{q}_i\|} \cdot \pi_{i,1.2}, & \text{if } j \in \text{C, D, F} \end{cases} \tag{4.2}$$

where q_i is a vector of length six which contains the number of reads mapped to each allele A-F as described above, and π, i and n are as previously. A single sample for z was taken for each isolate i. This 6-vector z_{ij} is then used as the observed data in the same model as above to infer π_i, and μ_c for the whole locus allele (A-F) rather than just the 5' end.

For the 5' allele (TRD1.j) a model using a single κ parameter rather than a κ indexed by tissue c was preferred (change in deviance information criterion $\Delta\text{DIC} = -0.523$ (Spiegelhalter et al., 2002)). For the 3' allele (TRD2.j), a model with a single κ parameter did not converge. A model with κ indexed by allele was used instead.

This simultaneously estimated the proportion of each colony pick with alleles A-F for both each individual isolate (π), and summed over all the samples in each niche (μ). I applied this over i samples and c niches (in this case c can be carriage/nasopharynx or CSF). The difference in mean of μ (corresponding to the mean allele frequency over all sample pairs for each allele) shows whether alleles are selected for in carriage or invasive disease, however as the confidence intervals overlapped for alleles, no particular allele was associated with invasive disease or carriage isolates. I also checked the diversity of alleles present in each sample by calculating the Shannon diversity index for each sample using the π vector. The median diversities were not significantly different (carriage 0.94; invasive 1.00).

The finding that *ivr* allele does not associate with invasive disease is at odds with the interpretation of Manso et al. (2014) that the capsule expression changes caused by each allele (through genome-wide methylation profile changes) are central to colonisation and disease. I found that, in clinical cases of meningitis, the allele of the *ivr* locus continues to be phase variable regardless of the niche the bacteria are in. Its purpose is likely to defend against phage (Croucher, Coupland et al., 2014), with little effect on disease course in natural human infection.

4.5 Genetic adaptation over the course of single infections

This section concerns whether the invasive pneumococcal population accumulates mutations as it moves from carriage, through blood to the CSF, and if it does whether this mutation represents adaptation to either of these niches. By sampling the same population longitudinally the issue of population structure is not an issue as for the convenience samples of cases and controls collected for GWAS, which will not be from the same population of bacteria. I called variation between pairs of samples (table 4.8), and looked for convergent evolution between different cases and/or signals of adaptation to a specific niche.

Organism	Number of pairs sequenced		Mean coverage
	blood/CSF	nasopharynx/CSF	
S. pneumoniae	674	6	91.7x
N. meningitidis	195	48	96.6x

Table 4.8: The number of paired samples analysed from the MeninGene study, and the average sequencing coverage.

I made assumptions about the evolution of bacteria within the host, under which I discuss the power of pairwise comparisons between single colonies taken from each niche to capture repeated evolution occurring post-invasion:

1. There is a bottleneck of a single bacterium upon invasion into the first sterile niche (usually blood), which then founds the post-invasion population (Gerlini et al., 2014; Moxon & Murphy, 1978).

2. A large invasive population is quickly established, as the population size approaches the carrying capacity of the blood/CSF. The population size is large enough for selection to operate efficiently.

3. As infection occurs in a mass transport system, populations are well mixed without any substructure. Therefore, the effective population size equals the census population size.

4. The bacterial growth rate within blood and CSF is similar.

Initially the population size is small, so selection is inefficient and the population-wide mutation rate is low. However, the eventual carrying capacity (the maximum number of cells) of the blood and CSF are large enough ($> 1.5 \times 10^5$ colony forming units (CFUs)) (Brown et al., 2004; La Scolea & Dryja, 1984) for beneficial mutations to fix rapidly. Due to the short generation time of around an hour (Allegrucci et al., 2006), this carrying capacity is reached early in the course of the disease (after 1-2 days) (Gang et al., 2015).

Crucially, population sizes where selection acts efficiently (Patwa & Wahl, 2008) are reached even earlier than this – a few hours after invasion. Therefore, mutations with a selective advantage occurring after the first stages of infection will eventually become fixed in the niche's population. So, sequence comparison between colony picks from each niche is likely to find adaptation that has occurred post invasion.

Similarity of the bacterial growth rate within blood and CSF is an important assumption because in 45% of the pneumococcal cases there was evidence that CSF invasion happened before blood invasion (patients had a documented prior CSF leak, otitis media or sinusitis (Brouwer, Heckenberg et al., 2010; Heckenberg et al., 2012)). This allowed me to search for post-adaptation invasion that happens in either direction in this species. I investigated the validity of this assumption using analysis of data on the *ivr* locus (section 4.5.4).

In carriage samples, although the population size is small (Y. Li, Thompson et al., 2013) carriage episodes can persist over many months (chapter 3), therefore allowing the potential for mutations conferring an advantage in an invasive niche to arise. Additionally, during carriage there is known to be population wide diversity (Cremers et al., 2014) and in some cases competition between strains (Cobey & Lipsitch, 2012). I only had access to the sequence of a single strain sampled from this diverse pool, which means I had less power to detect mutations either side of the bottleneck. Combined with the small sample size, this means only adaptive mutations with large selective advantages could be discovered in this part of the study.

Finally, I considered whether the culturing process will bias the results. In *S. pneumoniae* I found that two additional passages of the previous sample pair resulted in one additional insertion. In *N. meningitidis* a low rate of variation and no selection on phase-variable regions and no variation of other regions have been observed during the culture steps (Fransen et al., 2009; van der Ende et al., 1995; van der Ende et al., 2000). I therefore concluded that there will be minimal bias introduced during culturing, and that which is introduced will increase the frequency of mutations between pairs without bias towards either blood or CSF. Due to the higher power to detect variation between the blood and CSF, I present those results first in section 4.5.2, and the carriage/CSF results in section 4.5.5.

4.5.1 Reference free variant calling

As the amount of variation beween blood and CSF isolate pairs is very low, I needed to ensure I had sufficient power to call variants and did not suffer from an elevated false negative rate. I used the same simulation setup as in section 4.2, except generated an average of only 200 mutations between 100 simulated sample pairs.

To avoid reference bias, and missing variants in regions not present in an arbitrarily chosen reference genome, I then performed reference free variant calling between all sequence pairs of isolates using two methods: the 'hybrid' method (Uricaru et al., 2014)

and Cortex (Iqbal et al., 2012). The former uses de novo assembly of the CSF sequence reads, mapping of reads from both the blood and CSF samples back to this sequence, then calling variants based on this mapping. Cortex uses an assembly method that keeps track of variation between samples as it traverses the de Bruijn graph.

In the hybrid method I used the SPAdes assembly of the CSF sample as the reference, then mapped reads from both members of the sample pair to this sequence using SNAP (Zaharia et al., 2011) followed by variant calling with bcftools v1.1 (H. Li, 2011) using the command:

```
samtools mpileup -C 50 -m 2 -F 0.0005 -d 1000 -t DP,SP -g -
    p -L 1000 -f assembly.fa mapping.bam | bcftools call -vm
     -P 1e-3 samples.txt
```

I filtered variants with QUAL < 50, MQ < 30, SP > 30, MSQB < 0.001, RPB < 0.001 or DP < 4 out.

For Cortex I first error corrected sample reads using quake (Kelley et al., 2010) to prevent false positive calls supported by very low coverage of reads. I then used the joint workflow of cortex with each set of corrected reads in its own path in the de Bruijn graph, and bubble calling was used to produce a second set of variants between samples. SNPs in the error corrected reads were also called using the graph-diff mode of SGA (Simpson & Durbin, 2012).

I then called variants between these sequences and a draft R6 assembly from simulated read data using both of the above methods; comparison with the mutations known to be introduced allowed power and false positive rate to be calculated – separately for SNPs and INDELs.

In addition to in silico simulation, I cultured blood/CSF paired strains 4038 and 4039 (Croucher, Mitchell et al., 2013) and resequenced them using the same 100bp Illumina paired end sequencing as the rest of the isolates in the study. The genomes of strains 4038 and 4039 have been exhaustively analysed using multiple sequencing technologies (Illumina, 454 and capillary sequencing), so represent high quality positive control data to assess the calling methods. I tested both methods on these data.

The highest power was achieved using hybrid mapping for SNPs and Cortex for INDELs: median power for calling SNPs was 90% using hybrid mapping, and 74% for INDELs using cortex (fig. 4.7a). SGA recovered few true variants. I therefore used this combination of methods, mapping for SNPs and cortex for INDELs, across all samples. When applied to the paired strains 4038/4039 the same mutations as originally reported are recovered, plus a 37bp insertion in *cysB* which was found to be introduced during culturing.

I used simulations to compare against a simple method of mapping against an arbitrary reference, in this case TIGR4 (Tettelin et al., 2001). I found my reference free method has

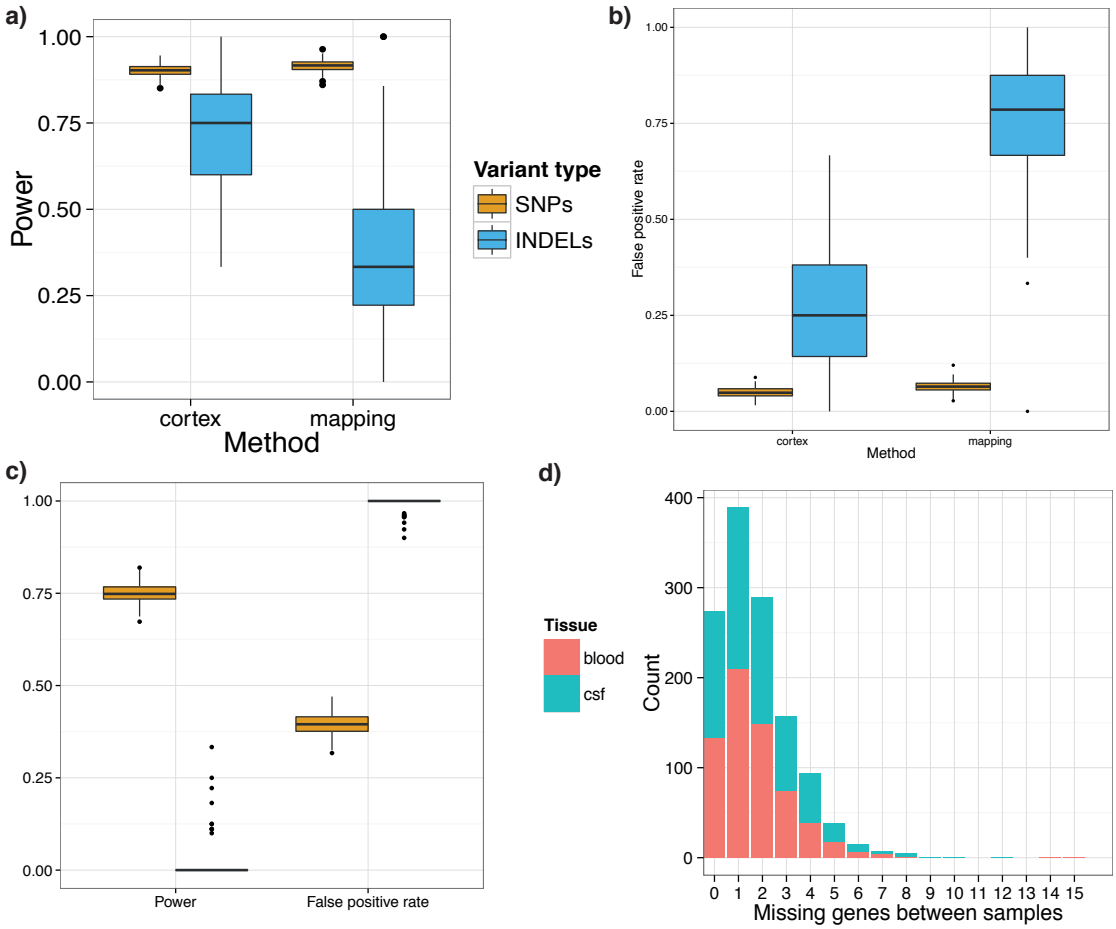


Figure 4.7: Performance of variant calling methods. SNPs (gold) and INDELs (blue) are shown separately. **a)** Boxplot of power (recall) for each method of variant calling for 100 simulated samples. **b)** shows the false discovery rate. **c)** Boxplot of power and false positive rate for reference based calling. Run on the same 100 simulated samples as a), calculated by number of false positives/number of true positives. **d)** Count of annotated genes present in blood but not CSF (red) or vice-versa (turquoise) between the 673 *S. pneumoniae* samples. The level of variation is inflated due to frequent misannotation of coding sequences (CDS)s.

greater power, especially for INDELs (fig. 4.7c), and a markedly reduced false positive rate. I also tested an assembly method alone to compare gene presence and absence, but this too suffered from a vastly elevated false positive rate (fig. 4.7d).

Variant direction and effect annotation

To be able to compare between samples using a consistent annotation, I mapped the called variants to the ATCC 700669 reference (Croucher et al., 2009) for *S. pneumoniae*, and MC58 reference (Tettelin et al., 2000) for *N. meningitidis*. This was done by taking a 300 base window around each variant and using blastn on these with the reference sequence. 'Directionality' was then relative to the reference used, and a binomial test with $\lambda = 0.5$ was used to test significance. I used VEP (McLaren et al., 2010) to annotate consequences of each SNP as synonymous, non-synonymous, or stop-gained and INDELs as frameshift or inframe.

4.5.2 No repeated post-invasion adaptation in coding regions across species

For each species I then counted the number of variants of any type between each blood/CSF isolate pair taken from a patient (fig. 4.8). In *S. pneumoniae* 452 of 674 paired samples (67%) were identical. The distribution of number of variants between isolate pairs is roughly Poisson (mean = 0.547), excluding outliers. Variation between *N. meningitidis* pairs also followed a roughly Poisson distribution (mean = 2.34), which when compared to *S. pneumoniae* showed a higher number of variants between blood and CSF isolates (Wilcoxon rank-sum test, W = 25 790, $p < 10^{-10}$) such that most pairs have at least one variant between the blood and CSF samples.

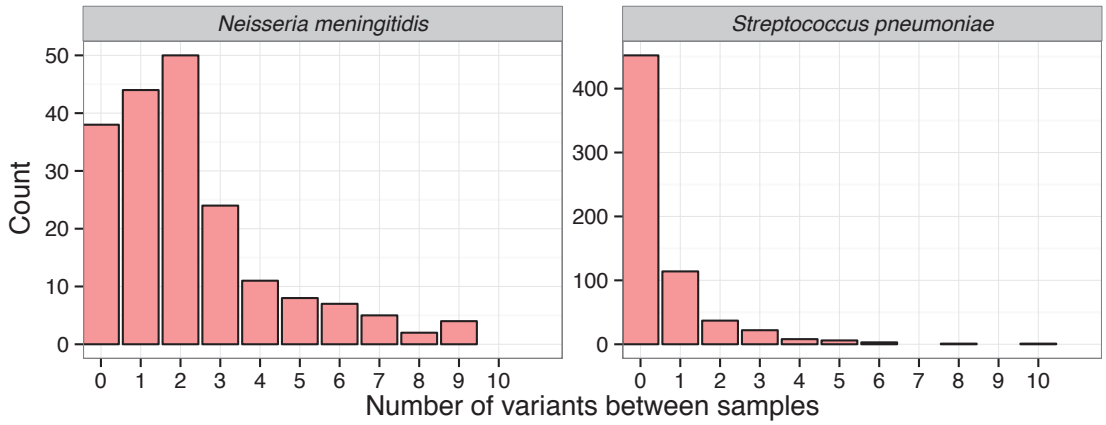


Figure 4.8: Histograms binned by number of variants between a blood/CSF sample pair, for both pathogens. Total pairs analysed in table 4.8. SNPs are from mapping, INDELs are from cortex. Three *S. pneumoniae* and one *N. meningitidis* sample with over 10 variants are not shown.

To test whether certain genotypic backgrounds were associated with a higher number of mutations that occurs post-invasion, I performed a linear fit of each MLST against number of mutations between blood and CSF isolates. I Bonferroni corrected the p-values of the slope for each MLST; at a significance level of 0.05 no MLST was associated with an increased number of mutations.

In both species, the mutations that do exist, if they cause the same functional change, could represent a signal of adaptation. To determine whether this is the case, the number of mutations in each CDS annotation was counted. I then performed a single-tailed Poisson test using the genome wide mutation rate per base pair multiplied by the gene length as the expected number of mutations. The resulting p-values were corrected for multiple testing using a Bonferroni correction with the total number of genes tested as the m tests; I have reported results with $p < 0.05$ in table 4.9.

Gene name	Gene length (bp)	Blood/CSF mutations	p-value
pde1 (SPD_2032)	1973	19	$< 10^{-10}$
dltD (SPD_2002)	1269	13	$< 10^{-10}$
dltB (SPD_2004)	1245	12	$< 10^{-10}$
dltA (SPD_2005)	1551	11	$< 10^{-10}$
clpX (SPD_1399)	1233	7	1.3×10^{-8}
wcaJ (SPD_1620)	693	6	3.4×10^{-8}
cysB (SPD_0513)	909	5	1.6×10^{-5}
cbpJ	1122	5	4.7×10^{-5}
amiC (SPD_1670)	1332	4	6.0×10^{-3}
marR	435	3	9.6×10^{-3}
fhuC	519	3	1.6×10^{-2}

Table 4.9: Genes containing significantly repeated mutations between blood and CSF isolate pairs in *S. pneumoniae*. Ordered by increasing p-value; locus tags refer to the D39 genome, if present.

The *dlt* operon, responsible for D-alanylation in teichoic acids in the cell wall (Deininger et al., 2007; Habets et al., 2012; Kovács et al., 2006), was the most frequently mutated locus: 36 mutations in 31 sample pairs (Poisson test $p < 10^{-10}$). This occurred in only 5% of samples, so adaptation to a niche due to variation in genes is not common. To investigate whether this represented adaptation to either blood or CSF, I annotated the effect of these variants, and determined whether they were specific to a niche. I mapped them to the R6 *S. pneumoniae* strain, which has a functional *dlt* operon and was therefore assumed to be the ancestral state. There was no directionality to the mutations: 19 occurred in the blood, and 11 in the CSF ($p = 0.2$). Only seven of the patients infected by these strains showed signs of blood invasion before CSF invasion (sinusitis or otitis); this also did not show directionality. I have plotted the position and nature of the mutations in fig. 4.9. Most of these mutations would be expected to cause LoF in the operon. Though this suggests this locus has a deleterious effect in invasive disease generally, the lack of directionality to the mutations means it does not show evidence of adaptation to either the blood or CSF specifically.

The next most significantly mutated gene was *pde1*. The *pde1* gene was first found to be essential for growth in an experimental meningitis model (Molzen, Burghout, Bootsma, Brandt, Der Gaast-De Jongh et al., 2011); further study by Cron et al. (2011) showed that *S. pneumoniae* mutants with *pde1* (SP2205 in TIGR4; SPD2032 in D39) and its paralogue *pde2* (SP1298 in TIGR4; SPD1153 in D39) knocked out exhibited reduced host cell adherence and attenuated virulence in a mouse model of meningitis. Following work confirmed that Pde1 acts as a phosphodiesterase, cleaving c-di-AMP into pApA (Bai et al., 2013; Kuipers et al., 2016). These signalling molecules are known to have broad effects

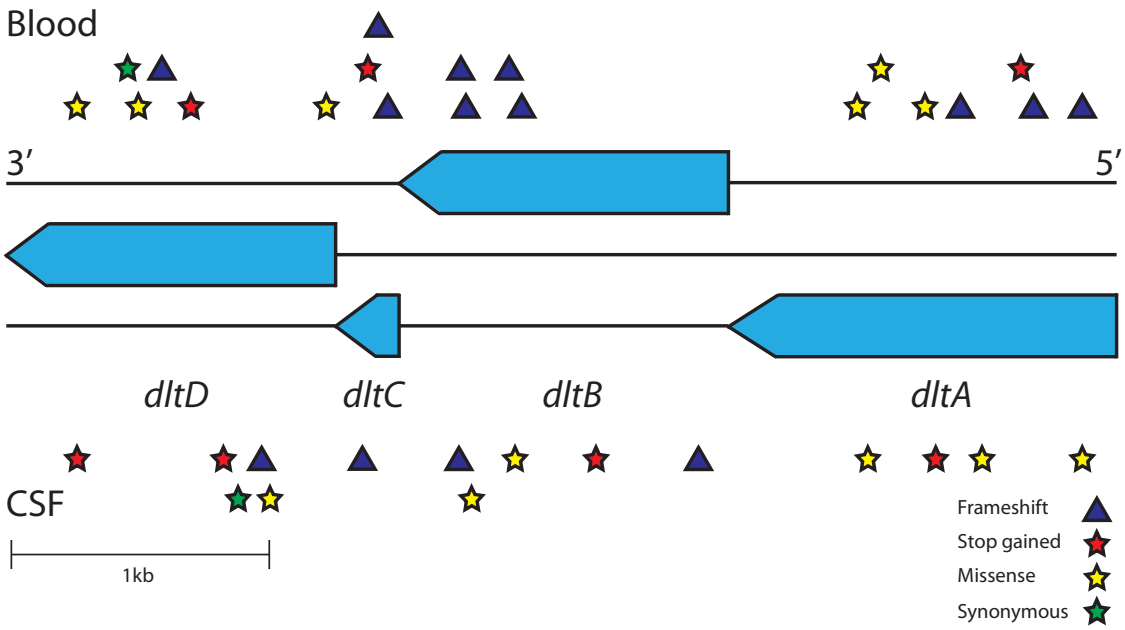


Figure 4.9: Mutations observed between all paired samples in the dlt operon. The operon consists of four genes in the three reading frames of the reverse strand. Mutations, displayed by type, in the blood strains are shown above the operon, and in the CSF strains below the operon.

on the cell (Tamayo et al., 2007) and were again shown to affect growth and virulence in a mouse model of pneumonia. In both studies, the authors suggested that these proteins are promising vaccine targets.

I therefore checked whether *pde1* appeared to be under selection in the sampled population. The ratio of nonsynonymous to synonymous mutations was neutral (dN/dS = 0.89) and contained variants with a SFS similar to that of other genes (fig. 4.10a and b; Tajima's $D = -1.44$; $p = 0.67$). However, as all the within-host mutations were nonsynonymous, this implied that selection may act on *pde1* during the course of invasive disease. I then computationally predicted the effect of the 19 mutations observed to occur in *pde1* using `SnpEff` and `PROVEAN` (Cingolani et al., 2012; Choi et al., 2012), and have plotted these along with the predicted functional domains in fig. 4.10c. Of these mutations, 14 are predicted to change protein function, without causing LoF. The mutations are not evenly distributed across the gene and are mostly clustered in the DHH family domain or just before it. While this does not allow a singular interpretation of the effect of these variants on gene function, this is consistent with selection acting on *pde1* during meningitis. This supports the conclusion of Cron et al. (2011) that *pde1* is essential for virulence, and lends credence to the idea it may be an effect component of a pneumococcal protein vaccine.

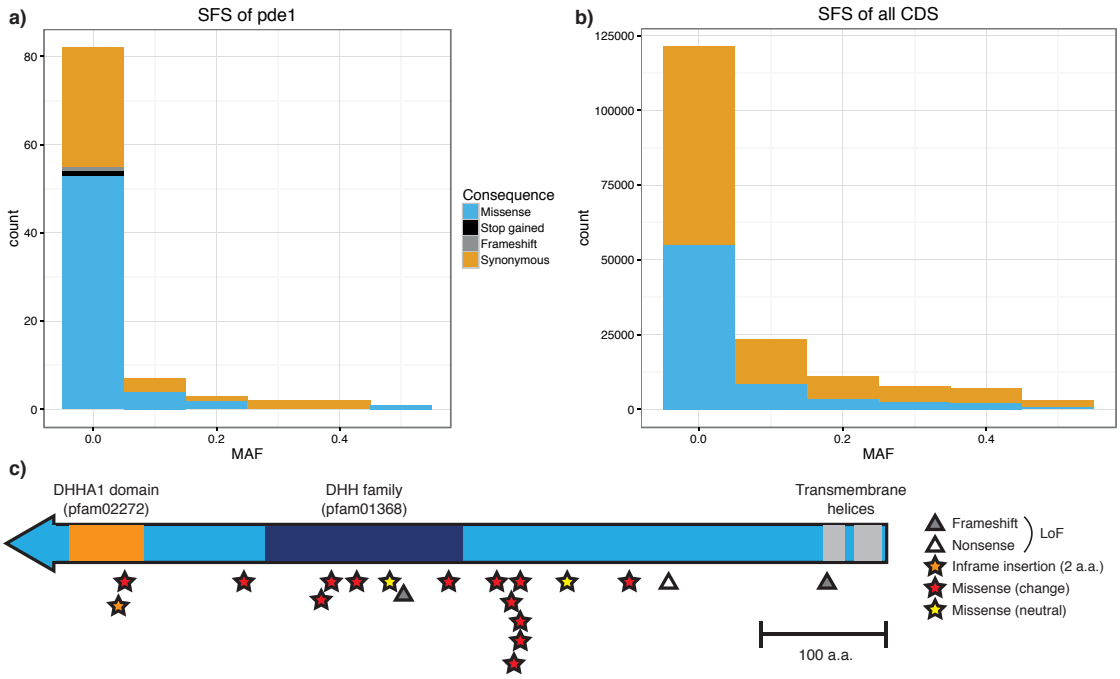


Figure 4.10: Evidence of selection on *pde1* during meningitis. Panels a and b show the SFS of mutations in just *pde1* and in all CDS, respectively. Variants are coloured according to the predicted effect. Panel c shows the positions and predicted effects of mutations observed in *pde1* during cases of meningitis and predicted pfam domains.

In all the other genes in table 4.9 the variants are non-synonymous SNPs distributed evenly between blood and CSF, therefore also showing no adaptation specific to either niche.

The most frequently mutated genes between pairs in *N. meningitidis* are shown in table 4.10. Top ranked are those relating to the pilus: *pilE* (19), *pilC* (6) and *pilQ* (4). Pilus genes are associated with immune interaction (Wörmann et al., 2014), and are therefore expected to be under diversifying selection; an excess of non-synonymous mutations (dN/dS = 1.39; p = 0.024) was consistent with this. The other notable gene with more mutations than expected in *N. meningitidis* was *porA*, encoding a variable protein which is a major determinant of immune reaction (Russell et al., 2004), in which 12 samples had frameshift mutations in one of two positions. Phase variation in the gene's promoter region, affecting its expression, is discussed in more detail below.

Gene name	Gene length (bp)	Blood/CSF mutations	p-value
pilE (NMB0018)	384	18	$< 10^{-10}$
lgtC	189	16	$< 10^{-10}$
hyaD	327	14	$< 10^{-10}$
oatA	1869	19	$< 10^{-10}$
hpuB (NMB1668)	2382	17	$< 10^{-10}$
porA (NMB1429)	1178	12	$< 10^{-10}$
lgtA (NMB1929)	1050	10	$< 10^{-10}$
kfoC	360	7	$< 10^{-10}$
cotSA	1134	7	9.2×10^{-9}
ssa1	3252	6	3.9×10^{-4}

Table 4.10: Genes containing significantly repeated mutations between blood and CSF isolate pairs in *N. meningitidis*. Ordered by increasing p-value; locus tags refer to the MC58 genome, if present.

The mutations in table 4.10 showed no association with blood or CSF specifically, so do not represent adaptation to either niche. Genetic variation in *pilE*, *hpuA*, *wbpC*, *porA* and *lgtB* within host has been observed previously in a single patient with a hypermutating *N. meningitidis* infection (Omer et al., 2011). These coding sequences overlap with those in table 4.10, which also suggests an elevated background mutation rate in these sequences, rather than strong selection between the blood and CSF niches.

Finally, I tested whether the increased mutation rate in the genes in tables 4.9 and 4.10 was associated with a particular genotype. I performed a logistic regression for each gene with over ten mutations reaching significance in the Poisson test, coding samples as one and zero based on whether they had a mutation in the gene being tested or not: no genes being mutated post invasion were associated with an MLST.

Copy number variation

I called CNVs between samples by first mapping each species to a single reference genome (ATCC 700669), then fitting the coverage of mapped reads with a mixture of Poisson distributions (Klambauer et al., 2012) as in section 4.3. Using windows of 1kb, I ranked regions by the number of sample pairs containing a discordant CNV call, as defined by the integer copy number being different between blood and CSF samples. I then inspected the top 5% of these regions.

In *S. pneumoniae* the most frequently varying region was due to poor quality mapping of a prophage region. The only other region with $p < 0.05$ was a change in copy number of 23S rRNA seen in a small number of sample pairs. In *N. meningitidis* mismapping in the *pilE*/*pilS* region accounts for the only CNV change.

4.5.3 No evidence for repeated adaptation in intergenic regions in *S. pneumoniae* and *N. meningitidis*

The previous result suggesting adaptation from blood to CSF was an intergenic change affecting the transcription of the *patAB* genes, encoding an efflux pump (Croucher, Mitchell et al., 2013). In general it is known that in pathogenic bacteria a common form of adaptation is mutation in intergenic regions, which may affect global transcription levels, causing a virulent phenotype (Gripenland et al., 2010; Johansson et al., 2002), antimicrobial resistance (Sreevatsan et al., 1997) and changing interaction with the host immune system (Magnusson et al., 2007). Changes in these regions have previously been shown to display signs of adaptation during single cases of bacterial disease (Marvig et al., 2014).

I therefore separately investigated the mutations in non-coding regions. Analysing the positions of these mutations required a consistent co-ordinate system across all sample pairs. To achieve this, I remapped the co-ordinates of each variant discovered in an intergenic region to the co-ordinates of the ATCC 700669 reference genome. I used the population matched carriage isolates as the ancestral state to determine whether these mutations occur in the blood or CSF isolate.

Figure 4.11 shows all mutations plotted genome-wide in *S. pneumoniae*. The peaks correspond to mutations in genes described in table 4.9. In the remaining 121 mutations in non-coding regions I observed no clustering by position. Over all pairs of samples, intergenic mutations were spread between blood and CSF isolates when compared to a carriage reference. This suggests none of the intergenic mutations are providing a selective advantage in either invasive niche.

The mutations in *N. meningitidis* are plotted in fig. 4.12, 110 of which were in non-coding regions. I observed enrichment (> 1 mutation), but no niche specificity, in the upstream region of six genes. These mutations are listed in table 4.11. Some of the mutations upstream of *porA* and *opc* are in phase variable homopolymeric tracts, which are discussed more fully in section 4.5.4. The other mutations are upstream of the adhesins *hsf*/NMB0992 and NMB1994, which are involved in colonisation (Hung & Christodoulides, 2013) and immune interaction during invasion (Griffiths et al., 2011), and *frpB*/NMB1988 which is a surface antigen involved in iron uptake (Delany et al., 2006). Differential expression of these genes may be an important factor affecting invasion, but the mutations I observed that may affect this do not appear to be specific to blood or CSF.

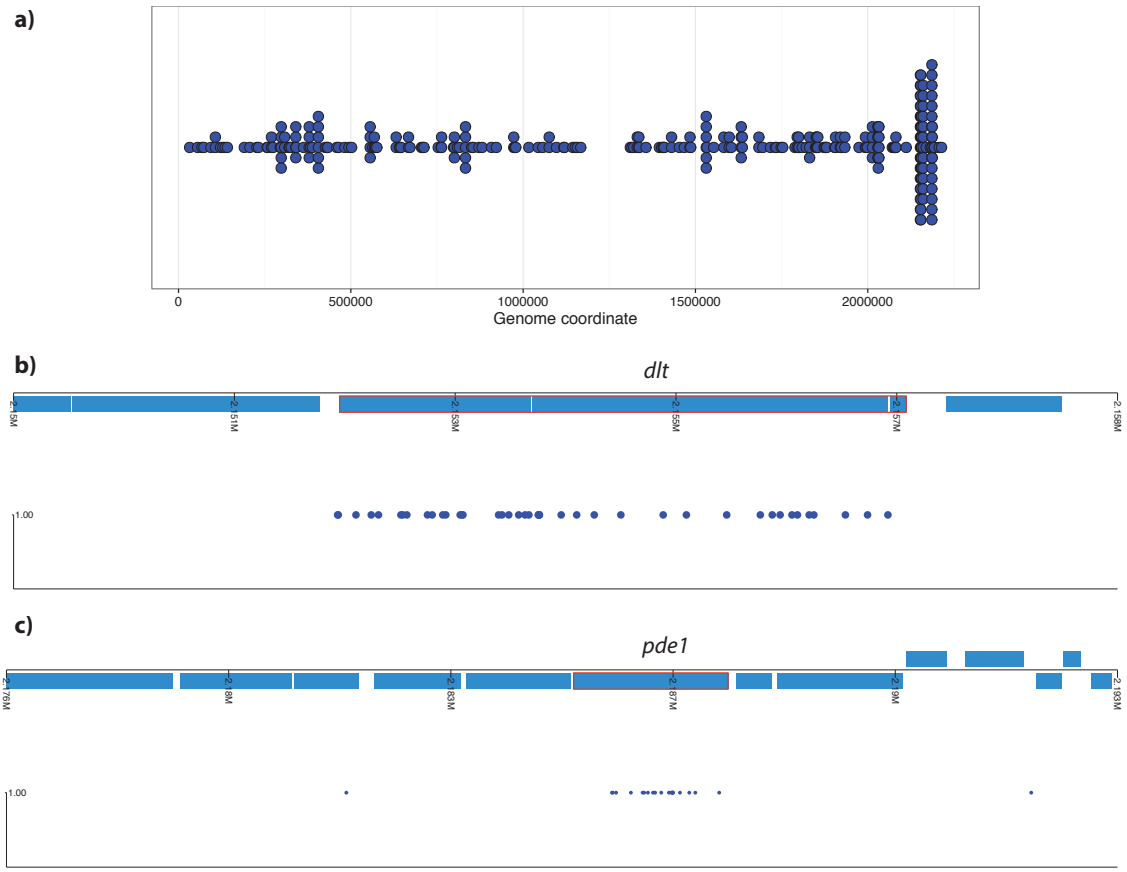


Figure 4.11: Mutations observed between all *S. pneumoniae* pairs, overlaid onto the Spn23F reference. Each blue point on the lower row corresponds to a SNP or INDEL variant observed between at least one sample pair. The blocks in the upper row represent CDSs, lying above or below the central line depending on whether they are on the forward or reverse strand respectively. The panels show **a)** whole genome (stacked, grouped by 1 000 bp windows); **b)** *dlt* operon (four genes in the centre, from 2 152 238 to 2 156 543 base pairs); **c)** *pde1* (gene in the centre from 2 185 398 to 2 187 371 base pairs).

Coordinates	Downstream gene	Blood/CSF mutations
1468329–1468331	*porA* (NMB1429)	7
1072215–1072328	*opc* (NMB1429)	7
1008872–1008985	*hsf* (NMB0992)	6
1315621–1315672	NMB1299	6
2092257–2092552	*frpB* (NMB1988)	5
2100124–2100258	NMB1994	4

Table 4.11: Intergenic regions containing significantly repeated mutations between CSF and blood isolate pairs in *N. meningitidis*. Ordered by increasing number of mutations; coordinates refer to the MC58 genome.

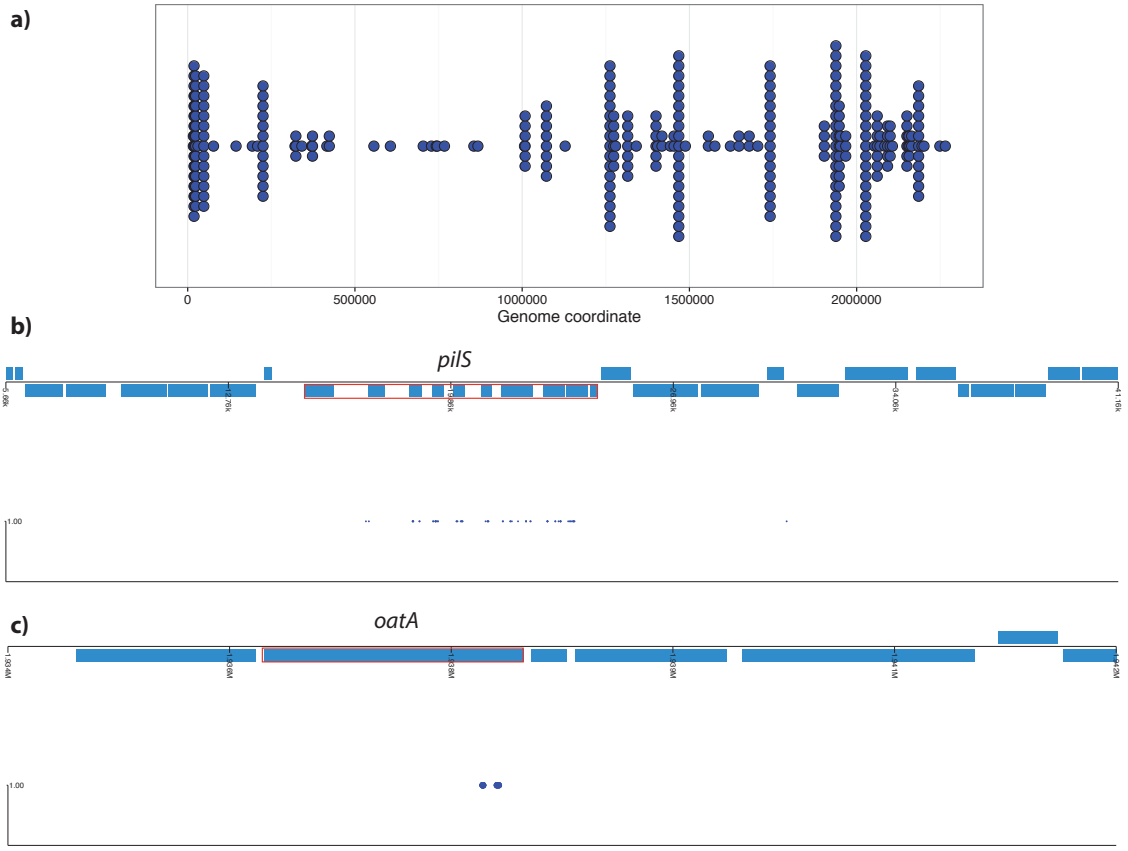


Figure 4.12: As fig. 4.11. **a)** whole genome. **b)** pilus encoding genes. Mapping to the MC58 reference places these incorrectly in the unexpressed *pilS* cassette; compared to the reference the isolates have recombined between *pilS* and the expressed *pilE*. **c)** *oatA*.

4.5.4 No evidence for repeated adaptation in phase variable regions in *S. pneumoniae* and *N. meningitidis*

Phase variable regions, which may also be intergenic, can mutate rapidly and are known to be a significant source of variation in pathogenic bacteria (Bucci et al., 1999). This mutation is an important mechanism of adaptation (Moxon et al., 1994), and meningococcal genomes in particular contain many of these elements (Snyder et al., 2001).

In *N. meningitidis* I observed six samples with single base changes in length of the phase-variable homopolymeric tract in the *porA* gene's promoter sequence, and five samples with the single base length changes in the analogous promoter sequence of *opc*. While changes in the length of these tracts will affect expression of the corresponding genes, both of which are major determinants of immune response (Sarkari et al., 1994; van der Ende et al., 2000), the tract length does not correlate with blood or CSF specifically. Consistent with this, *porA* expression has previously been found to be independent of whether isolates were taken from CSF, blood or throat (van der Ende et al., 2000).

In *S. pneumoniae* I was interested in whether the allele of the phase variable *ivr* locus discussed in section 4.3.2 was associated with either the blood or CSF niche specifically, as this could be a sign of adaptation. As the locus inversion is rapid and occurs within host,

we first ensured that cultured samples are representative of the original clinical samples using PCR quantification of each allele. We therefore extracted DNA from a subset of 53 of 674 paired clinical CSF samples and the respective bacterial isolates.

Allele prevalence was quantified using a combined nested PCR protocol based on PCR amplification of the *ivr* locus (Manso et al., 2014). Allele prevalence was identical between the original clinical sample and cultured bacteria in 50 out of the 53 samples. The predictive power of the *in vitro* detected *ivr* allele prevalence in a pneumococcal culture for the original allele prevalence within the clinical sample was therefore sufficient to draw conclusions about adaptation from.

I then used the mapping method described in section 4.3.2 to determine the allele for all the paired samples from the read data. 621 sample pairs had reads mapping to *hsdS* from which an allele can be called. However, as even a single colony contains heterogeneity at this locus, simply taking the allele with the most reads mapping to it in each sample gave a poor estimate of the overall presence of each allele in the blood and CSF niches. To take into account the mix of alleles present in each sample, and to calculate confidence intervals, I used the same hierarchical Bayesian model for the *ivr* allele used for GWAS in section 4.4.3. This simultaneously estimated the proportion of each colony pick with alleles A-F for both each individual isolate (π), and summed over all the samples in each niche (μ). I applied this over i samples and c niches (in this case c can be blood or CSF).

For each pair of blood and CSF samples the difference in allele prevalence $\pi_{\text{CSF}} - \pi_{\text{blood}}$ was calculated. All *S. pneumoniae* samples had a difference in mean of at least one allele (as the highest posterior density (HPD) overlaps zero), highlighting the speed at which this locus inverts. While this means that between a single CSF and blood pair the allele at this locus usually changes, it is the mean of μ_c (corresponding to the mean allele frequency in each niche over all sample pairs) which tells us whether selection of an allele occurs in either the blood or CSF more generally. This is plotted in fig. 4.13. As the HPD overlap, no particular allele is associated with either blood or CSF *S. pneumoniae* isolates.

Manso et al. (2014) showed in a murine invasion model that an increase in proportion of alleles A and B occurs over the course of infection. I did not observe the same effect in these clinical samples, though the large confidence intervals from the mathematical model suggest that genomic data with a small insert size relative to the size of repeats in the locus is limited in resolving changes in this allele. A small selective effect of *ivr* allele between these niches would therefore not be detected, but strong selection for a particular allele (odds ratio > 2) can be ruled out.

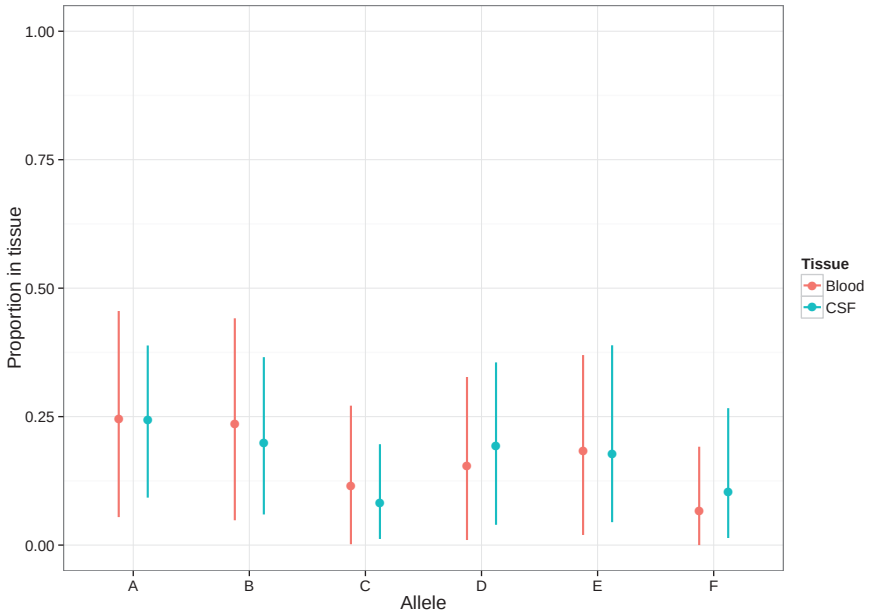


Figure 4.13: Mean and 95% HPD for μ_c. This shows the proportion of each allele present in each of blood (red) and CSF (turquoise) tissues pooling across all samples.

Diversity of *ivr* allele within samples

As the speed of inversion is rapid, I used the subsequent polymorphism of this locus to evaluate the assumptions about diversity of the bacterial population within each niche. I calculated the Shannon index of each sample's vectors μ_{CSF} and μ_{blood} to measure diversity of the sample in each niche. The mean Shannon index across CSF samples was 1.01 (95% HPD 0.39-1.51) and 0.98 (95% HPD 0.35-1.55) in the blood (fig. A.14). Looking at each sample pair individually, the difference between diversity in each niche appeared normally distributed with a mean of zero. Together, these observations suggested a similar rate of diversity generation in each niche. This is in line with the assumption that the two populations have similar mutation rates, and a similar number of generations between being founded and being sampled.

4.5.5 Carriage and invasive disease sample pairs show some evidence of repeated adaptation

Using the same methods, I also analysed pairs of genomes from 54 patients that were collected from the nasopharynx and CSF. Six of these were *S. pneumoniae*. In these samples, I detected only one sample with any variation (fig. 4.14), which was a two base insertion upstream of the *gph* gene. This is similar to the amount of mutation observed between the blood and CSF isolates, which is expected given the similar sampling timeframes. While I found that a functional *dlt* operon appears to have a deleterious effect in invasive disease, I did not observe mutation between the carriage and disease samples. However, this was expected given the small number of carriage samples relative to the effect size detected for this operon.

Between the remaining 48 *N. meningitidis* carriage and CSF isolate pairs small numbers of mutations were common. I went on to search for regions enriched for mutation, however in 8 samples I observed large numbers of mutations clustered close together (fig. 4.14). These represented single recombination events, so when analysing genes enriched for mutation I counted each recombination as a single event (Croucher, Page et al., 2015; Maiden et al., 1998).

Table 4.12 shows the results of this analysis. Similar genes are mutated as in the blood/CSF pairs, again with no specificity to either niche. In phase variable intergenic regions, I observed four sample pairs with an insertion or deletion in the *porA* promoter tract with no niche specificity. Otherwise, none of the regions above showed enrichment for mutation in either niche. These observations support the theory that these genes mutate at a higher rate but do not confer a selective advantage in any of the three niches studied.

Gene name	Gene length (bp)	Carriage/CSF mutations	p-value
lgtA (NMB1929)	1050	6	5.0×10^{-7}
oatA	1869	6	1.5×10^{-5}
hyaD	327	4	2.6×10^{-5}
pilE (NMB0018)	384	4	3.8×10^{-3}
pilT (NMB0052)	1131	4	3.5×10^{-3}
dca (NMB0415)	444	3	1.1×10^{-2}

Table 4.12: Genes containing significantly repeated mutations between nasopharyngeal and CSF isolate pairs in *N. meningitidis*. Ordered by increasing p-value; locus tags refer to the MC58 genome, if present.

A notable exception to this is the *dca* gene, a phase variable gene involved in competence in Neisseria gonorrhoea but of unknown function in N. meningitidis (Snyder et al., 2001; Snyder et al., 2003), in which all mutations are protein truncating variants in the invasive isolate. Similarly, though not reaching significance (due to the long length of the genes) were the *ggt* (NMB1057) and *czcD* (NMB1732) genes in which three recombina-

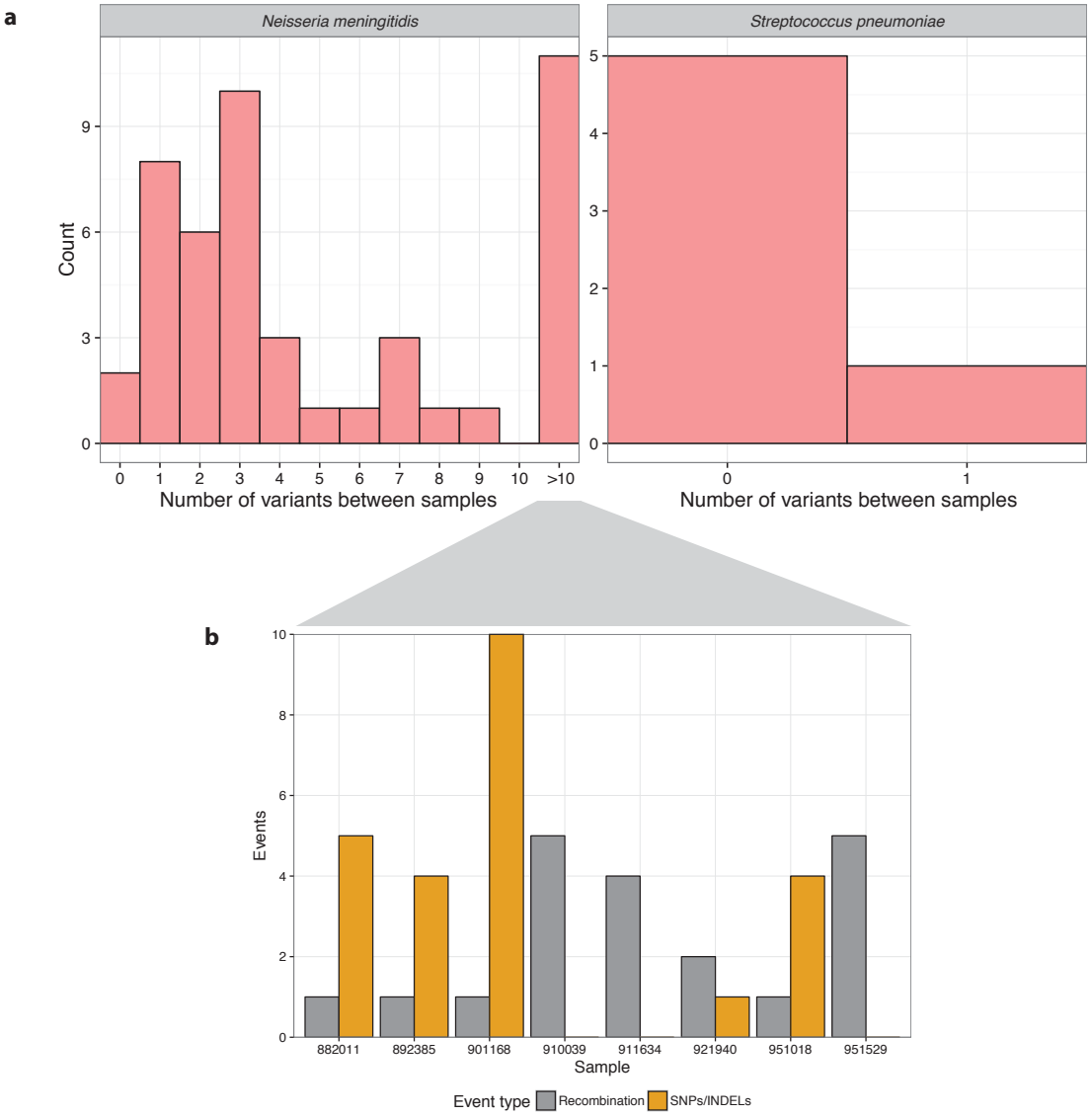


Figure 4.14: Histograms binned by number of variants between a carriage/CSF sample pair, for each bacterial species. **a)** As fig. 4.8. In *N. meningitidis* eleven samples with over ten variants between them due to recombination events are grouped. **b)** The number of recombination and SNP/INDEL events in samples in the group with over ten detected variants.

tions occurred, all of which were in the invasive isolate of the pair.

The mutations in these three genes therefore may confer a selective advantage in the invasive niche; the sequence at these loci in the invasive strains are the same as the MC58 reference, an invasive isolate itself. *ggt* has previously shown to be essential for *N. meningitidis* growth in CSF in rats (Takahashi et al., 2004), and metal exporters such as *czcD* have been shown to increase virulence in a mouse sepsis model (Veyrier et al., 2011). More such paired carriage and invasion samples would be needed to confirm if this is the case in human invasive disease.

4.6 Conclusions

In this chapter I have used a population of *S. pneumoniae* genomes to determine the contribution of naturally occurring bacterial variation to the progression of meningitis from asymptomatic carriage through blood invasion to CSF invasion. I first used a variety of bioinformatic methods to catalogue as wide a variety of variants as possible, particularly those which have previously been associated with virulence.

Using these variants and a matched collection of carriage and invasive isolates I found that the bacterial genome is crucial in determining invasive potential, with serotype likely to be the main factor. However, I did not find any evidence that the bacterial genome contributes to severity or outcome of disease. Using GWAS of both common and rare variants I found many regions and genes to be associated with invasive disease, independent of genetic background. Some of these have been previously described, whereas this is the first time others have been associated with invasive human disease. Genes involved in capsule synthesis, *yhfE*, RumA2, bacteriocins, *nanA* and *nanB* were associated with invasiveness using both common and rare variants, as well as analysis of selection.

The rare variant burden test found some well known virulence factors, showing that large effect size LoF mutations generated in lab mutants exist in the natural population, and further can affect disease in human infection. Common variants with smaller effect sizes may be the most interesting result of this approach in future, as the smaller effect sizes are harder to discover with bottom-up approaches, and their higher frequency in the population may make them more appealing vaccine targets.

I did not find evidence for association with invasiveness for some previously described variants. I did not find that the *ivr* allele was associated with invasive disease, suggesting that its function is to defend against highly variable prophage and that the variable capsule expression it can produce are not selected for in natural disease. The three antigen alleles were not associated with invasiveness, suggesting the allelic variants are a general form of diversifying selection without specific forms having a differing fitness in carriage or invasion.

These hits, as they rely on a single study population, are susceptible to batch effects specific to the Dutch setting or due to sampling bias of the collection. The association of positive controls such as capsule is reassuring, but replication in an independent population is necessary before further interpretation. The hits I have reported here will be useful for meta-analysis when further sampling and GWAS is performed.

As well as large scale population differences, previous studies have shown that substantial levels of genomic DNA sequence variation occur in bacteria colonising or infecting human hosts (Eyre et al., 2013; Kennemann et al., 2011; Morelli et al., 2010) and suggest that some of this variation may be due to selective adaptation (Croucher, Mitchell et al., 2013; Jorth et al., 2015; Marvig et al., 2014; L. Yang et al., 2011; Young et al., 2012). Such

adaptations during invasive bacterial disease could lead to new insights into the processes of pathogenesis with the potential to inform therapies (Sudip Das et al., 2016; Didelot et al., 2016), which would be difficult to assess with GWAS due to the rapid disease progression. I have searched for variation in *S. pneumoniae* and *N. meningitidis*, by comparing the pan-genomes from bacteria isolated from both blood and CSF from the same individuals in 869 bacterial meningitis cases. The genetic background within-host is the same, so this comparison could be performed without population structure correction.

I found overall that blood and CSF isolates have very similar genetic sequences. The mutations observed are not randomly distributed throughout the genome, but are instead randomly distributed between blood and CSF isolates. These mutations are therefore an observation of a higher mutation rate in these regions during invasion (for example the pilus in *N. meningitidis*, which is known to be under diversifying selection) but not repeated adaptation to either niche. This study indicates that the previous observation of variation between blood and CSF isolates from a single case of meningitis (Croucher, Mitchell et al., 2013) was a rare event most likely driven by antibiotic selection pressure during treatment. The large sample size means that this eliminates the need to search for bacterial diversity between invaded host niches (blood and CSF) when trying to explain pathogenesis of meningitis, which is a tempting analysis for reference labs with both sets of samples available. However, my comparison between the genomes of carriage and invasive isolates did show some weak signals of adaptation. I found that *dlt* appeared to be deleterious in invasion, and that selection appeared to be acting on *pde1* during invasion. These genes were not associated with invasiveness in the GWAS, which may be due to insufficient power or population stratification.

I went on to analyse 54 samples comparing carriage and invasive isolates from the same patient. Though the sample size was lower, and fully sampled diversity within the nasopharynx was not available, I was able to get an insight into potential genetic differences between bacteria in these niches. I saw some of the same genes that mutate rapidly between blood and CSF isolates also do this between carriage and invasion. This supports the conclusion that these genes have a higher mutation rate, rather than giving a selective advantage to a niche. However the power in these comparisons was limited by sample size and single colony sequencing, so comparison with GWAS results is not possible.

In the next chapter I will perform a similar analysis on the effect of host genetics on bacterial meningitis, starting with the proportion of variability attributable to common host genetic variation for invasiveness and disease severity. Together, this will give an overall picture of host and pathogen genetics affecting pneumococcal meningitis.

Chapter 5

Human genetics contributing to invasive pneumococcal disease

Declaration of contributions

Jeff Barrett, Diederik van de Beek and Stephen Bentley supervised this work. Data collection: Diederik's group designed and ran the Dutch MeninGene study; Thomas Benfield designed and ran the Danish study; Matthjis Brouwer and Bart Ferwerda compiled clinical metadata for the Dutch cohort; Lars Henrik Ängquist, Thorkild Ingvor Arrild Sørensen and Ellen Aagaard Nøhr provided quality controlled genotype data for the GOYA study; Alexander Mentzer and Julian Knight provided summary statistics for the GenOSept study; Chao Tian and David Hinds provided summary statistics for the 23andme data. Philip Kremer performed re-impuation of the *CFH* region. I performed all other analyses.

5.1 Introduction

The previous chapter has considered variation present within the pneumococcal genome that is associated with colonisation and invasive disease, while mostly treating the infected hosts as identical, with the exception of section 3.6 where I showed infant age and previous colonisation were both associated with carriage duration. However, the hosts are in reality heterogeneous: as epidemiological parameters such as contact network (Dagan et al., 2002; P. C. Hill et al., 2010), vaccination status (Klugman, 2001), co-infections (McCullers, 2006; Siegel et al., 2014; Cohen et al., 2013) host age and immune response (Cobey & Lipsitch, 2012) have all been shown to affect invasive pneumococcal disease.

However, as well as varying in these ways, humans differ in the sequence content of their genomes. The contribution of human genetics to adult pneumococcal meningitis is presently unknown – both whether it affects the disease at all, and if so which specific regions of the genome contribute to the effect. Twin studies (Jepson, 1998; Burgner et al., 2006), linkage studies (Abel & Dessein, 1997) and then GWAS studies have all suggested a role for human variation for many bacterial diseases (Chapman & Hill, 2012). Association of HLA allele as well as other regions have been found. Despite likely being selected against over human history, variants pre-disposing to bacterial diseases as stable and enduring as tuberculosis have been found (Curtis et al., 2015; Sveinbjornsson et al., 2016).

I start this chapter by using genotype data from the MeninGene (section 1.1.4) cohort to calculate the heritability of susceptibility to and severity of meningitis (section 5.2). After I found that human genetics is expected to explain the variation in these traits, I performed a GWAS for each trait to find specific regions of the genome associated with bacterial meningitis and its progression. To obtain more evidence for the associations, and increase power, I then performed the same analysis in two additional cohorts, and finally meta-analysed the results of all of the studies with a further two previous cohorts for which we obtained summary statistics.

In section 5.3 I bring host and pathogen genetics together by performing a genome to genome analysis, using cases of pneumococcal meningitis from the MeninGene cohort where both the pathogen genome and corresponding host genotype was available. Rather than looking for human variants which affect meningitis susceptibility and severity regardless of the bacterial variation, this section attempts to find specific bacterial variation which correlates with specific host variation to contribute to disease. This can be considered an interaction, between the genomes. As interactions between host and pathogen proteins are known to be important in pathogenesis (Lambris et al., 2008; Serruto et al., 2010), this is a plausible avenue to explore and may further determine the genetic architecture contributing to infection in clinical cases of disease.

5.2 GWAS of human variation associated with meningitis

The MeninGene collection was built up in three batches over the course of this work: the final numbers along with each phenotype are shown in table 5.1. As the collection includes all consenting adults with culture-proven meningitis, all causative pathogen species are included in the collection. My analysis so far has mostly been restricted to pneumococcal meningitis, as being the most common cause of meningitis in adults it is the most well powered. However in this chapter I will also consider meningitis as a whole, which also includes cases caused by *N. meningitidis*, *L. monocytogenes* and *H. influenzae*. As well as microbiological data, clinical information has been collected for most cases, allowing an association of disease severity as in section 4.4. For the association I used genotype data from the ALS (van Es et al., 2009) and B-PROOF (van Wijngaarden et al., 2011) as population matched controls, all of whom were adults.

Cohort	Country	Age	Data	Samples	Phenotype
MeninGene	Netherlands	Adults	Illumina Omni array	1 149	Meningitis
				732	Pneumococcal meningitis
				277	Unfavourable outcome
ALS & BPROOF	Netherlands	Adults	Illumina Omni array	4 836	Controls
Benfield	Denmark	Children	Illumina Omni array	353	Pneumococcal meningitis
				873	Pneumococcal bacteremia
				473	Controls
GOYA	Denmark	Young adults	Illumina quad array	2 805	Controls
23andme	European	All	Summary statistics	842	Bacterial meningitis
				82 778	Controls
GenOSept	European	Adults	Summary statistics	220	Pneumococcal bacteremia
WTCCC	UK	Adults	Summary statistics	2 244	Controls

Table 5.1: Summary of cohorts with available human genotype data. The first section shows cohorts with full genotype data where I performed a GWAS; the second section is cohorts with the summary statistics from an existing GWAS used in meta-analysis only. Sample numbers are after the QC in section 5.2.1.

I also used data from Danish children with invasive pneumococcal disease (referred to here as the Benfield cohort). Using archived blood spots in the Danish national biobank, we extracted DNA for genotyping from cases of children with pneumococcal meningitis and bacteremia, as well as 473 population controls. As additional population matched controls I obtained the genotypes of controls from the GOYA study, which randomly sampled 2 805 healthy Danish young adults (Paternoster et al., 2011).

Finally, summary statistics were available from two existing studies. The first, performed by 23andme, gave participants a questionnaire on infectious diseases. Those responding yes to the question 'Have you ever had bacterial meningitis?' were classified as cases, and those responding no as controls ('I'm not sure' was also an option, and these responders were excluded from further analysis). The analysts performed a logistic

regression at all imputed SNPs using age, sex and the first four principal components as covariates (Tian et al., 2016). The second is the unpublished GenOSept study which included 220 adults with sepsis, who suffered shock in intensive care unit (ICU) and were either blood culture positive from pneumococcus, or were positive from pneumococcal antigen in their urine. The analysts used controls from WTCCC (Burton et al., 2007) and performed a regression at all imputed sites using a linear mixed model as implemented in `gemma` (Zhou & Stephens, 2012).

5.2.1 Genetic data processing

In this section I describe the set of steps I took to prepare genotyping intensity data for GWAS analysis. From the Dutch cohort there were initially 905 cases available from the collection since the Meningene study began, with a second batch of 94 new cases covering a subsequent winter, and a final third batch of 178 new cases covering a subsequent two winters. As controls, 1 981 samples from the ALS study, and 2 898 from the B-PROOF study were available from the start. From the Danish collections, 373 meningitis cases and 475 controls were available as called genotypes, and we genotyped 904 additional samples with pneumococcal bacteremia. I also applied for access to 2 817 samples from the GOYA study, which I received as quality controlled genotype calls.

The following analysis was completely repeated four times to arrive at the final SNP calls used in the association study. The processing steps and cut-offs used were the mostly same for all of these genotyping runs, however I do point out where steps differed based on cohort or run, and where cohorts or runs have been merged. Throughout, I have used a combination of plink v1.9 (Purcell et al., 2007; Chang et al., 2015) and my own perl scripts (https://github.com/johnlees/bioinformatics) to convert between different data formats.

Genotype calling

Genotyping arrays have hundreds of thousands of SNP probes, allowing for a relatively cheap assay of all common ($>$ 5% MAF) positions in the human genome. For each variant, there is a red florescently tagged probe which binds to the A allele, and a green probe which binds to the B allele. By comparing the relative intensities of these two colours across a large number of samples a genotype probability can be assigned to each sample in the run.

We processed raw genotyping data using Illumina's Beeline software to produce normalised intensity files. In these files, for each sample an x and y intensity is recorded at every SNP typed by the array, proportional to the amount of the A and B allele present. In the ideal case a sample homozygous for A would have high x and low to no y intensity, whereas a sample homozygous for B would have the opposite. Heterozygous samples would have half of each intensity. In practice the intensities are distributions (fig. 5.2),

and the best way to produce genotype calls is to plot x against y for many samples, and find three discrete clusters. As a final complication, at any given site some samples will act anomalously and either fail to produce an intensity, or worse produce an extra cluster which confounds the identification of the real genotype clusters. Such samples should be assigned a missing genotype at these sites.

As high quality genotyping is important for downstream imputation and any eventual fine-mapping (Spain & Barrett, 2015), I used optiCall (Shah et al., 2012) to deal with these issues and produce genotype calls for all the samples with genotype intensity data. This method has been shown to throw away fewer correctly typed variants than other methods, and produce more accurate calls overall. The algorithm first samples random intensities from across the genotyping run to generate priors of where the three genotype clusters are centred, then for each variant uses an EM algorithm to adjust class membership based on these priors and the observed data.

I ran optiCall using default settings on a per chromosome basis separately for each genotyping run, using the sample sex as a covariate. In the second and third rounds of Dutch case samples, each batch contained fewer than 200 samples. So at the rarer end of the SFS, less than one sample is expected to be in the homozygous rare category. While optiCall is robust to missing classes in rare variation, it needs reliable prior information to do so. To ensure high quality calling of these runs I therefore:

1. Combined the meningitis samples with intensity data from a run of 41 samples from a European population on the same platform, used by another study.
2. Treated the run ID as a covariate in optiCall.
3. Used chromosome 1 to generate priors for all other chromosomes, as it contains the most number of variants.

After calling, I discarded the samples from the other study. I will cover direct assessment of genotype call quality in section 5.2.1.

Quality control of genotype data

When performing QC of the called genotype data I followed the advice of C. A. Anderson et al. (2010), though using more modern and faster algorithms where appropriate. I first merged the first two runs of Dutch cases and controls, giving five sample sets to QC (Dutch combined, Dutch case batch three, Danish meningitis combined, Danish bacteremia and Danish controls).

For all these datasets, I performed the following basic QC steps using `plink`:

1. Predict sample sex using genotypic data (heterozygosity rate on X chromosome). Where discordant with recorded phenotypic sex, or the phenotypic sex was missing, I replaced it with the predicted value.

2. Remove samples with an overall heterozygosity rate above three standard deviations from the mean.

3. Remove samples with $> 3\%$ of genotypes missing.

4. Remove markers with $> 5\%$ of genotypes missing.

5. Remove markers with a significantly different call rate between cases and controls ($p < 10^{-5}$).

6. Remove markers with MAF $< 1\%$.

7. Remove markers out of Hardy-Weinberg equilibrium (HWE) ($p < 10^{-5}$).

Failing samples were removed before failing markers, to maximise the number of markers retained. Steps 2–5 remove those samples and markers which have not been genotyped well on the array, whereas step 6 removes those markers with insufficient power to inform imputation or association. Step 7 is useful in discarding genotype failures as almost all markers are close to being in HWE, so the number of samples in each genotype group can be related to the MAF. Departures from HWE are mainly due to genotyping failures, where clusters have been incorrectly merged or labelled. However, while a good first step, this step is not sufficient to remove all genotyping failures.

I then estimated sample ancestry and relatedness within each collection. To estimate degree of relatedness between samples I used KING with default settings (Manichaikul et al., 2010). For ancestry, I first removed palindromic SNPs (A/T or C/G) to minimise strand issues, and merged with the genotype data with 270 individuals from four different ancestries released as phase II of the HapMap project (International HapMap Consortium, 2005; International HapMap Consortium et al., 2007). I then used `eigenstrat` to perform PCA on the merge of samples and hapmap to identify and control for ancestry (A. L. Price et al., 2006).

I did not immediately discard these samples as they can still be included in a linear mixed model to increase discovery power (Lippert et al., 2011; Zhou & Stephens, 2012). Instead, I only removed identical samples, and recorded those which were related as third-degree or closer, and samples of non-European ancestry (PC1 < 0.07 in the hapmap projection; fig. 5.1). These were only removed in downstream analyses requiring unrelated samples from the same population.

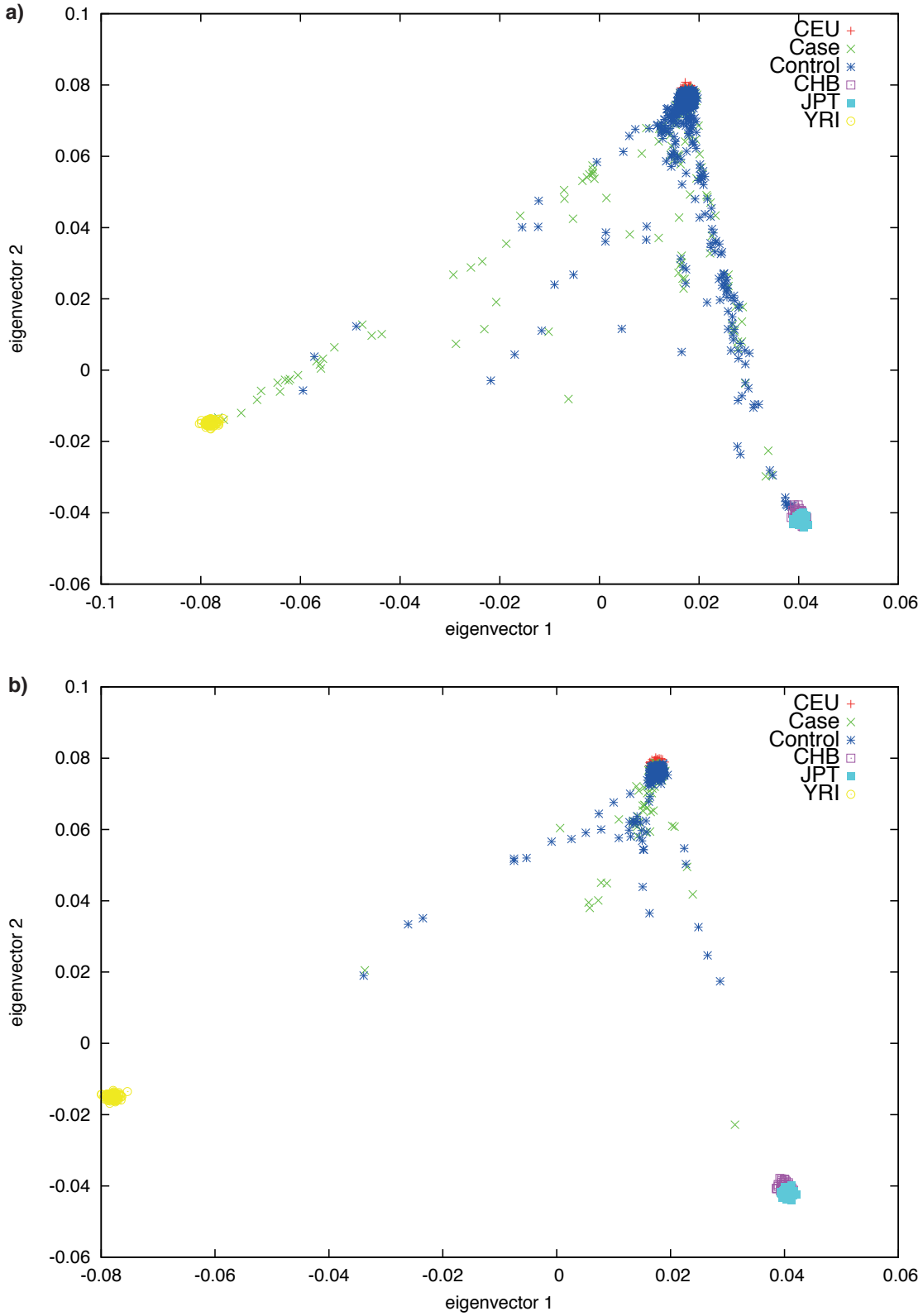


Figure 5.1: Projection of samples onto first two principal components of case (green crosses) and control (blue stars) samples from **a)** the Netherlands and **b)** Denmark with HapMap phase I populations. HapMap populations are 3 (red crosses) – CEU, European; 4 (pink squares) – CHB, Han Chinese; 5 (turquoise squares) – JPT, Japanese; 6 (yellow squares) – YRI, Yoruba Nigerians.

Using this first pass of QC, I performed an initial association test at all passing sites using a logistic regression. I removed all population divergent and third-degree or closer related samples, and fitted the basic model

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{I}-\boldsymbol{y}}\right) = \boldsymbol{X}\boldsymbol{\beta} \tag{5.1}$$

at every marker, where $\boldsymbol{y}$ is the vector of binary phenotypes, $\boldsymbol{X}$ is the additive model matrix of genotypes (0 for homozygous common; 1 for heterozygous; 2 for homozygous rare) and β is the fitted slope. Using the Wald test p-values I found 226 sites suggestively associated with the susceptible phenotype $p < 10^{-4}$, and manually inspected the genotype cluster plots using Evoker (https://sourceforge.net/projects/evoker/files/). Many of these plots were miscalled in one or more cohorts, though in such a way that the HWE p-value managed to pass the filter set earlier. Some examples of faulty calling are shown in fig. 5.2 – all such identified variants were removed prior to downstream analysis and imputation. In addition, I performed an association within the control group, using the ALS study as cases and the B-PROOF study as controls. As there should be no overall phenotypic difference between these cohorts any significant results are likely artefacts from genotyping batch or incorrect calling (Burton et al., 2007). I therefore removed all markers with $p < 5 \times 10^{-8}$.

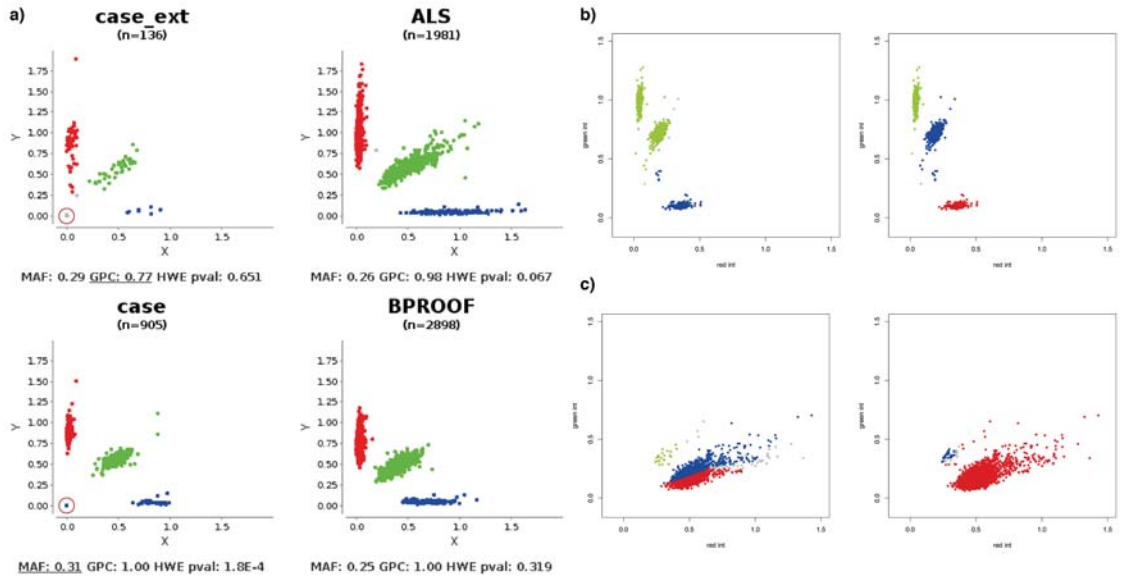


Figure 5.2: Examples of manual quality control of genotype cluster plots. All were removed rather than recalled. **a)** Evoker view of rs9516252. In cases missing genotypes have been mistakenly called as homozygous rare, whereas in cases-ext they were correct (red circles). **b)** A common mode of failure when cluster centres are not near the average. Left: incorrect identification of only two clusters at rs2717808. Right: corrected identification of three clusters. **c)** A common mode of failure when there are only two clusters at low MAF. Left: incorrect split into three clusters at rs17876189. Right: corrected identification of two clusters.

Imputation of untyped variants

To increase the power of GWAS it is common practice to impute the allele of untyped common variants in LD with those directly typed by the array, using haplotype information from whole genome sequenced population cohorts (Stranger et al., 2011). By finding overlap between genotyped alleles and haplotypes drawn from the population at these positions while taking into account population level LD it is possible to assign a probability of each genotype at all known variable positions. This increases the number of locations at which association can be tested for, mitigating the loss of low quality markers, and giving more information around signals of association. During genotype imputation all sites in the reference panel are assigned a most likely allele. At many common sites imputation accuracy is good ($R^2 = \sim 0.9$), and accuracy can be assessed through the INFO score which assesses how much information has been added at each position over the worst case of assigning the population MAF.

Humans are diploid organisms: they inherit one copy of a chromosome from their mother and the other from their father. However, as imputation works with haplotypes, a linear sequence along a single inherited chromosome, input genotypes must first be 'phased' into haplotypes. Phased data ensures that heterozygous SNPs are assigned to the chromosome they came from: for example if two alleles A/B were called as heterozygous and were next to each other possible haplotypes would be AA + BB or AB + BA (fig. 5.3). Data can be directly phased by barcoding which DNA molecules are being sequenced (Borgström et al., 2015), or by sequencing the sample's ancestors (mother and father). With genotype arrays used for GWAS direct phasing is not possible, but phased population reference panel data can be used to statistically estimate the most likely haplotypes of the input data (Delaneau et al., 2013; Loh et al., 2016).

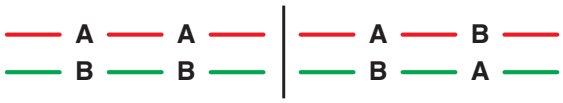


Figure 5.3: Demonstration of the effect of phasing. The subject is heterozygous for an A/B allele at two positions. The left panel shows one possibility, where the maternally inherited haplotype (red chromosome) is AA and the paternally inherited haplotype (green chromosome) is BB. The right panel shows the other possibility, of AB and BA haplotypes. Though there are another two possibilities gained from switching the parents, phasing does not distinguish these.

I performed phasing and imputation of variants using two methods. The first method, which I performed with the first batch of Dutch cases and controls, used the software `shapeit2` (Delaneau et al., 2013; O'Connell et al., 2014) and `impute2` (B. N. Howie et al., 2009; B. Howie et al., 2011) directly. I first merged the data, working with a file per chromosome across all case and control samples, then performed phasing with `shapeit2`. It is common to use the 1000 Genomes Project as the reference panel, as it contains a large collection of diverse haplotypes (1000 Genomes Project Consortium et al., 2015). It has

been shown that using a population specific reference panel can further increase imputation accuracy due to better matching, longer haplotypes being present between the reference panel and genotyped subjects (The Genome of the Netherlands Consortium, 2014). I therefore used `impute2` in reference panel merging mode, using both 1000 Genomes phase 3 (5 008 haplotypes) and The Genome of the Netherlands (GoNL) (998 haplotypes) as references to try and attain the best possible imputation accuracy for Dutch samples. I wrote a pipeline to automatically perform the imputation over a cluster system using this method by working in parallel on chunks of 2.5Mb at a time with a 250kb buffer to avoid loss of accuracy at the ends of each chunk, and automatically resubmitting failed jobs with more memory or wall-time as appropriate.

As more data became available later through the project, more efficient methods and sophisticated interfaces to phasing and imputation became available. Faster phasing became possible with `eagle2` (Loh et al., 2016) and faster imputation with `PBWT` (Durbin, 2014). This allowed the collection and use of the much larger and more diverse reference panel the haplotype reference consortium (HRC) (McCarthy et al., 2016). Though imputation accuracy is slightly lower than impute2, the efficient data structure and matching algorithm within `PBWT` allows rapid imputation even with the 63 000 haplotypes in release 1.1 of the HRC. The larger reference panel size overall gives good imputation accuracy, and includes both reference panels used in my previous imputation iteration. I therefore re-ran the phasing and imputation using this procedure, through the Sanger imputation server (https://imputation.sanger.ac.uk). Sex chromosomes were not included in this release, so all downstream analysis is of autosomes only.

To homogenise samples before imputation I used the HRC strand checking tool (http://www.well.ox.ac.uk/~wrayner/tools/#Checking). For each sample cohort, this checked whether alleles, strand of genotyping (which should all be on the positive strand, rather than the Illumina TOP strand), reference allele and MAF match with the reference panel. SNPs with MAF > 0.2 different from the reference panel are removed, which may assist with missed strand flips. I merged all samples with the same array version together (table 5.1) and then performed phasing and imputation.

Using the imputed data, I performed a final QC check on all the markers from the reference panel to remove low confidence sites. I re-applied the filters of MAF $> 1\%$ and HWE $p < 10^{-5}$, as well as removing any sites with an INFO score < 0.7 (suggesting poor imputation accuracy). After this step, 6.8M good quality SNPs were left for association. For phenotypes with lower numbers of cases (unfavourable outcome, genome to genome analysis) I applied a stricter MAF filter of $> 2\%$.

An initial association using eq. (5.1) revealed two quality issues not identified by the filters described. In both cases the issue was manifested by many highly significant p-values of markers, and non-significant values of those nearby and in LD with the lead variant. The first was a failure to match the strand between cases and controls, and in some

cases the imputation reference panel, at palindromic SNPs. At non-palindromic SNPs the reference strand is unambiguous and was correctly assigned by the strand checking tool, but at 1 722 (around 0.3% of genotyped positions) A/T or C/G SNPs with MAF $> 30\%$ neither allele or frequency mismatch could be used to disambiguate the genotype value. I used the Illumina genotype manifests data to ensure all genotypes were with respect to the positive reference strand rather than the Illumina TOP strand, and re-ran the imputation and subsequent QC on all affected cohorts.

The second issue was due to a mismatch of array design between cases and controls for the Danish bacteremia samples and GOYA controls. Despite performing separate QC and imputation of these cohorts to arrive at the same set of genotyped markers, a simple merge led to spurious association results. Although the imputation model in theory should allow for imputed sites to be merged when produced from different sets of calls, in practise subtle differences in genotyping quality and marker density for a large number of samples can easily lead to systematic differences between cases and controls. To match these two cohorts without introducing technical differences between them, I took the intersection of SNPs between the two panels and merged the genotype calls, then performed identical QC steps on the dataset as a whole. As this left only 291 830 markers ($\sim 50\%$ of that on a single array) I used `minimac3` via the Michigan imputation server (Sayantan Das et al., 2016) to perform imputation to the HRC, as this algorithm coped with the relative sparsity of markers better than `PBWT`.

Finally, as the *CFH* region was of particular interest given its previously reported association with meningococcal meningitis, we reimputed it for all the Dutch samples using `impute2`. In this mode we allowed `impute2` to infer the phasing during its MCMC which is far slower, but more accurate over this small region. This imputed data was used for meningococcal meningitis associations not reported here, and for the specific association with antigens in section 5.3.3.

5.2.2 Association results

Using the quality controlled genotype data I was able to perform three analyses on each cohort. The first was an estimation of heritability of each trait of interest, which represents the proportion of phenotypic variance explained by genetic variation. As in sections 3.3 and 4.4 I performed this calculation using different methods, as various technical limitations of each can bias estimates (Evans et al., 2017; Speed et al., 2017). All methods assume unrelated individuals with shared ancestry, so I filtered out these samples before performing heritability calculations.

I used the GCTA-GREML model, as implemented in `bolt-lmm` (Loh et al., 2015), which assumes normally distributed effect sizes with a variance equal to the genetic component of heritability σ_g^2 (J. Yang et al., 2010; J. Yang, Lee et al., 2011). Under this

assumption, restricted maximum likelihood optimisation of a LMM can be used to estimate h^2_{SNP}. This model does not adjust for LD, which in some cases may lead to underestimation of h^2_{SNP} (Speed et al., 2012). I therefore used LD-pruned SNPs as the input, and performed an additional heritability estimate with LDAK, which adjusts the weights of SNPs by their LD when calculating the kinship matrix used as the random effects in the linear mixed model.

After confirming that it is expected that a genetic contribution to the phenotype exists, I then ran an association scan. This performs a regression between variant and phenotype at every marker, though the use of an LMM allows ancestry and relatedness of samples to be included as random effects in the regression model. This means ancestrally divergent and related samples do not have to be completely removed, increasing the power to find associations without increasing type I error (A. L. Price, Zaitlen et al., 2010). It has previously been computationally prohibitive to fit this model to every imputed marker, but recent efficiency advances have allowed this technique to become commonplace (Lippert et al., 2011; Zhou & Stephens, 2012). I used `bolt-lmm` to perform the association (Loh et al., 2015), using LD-pruned genotyped markers to estimate the kinship matrix and random effects, and performing association at all genotyped and imputed sites. Where appropriate, I have included covariates such as immunocompromised status as fixed effects in the model.

The final question I wished to test using this data was whether there was evidence for difference of the genetic basis between similar sub-phenotypes of invasive disease. For example, is the association with *CFH* specific to meningococcal meningitis, or is it also shared by pneumococcal meningitis too? Overall, is there a difference in genetic susceptibility to different pathogens, or different manifestations of invasive disease? As the case numbers are low, these studies were underpowered to detect a difference through direct association of different sets of markers, or to calculate co-heritability. However, in such cases, performing an association between all cases and controls, and then between sub-phenotypes of cases may help test for an overall difference. Liley et al. (2017) have developed the subtest method which fits a mixture of Gaussians to the Z-scores from these two association tests, which compares the null model fit assuming no difference between subphenotypes and the alternative model when there is a difference. It can extract a p-value from the LRT which expresses the probability that the genetic basis for the subphenotypes are distinct. When running subtest I used the weights from LDAK to account for LD between associations, and performed 1 500 subsamples of 400 samples to generate the null-distributions of the test statistic.

Dutch cohort results

In the Meningene cohort I considered three different phenotypes: the susceptibility of adults to bacterial meningitis (using all cases), pneumococcal meningitis only, and severe (unfavourable clinical outcome) meningitis. In all of these associations I used immunocompromised status as a covariate (10% of cases) assuming that no controls were immunocompromised, as population prevalence is around 1% (van Veen et al., 2011; Harpaz et al., 2016).

The heritability analysis (table 5.2) showed that human genetic variation was expected to contribute to all of the phenotypes of interest. The size of the contribution varied, but was relatively high in comparison to other complex traits (Ge et al., 2017). In general LDAK estimated a higher heritability than GCTA-GREML, as expected from the structure of the models (Evans et al., 2017). Analysis using `subtest` as described above did not provide any evidence that pneumococcal meningitis was distinct from other bacterial meningitis (PLR = 0.25; $p = 0.75$) or that unfavourable outcome was distinct from overall meningitis susceptibility (PLR = 0.14; $p = 1.00$). However this may rely on relatively highly associated SNPs, which were not found with this few samples. Susceptibility to any meningitis has a significantly higher heritability than its sub-phenotypes, which also have heritability above zero. This is more consistent with some difference in genetic architecture between the phenotypes.

Phenotype	Method	Heritability	Error	Fit p-value
All meningitis	GCTA	0.418	0.064	2.4×10^{-6}
	LDAK	0.556	0.088	3.9×10^{-11}
Pneumococcal meningitis	GCTA	0.353	0.068	2.4×10^{-6}
	LDAK	0.416	0.096	3.9×10^{-6}
Unfavourable outcome	GCTA	0.192	0.067	2.8×10^{-5}
	LDAK	0.325	0.090	1.4×10^{-4}

Table 5.2: Human SNP heritability (h^2_{SNP}) of three meningitis phenotypes in Dutch adult cohort. Pneumococcal meningitis and unfavourable outcome are subsets from the 'all meningitis' phenotype. For each phenotype I estimated heritability using both GCTA-GREML and LDAK models, in every case there was evidence for heritability significantly above zero.

The Manhattan plots of the association results are shown in figs. 5.5 to 5.7. Across the three traits only one locus reached genome-wide significance: position 64680775 on chromosome 1, an intronic variant in *UBE2U*, was associated with unfavourable outcome (MAF = 0.43; OR = 1.62; p = 2.0×10^{-8}). *UBE2U* is part of the ubiquitin pathway (responsible for degrading proteins in the cell) (Gregory et al., 2006), but has not previously been associated with any other disease or trait. The signal also spanned *ROR1* (fig. 5.4), a protein of unknown function (Bainbridge et al., 2014) which has previously

been associated with cancers (Reddy et al., 1996) and pulmonary function (Lutz et al., 2015). Signals suggestive of significance for each trait are reported in table 5.3. Despite the lack of association from meningitis susceptibility, the heritability estimates above suggest that meta-analysis with more samples should be able to find associations with lower OR and MAF. I otherwise delay a detailed interpretation of results until they are replicated in an independent study and reach genome-wide significance in section 5.2.3.

Phenotype	Position	Effect allele	MAF	OR	p-value	Annotation
All meningitis	chr6:153582990	T	0.42	1.27	7.2×10^{-8}	Upstream of *RGS17*
Pneumococcal meningitis	chr6:117624549	G	0.46	0.77	8.8×10^{-7}	*ROS1* intron
	chr18:48403560	T	0.43	0.65	7.6×10^{-8}	*ME2/ELAC1/SMAD4*
	chr22:47506160	G	0.33	0.74	5.5×10^{-7}	*TBC1D22A* intron
Unfavourable meningitis	chr1:64680775	A	0.43	1.62	2.0×10^{-8}	*UBE2U/ROR1*
	chr4:182823804	A	0.33	1.58	4.1×10^{-7}	*AC108142.1* intron
	chr9:37382231	A	0.07	2.36	6.7×10^{-7}	*ZCCHC7/GRHPR*

Table 5.3: Signals of association in the Dutch cohort. I report the lead SNP at each associated locus with MAF $> 5\%$ and $p < 1 \times 10^{-6}$, and nearby annotated genes. The suggestive signal in all meningitis cases was also present when restricted to pneumococcal cases, albeit with a lower p-value of 3.9×10^{-7}.

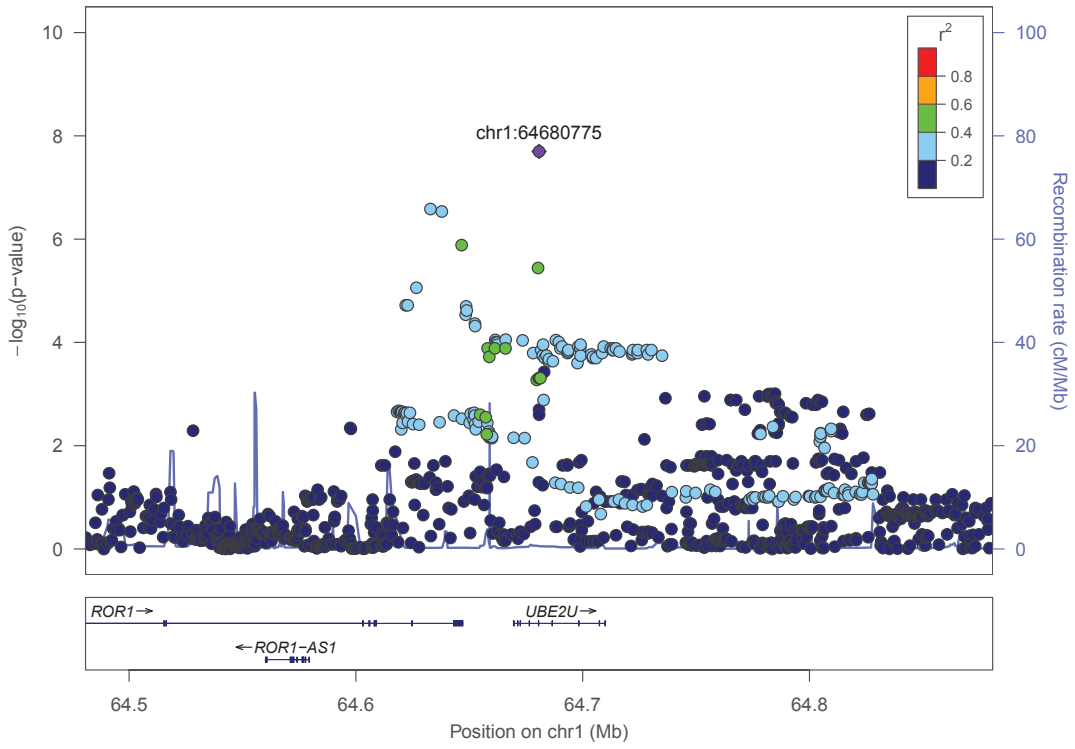


Figure 5.4: Locuszoom plot (Pruim et al., 2010) of association on chromosome 1 with unfavourable outcome, which is a zoom of the Manhattan plot on the locus. The lead SNP is a purple diamond, other markers are circles coloured by their r^2 with the lead SNP to show LD. The bottom panel shows annotated genes in the region, with exons as boxes and introns as lines. Recombination rate in cM/Mb is plotted as a pale blue line.

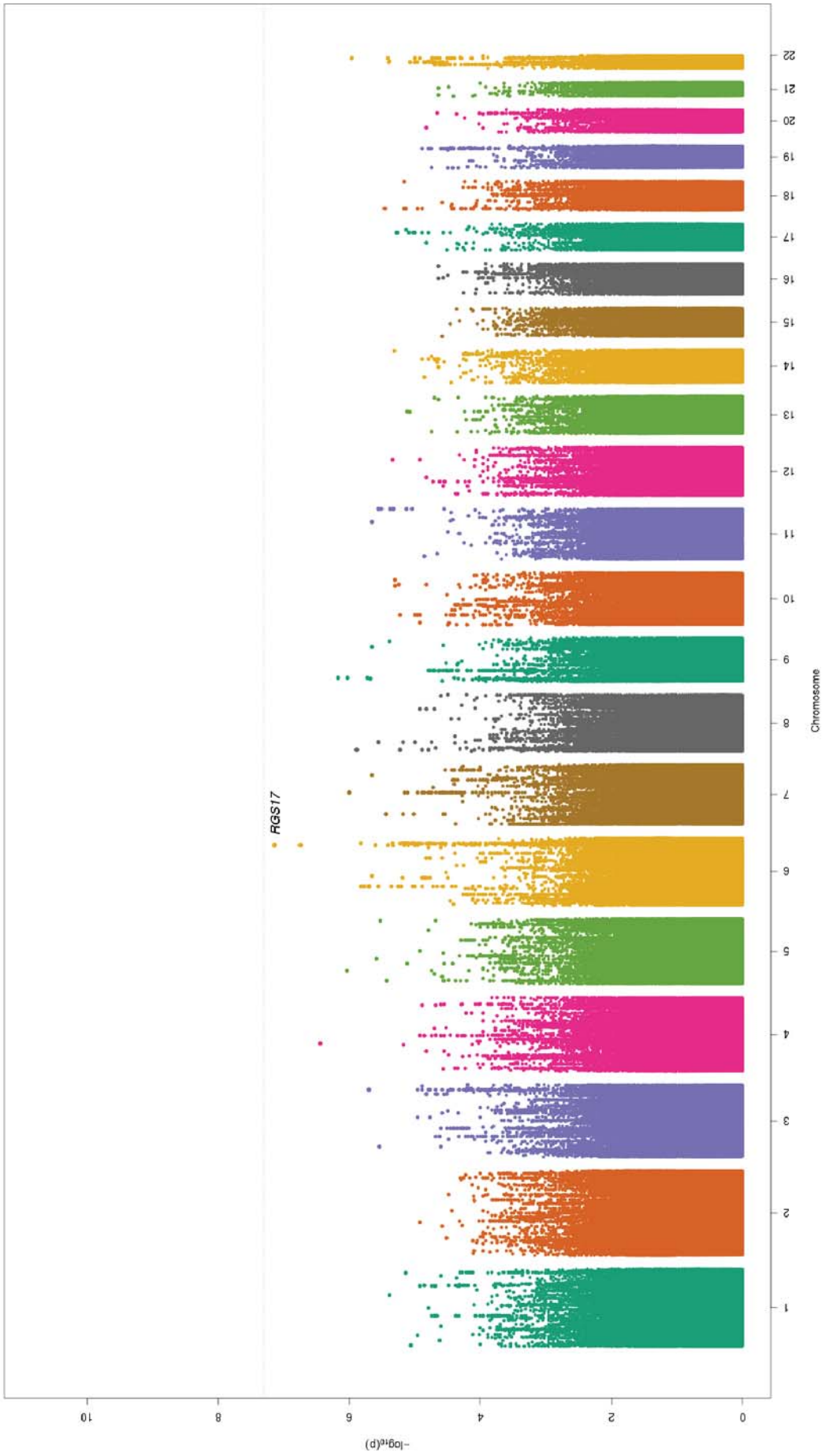


Figure 5.5: Manhattan plot from GWAS of all Dutch meningitis cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Suggestive results from table 5.3 are annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

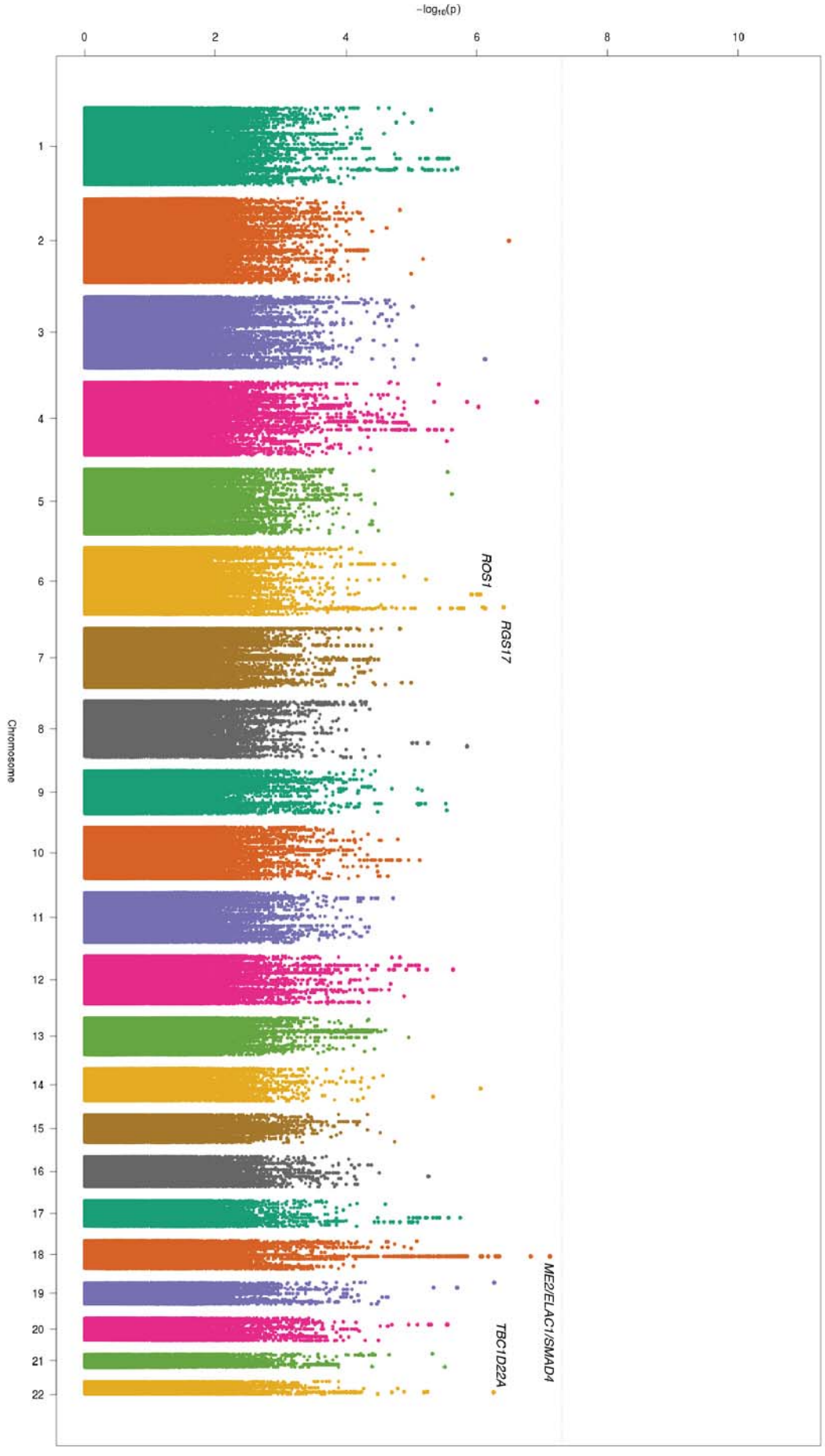


Figure 5.6: Manhattan plot from GWAS of Dutch pneumococcal meningitis cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Suggestive results from table 5.3 are annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

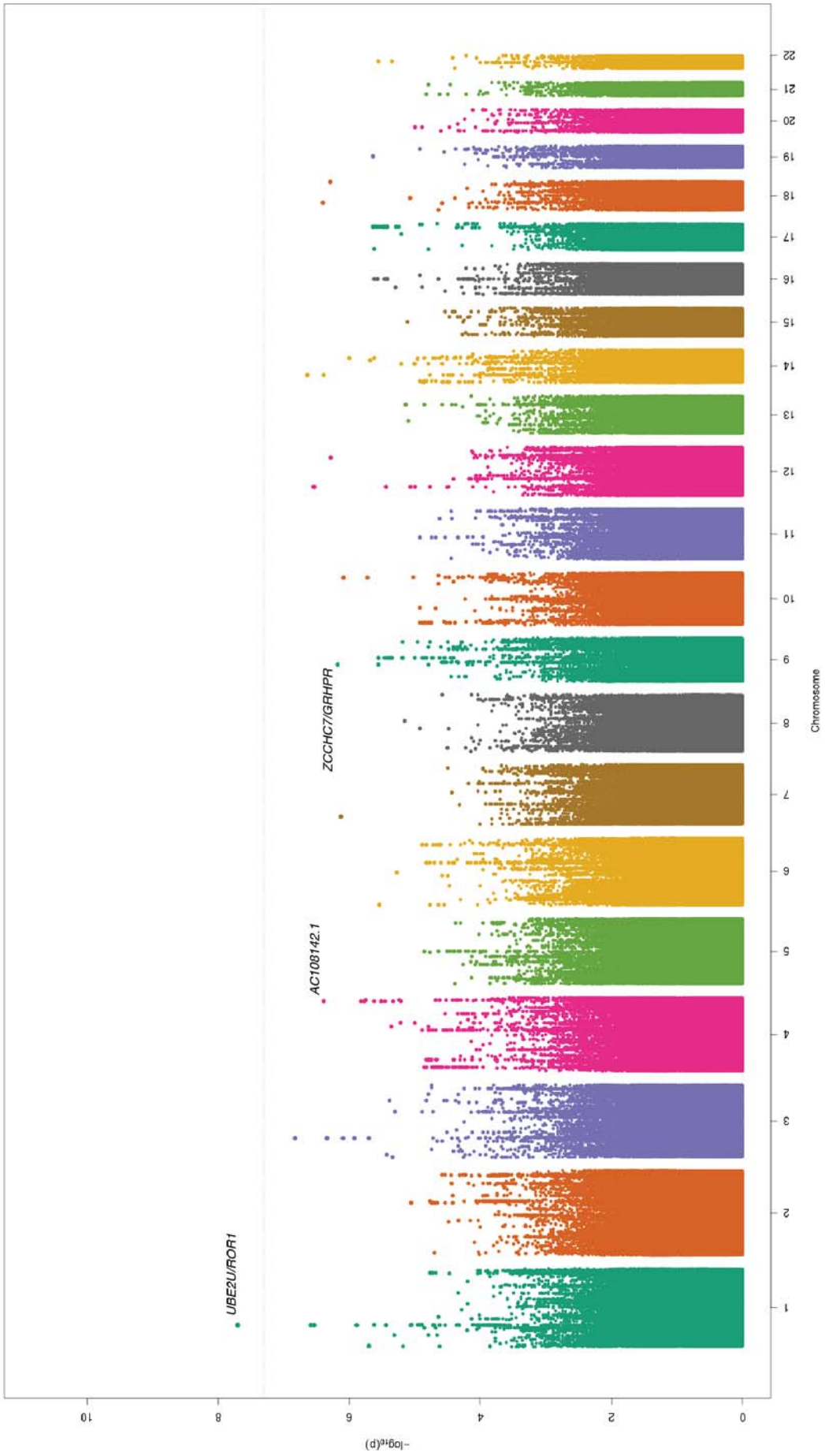


Figure 5.7: Manhattan plot from GWAS of all Dutch meningitis cases with an unfavourable outcome. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Suggestive results from table 5.3 are annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

Danish cohort results

Once again analysis of all invasive pneumococcal disease, pneumococcal meningitis and pneumococcal bacteremia suggested a heritable component to each of these phenotypes (table 5.4), with estimates consistent with the Dutch study (although with wider confidence intervals, due to the smaller number of samples). Subtype did not provide any evidence that bacteremia and meningitis are genetically distinct phenotypes (PLR = 311; $p = 0.60$), as associations between the phenotypes followed a similar profile. No genome-wide significant associations were found for either pneumococcal meningitis or pneumococcal bacteremia (figs. 5.8 and 5.9). The only suggestive association (MAF $> 5\%$ and $p < 1 \times 10^{-6}$) was found on chromosome 14 at 67181537 (MAF = 0.14; OR = 0.45; p = 2.2×10^{-7}) in an intron of *GPHN*.

Phenotype	Method	Heritability	Error	Fit p-value
Invasive pneumococcal disease	GCTA	0.259	0.081	1.3×10^{-5}
	LDAK	0.285	0.092	8.5×10^{-4}
Pneumococcal meningitis	GCTA	0.727	0.451	5.1×10^{-7}
	LDAK	0.849	0.569	7.3×10^{-2}
Pneumococcal bacteremia	GCTA	0.371	0.098	1.4×10^{-5}
	LDAK	0.575	0.113	2.1×10^{-7}

Table 5.4: Human SNP heritability (h^2_{SNP}) of three pneumococcal phenotypes in Danish children cohort, as in table 5.2. Pneumococcal meningitis and bacteremia are subsets of the overall category of invasive disease.

5.2.3 Meta-analysis of four studies

An important step in GWAS is to confirm the results using an independent study population. As well as avoiding possible batch effects from a single cohort, this also increases sample size and power at true associations with an OR/MAF too low to find in the initial study. Here I did this analysis for meningitis susceptibility, which had the most total samples available. I used the summary statistics (p-value and β) that I generated from the Dutch and Danish cohorts, as well as summary statistics I received from 23andme and GenOSept (table 5.1).

I performed the meta-analysis between these studies using METAL (Willer et al., 2010). At each site the beta values (effect sizes and direction) and p-values from each study are converted into z-scores, which are then combined as a weighted sum with the weights given by the number of samples N in each study. This combined z-score gives the meta-analysis p-value. Before doing this I made sure all marker positions and alleles were given with respect to the same reference, as the direction of effect is crucial. For the association

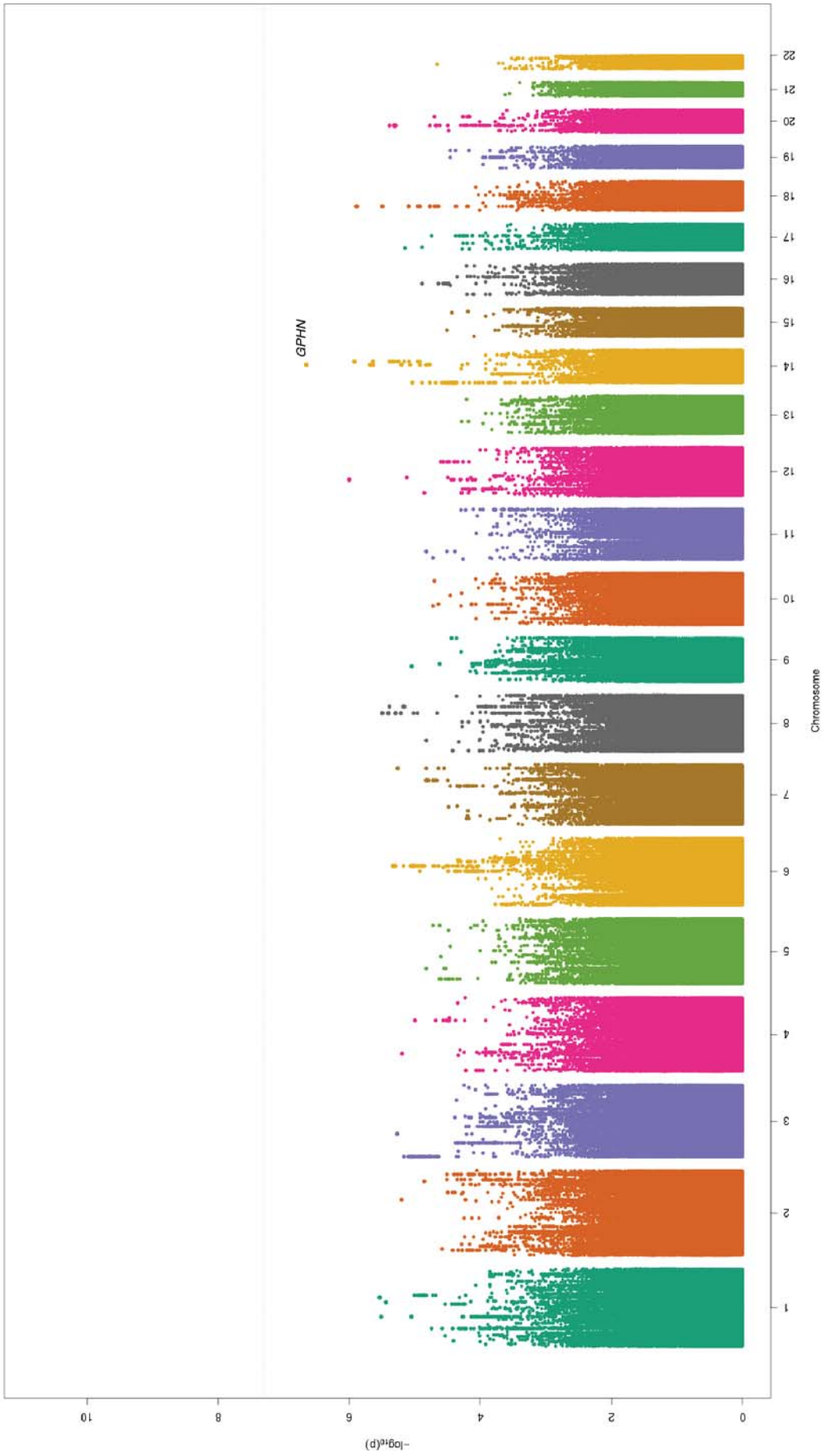


Figure 5.8: Manhattan plot from GWAS of Danish pneumococcal meningitis cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. The suggestive results is annotated with nearby genes. Genome-wide significance is at 5×10^{-8}.

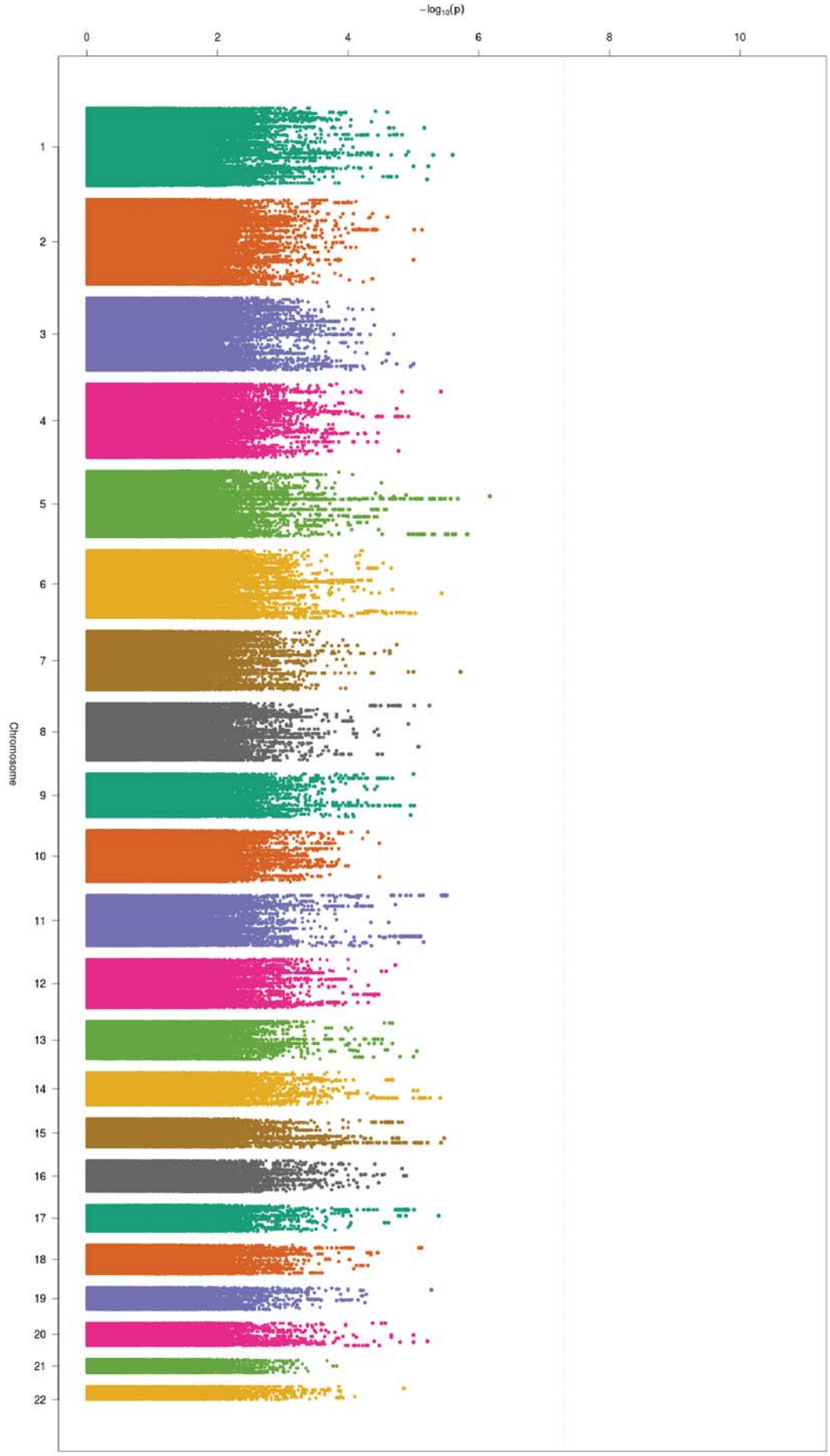


Figure 5.9: Manhattan plot from GWAS of all Danish pneumococcal bacteremia cases. Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Genome-wide significance is at 5×10^{-8}.

studies I performed using bolt-lmm I adjusted the beta values using the formula

$$\beta_{\text{adjusted}} = \beta \cdot \frac{1}{\pi \cdot (1-\pi)}$$
$$\text{where } \pi = \frac{N_{\text{cases}}}{N_{\text{cases}} + N_{\text{controls}}}$$

As the weight N for each study I used the effective sample size

$$N_{\text{eff}} = \frac{4}{\frac{1}{N_{cases}} + \frac{1}{N_{controls}}}$$

rather than the total number of samples, as some of the studies were highly biased to a larger number of controls (for example 23andme used 842 samples and 82 778 controls). I only included markers that had summary statistics from all studies in the meta-analysis (M = 5 627 710), to avoid effects of sample size heterogeneity in the final p-values.

Figure 5.10 shows the results of the meta-analysis genome-wide. No sites were significant in this analysis, and the additional data did not support the genome-wide significant hit in an intron of *CA10* reported by 23andme (Tian et al., 2016). A possible reason for these observations is due to heterogeneity of phenotype between the cohorts in the meta-analysis. The simple method used here assumes that sites must have the same direction of effect, and are independent observations of significance, and are on the same phenotype with no measurement error. However, the Dutch and Danish cohorts differ in that they analyse adult and childhood meningitis respectively, which differ in their immune system competence and their vaccination status (Imöhl et al., 2010; Rodrigo et al., 2014). GenOSept includes bacteremia cases, which may be different from meningitis specifically. Finally, 23andme uses self-reported status of bacterial meningitis. While self-reported data has generally been shown to be as good as hospital diagnoses for phenotype association, especially given the increased number of cases available, for difficult to diagnose infectious diseases such as lupus this has been shown not to be the case (Tian et al., 2016; Cortes et al., 2017). For bacterial meningitis cases have not been culture-proven, and may well be viral meningitis or not meningitis at all. If they are meningitis, most likely a wider range of pathogens compared to the other cohorts have been included.

A future analysis will include association statistics calculated from the UK biobank, which has a large collection of genotyped samples (N = 500 000) and hospital diagnoses for bacterial and pneumococcal meningitis. This may help to provide extra samples with a well-defined phenotype. Alternatively, modelling the heterogeneity in phenotype may help, though sample size is still likely to be a limiting factor.

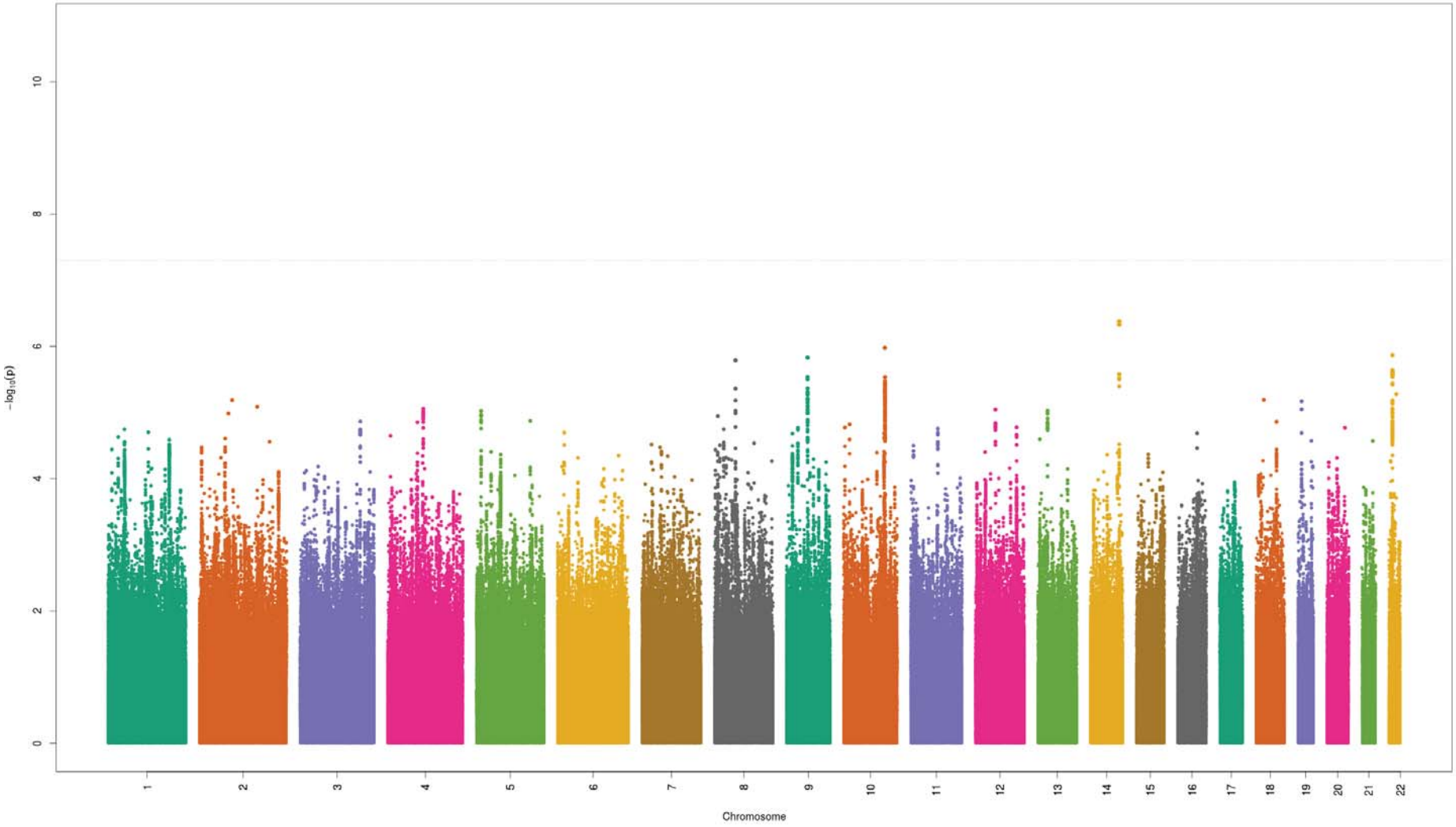


Figure 5.10: Manhattan plot from meta-analysis of meningitis susceptibility, including samples from the Dutch (adult meningitis cases) and Danish cohorts (child pneumococcal meningitis cases), 23andme (self-reported bacterial meningitis) and GenOSept (adult pneumococcal bacteremia). Y-axis is $-\log_{10}(p)$ of the association values, x-axis is ordered by marker position on each autosome. Genome-wide significance is at 5×10^{-8}.

5.3 Genome-to-genome analysis of host and pathogen variation

In this final section I aim to bring together data from chapter 4 and section 5.2 to search for genome-to-genome associations between the host and pathogen in cases of bacterial meningitis. By linking the datasets from the human and pathogen arm of the Meningene study and performing an association study between pairs of variants drawn from each genome over all these samples, I tested the hypothesis that certain bacteria are more likely to cause invasive disease in specific host genotypic backgrounds. This dataset is unique, and to the best of my knowledge the first time both host and pathogen have been sequenced for a bacterial infection. The present analysis does not require a phenotype, an advantage of such epistasis analyses (Skwark et al., 2017).

In viral infections, two previous analyses have been published attempting this analysis. Bartha et al. (2013) used host genotype and the infecting viral genome from 1 071 HIV patients to perform a logistic regression between every human SNP (of which there were ~ 7 million) and every viral amino acid (of which there were 3 000) while using the first two principal components to correct for viral population structure. The authors recapitulated the well known association with viral load and HLA allele, but were unable to find any new genome-to-genome links. They estimated having 80% power to detect a variant with MAF of 10% with an OR of 4.2 given their sample size and the number of tests being performed.

Azim Ansari et al. (2017) performed a similar analysis on 542 cases of hepatitis C infection. Again using imputed human genotypes and viral amino acids they performed a logistic regression between variants, using the first three principal components to control for human population structure, and the first ten to control for viral population structure. As well as finding expected associations with the HLA, they found a region of the viral genome associated with variability in *IFNL4*, though not quite reaching significance. However, the same human SNPs were found to be associated with viral load, for which the authors were able to conclude a link between the strength of selection acting on the viral population due to the *IFNL4* response, and the resulting fitness of circulating virions.

I wished to first remain agnostic to annotation or previous knowledge of host-pathogen interactions to attempt to uncover previously unknown genome to genome links in clinical cases of bacterial meningitis, following a similar design to the two viral studies. To do this, in section 5.3.1 I performed an association test between every genotyped human SNP and every bacterial mapped SNP/INDEL. However, given the small sample size and the large amount of variation between the two genomes, the power to overcome the multiple testing was very low for even moderate effect sizes. I therefore used unsupervised clustering techniques which use the correlation structure present in the bacterial population

to produce a lower dimensional representation of the bacterial genomic variation, lowering the multiple testing burden (section 5.3.2).

Finally I wished to test whether variation in host and pathogen protein which are well known to interact with each other is correlated in cases of disease (section 5.3.3). I used the detailed antigen calling already performed in section 4.3.1 as the bacterial variants, and tested for correlation with human variation. As these bacterial proteins are known to be broadly antigenic (Croucher et al., 2017), I tested not only the specific human gene involved in the interactions, but every imputed human variant to try and identify potential new interaction partners.

In the tests below I used the 460 samples which passed the QC filters from both sections 4.2 and 5.2.1. When performing associations on a sub-phenotype, as in splitting these samples into two based on cluster or antigen membership and testing human SNPs against this, I only tested those sub-phenotypes which contained at least 5% of samples. This avoided spurious results from testing rare (and underpowered) variants resulting from partitioning lower frequency variants into yet lower frequency phenotypes.

5.3.1 All by all variant association

To perform a correlation analysis between 7×10^6 imputed human variants and 1×10^5 requires around 10^{12} association tests, which even given the availability of a large number of CPU cores and the embarrassingly parallel nature of the problem is computationally challenging.

To approach this problem, I modified the SEER C++ code from chapter 2 to perform the association tests, as I had already optimised it for speed. I converted the VCF files with the human and pathogen variant calls to comma separated values (CSV) files, coding the human calls as 0, 1 or 2 based on the number of copies of the minor allele the genotype contained (the additive model). These CSV files then only contained the genotypes, and I stored site and sample level metadata in separate files – this separation allows much quicker processing of genotype data, especially when accessing specific chunks (Ganna et al., 2016). I extended the χ^2 test to a 3x2 table, and added efficient code for a 3x2 Fisher's exact test (https://github.com/chrchang/stats) which I applied when the assumptions of the χ^2 test were violated (by small expected values in the table counts, when MAF in either genome was low). I used a filter of $p < 5 \times 10^{-11}$ for this uncorrected test, which is equivalent to a Bonferroni correction with a significance level of $\alpha = 1$. I then tested the pairs of variants which passed this filter with a logistic regression, using the human SNP and the first three components of the bacterial MDS projection as the design matrix $\boldsymbol{X}$ and the bacterial variant as the response vector $\boldsymbol{y}$.

To parallelise the code I used 300 independent jobs. Each job first read in all the bacterial variants from the CSV file, and parsed these into a matrix stored in main memory.

The null log-likelihood for the logistic regression was calculated for each at this point, to avoid having to make this calculation multiple times when the same bacterial variant was tested against every human SNP. The chunk of human SNPs assigned to the job were then read in, and each one passing filtering was tested for association with every bacterial variant.

As the number of imputed human SNPs was still prohibitively large, I tested the genotyped human variants only. This is similar to testing an LD pruned subset of sites with the advantage that their genotype calls could be further investigated if they proved significant. Using this approach I tested 623 649 human SNPs for correlation with 113 059 associated bacterial variants (SNPs and INDELs from section 4.3). 1.8×10^{10} variant pairs passed filters of MAF $> 5\%$ in both human and bacterial population with $< 5\%$ of calls missing. Using 300 jobs the total computation time was 268 hrs, using 600Mb memory per job. 2 433 variant pairs passed the initial p-value filter for $p < 1$ when adjusting for multiple testing, but none of these were significantly associated at $p < 0.05$ when tested adjusting for bacterial population structure.

Due to the high multiple testing burden from the large number of variant pairs being tested, this number of samples would only detect strong correlations between genomic variants. This is plotted in fig. 5.11: assuming a MAF of 25% in each population, the sample size of 460 has 80% power to detect an epistatic effect with an odds ratio of 4. While bacterial population structure is less likely to be an issue for this analysis, it may still reduce the power to fine-map specific interactions. To find whether interactions exist at lower coupling strengths it would help to have more samples, as at sample sizes double this study the discovery power increases sharply. The number of samples is also currently too small to do a heritability analysis of the interaction effect.

While sample size fundamentally limits this analysis, there are some further steps to be taken. Firstly, the use of Direct Coupling Analysis has been shown to have greater power at detecting epistatic interactions in the *S. pneumoniae* genome than the simple χ^2 tests I have used (Skwark et al., 2017). However, an implementation of this which will scale to the size of the present problem does not exist. Instead, in subsequent sections I use a representation of the pathogen genome in a lower number of dimensions to attempt to reduce the multiple testing burden.

5.3.2 Reduced representation of pathogen genome

Given the difficulties encountered when testing every human variant against every bacterial variant, I wished to find a way to reduce the dimensionality of the problem. This problem is well known in eQTL studies, where both transcriptomic and genomic variation is measured, and an association is performed between the genetic variation and altered gene expression (Breitling et al., 2008; L. Franke & Jansen, 2009). One approach is to model the per-gene

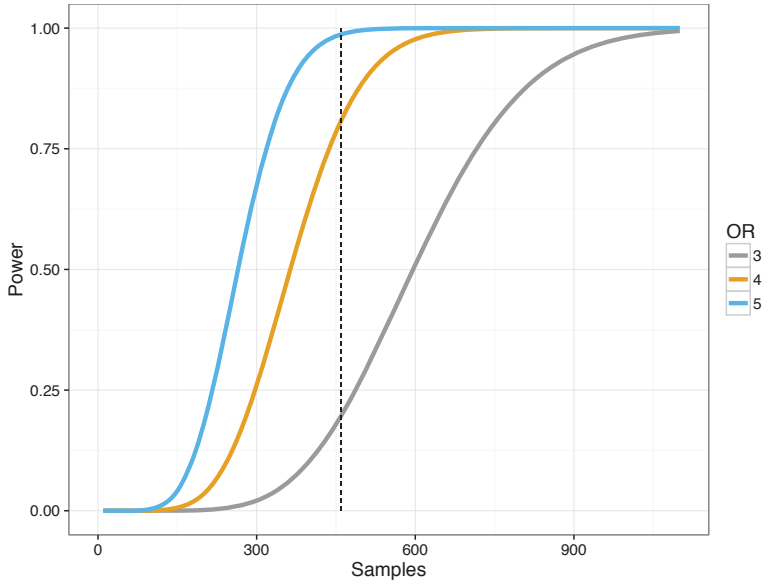


Figure 5.11: Power for detecting genome-to-genome interactions. Assuming no population structure effect, the power of detecting an correlation between genome positions at 25 % MAF at a range of ORs. The 460 samples I was able to use in this study is marked as a vertical dashed line.

levels of transcript variation as a smaller number of latent variables, each of which affect a number of transcripts. The simplest way to do this would be by PCA which would use the linear combinations of transcripts explaining most of the variation as the latent variables, though more sophisticated methods exist (Marttinen et al., 2013; Gillberg et al., 2016). In the present analogy, transcript variation corresponds to bacterial sequence variation, and the latent variables may combine these into features such as sequence type, serotype or antibiogram type.

A method which has been successfully used for this purpose is probabilistic estimation of expression residuals (PEER), which estimates latent factors and their per sample weighting from high dimensional input (Stegle et al., 2012). PEER's advantages over PCA are that: the latent factors estimated from the data do not have to be orthogonal, which may not always be biologically realistic; covariates can be included in the model fit such as batch effects or case/control status; the factors can be controlled to not be parallel with other known influences, for example serotype or sequence type.

I therefore ran PEER, learning 40 unobserved factors (though this is an unimportant setting, as automatic relevance determination is used to determine this from the data). The results are shown in fig. 5.12 – the first few factors can be seen to represent the large scale population structure, and some later factors represent finer scale population structure. I performed an association with all imputed variants against all the inferred factors, which gave uninflated results for the first twelve factors. Further factors gave spurious results at lower frequency variants.

While the PEER factors can be interpreted by the looking at the weights assigned to each input variants for the associated factor, I found this difficult to link directly to a biological

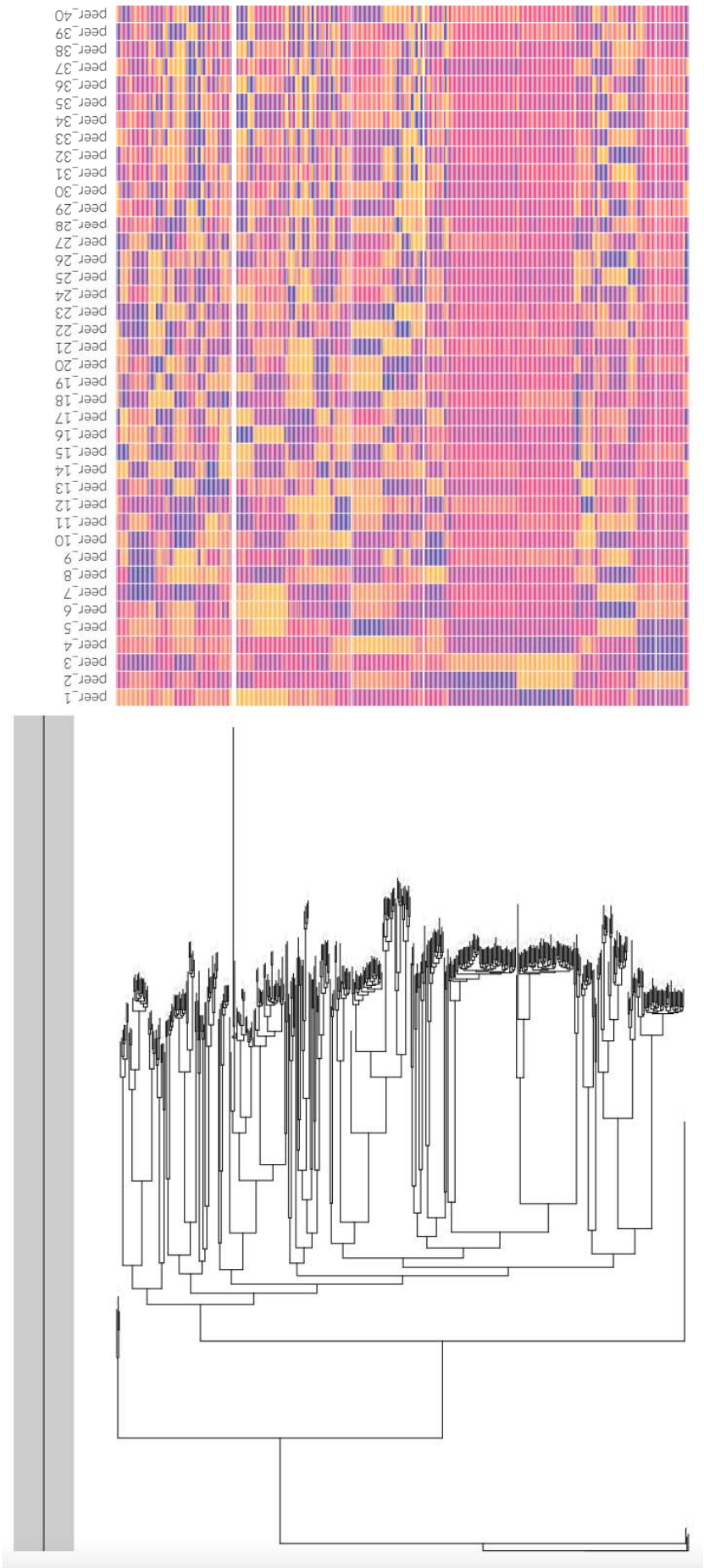


Figure 5.12: PEER factor analysis with $n = 40$ factors of the 460 *S. pneumoniae* genomes in the genome-to-genome analysis. The left panel shows the phylogeny drawn from the core-genome alignment with RAxML; the right panel shows the residual of each factor for the given sample on a continuous scale.

interpretation. Noting that the first components were describing population structure, I instead opted to instead test discrete population clusters for correlation with human variation as the interpretation of the bacterial variants was much more straightforward. This is essentially testing for lineage effects correlated with human variation, as the power to find locus effects is limited (as calculated above). I therefore created a core-genome alignment of these strains using `roary` as in section 4.3, and ran BAPS on this to generate population clusters. I found that the PEER components generally represented the same population structure as the BAPS clusters (fig. A.15).

Cluster	Serotype	Samples	Tested
1	4	17	-
2	-	145	✓
3	8/11A/33F	49	✓
4	10A/35F	22	-
5	23A/B/F	32	✓
6	6B	14	-
7	22F	39	✓
8	9N/15B/19A	47	✓
9	3	47	✓
10	7F	55	✓

Table 5.5: Number of samples in each population cluster. Cluster two is a polyphyletic 'bin' cluster. The dominant serotypes for each cluster, where they account for $> 50\%$ of the isolates, are listed.

Table 5.5 lists the ten clusters found in the data, and the dominant serotypes for each cluster. I ran an association with the BAPS clusters with at least 10% of samples in the subphenotype. The only result reaching genome-wide significance was an association between cluster eight (serotypes 9N/15B/19A) and variants on chromosome 10 (fig. 5.13). The lead variant is at position 134046136 on chromosome 10 (MAF = 0.27; OR = 4.28; p = 1.2×10^{-8}) located in an intron of *STK32C*, a serine/threonine kinase highly expressed in the brain. The high effect size estimated for the interaction is consistent with the power predicted in fig. 5.11.

5.3.3 Association of antigens

This section considers known interactions between host and pathogen proteins, and whether variation in the coding sequence or surrounding regions of each gene is correlated in cases of bacterial meningitis. *S. pneumoniae* has many virulence factors, some of which are known to interact with specific human proteins (Kadioglu et al., 2008). However, I was mostly interested in the interactions where the pneumococcal protein contains sequence variation, ideally somewhat independent of population structure. These regions have

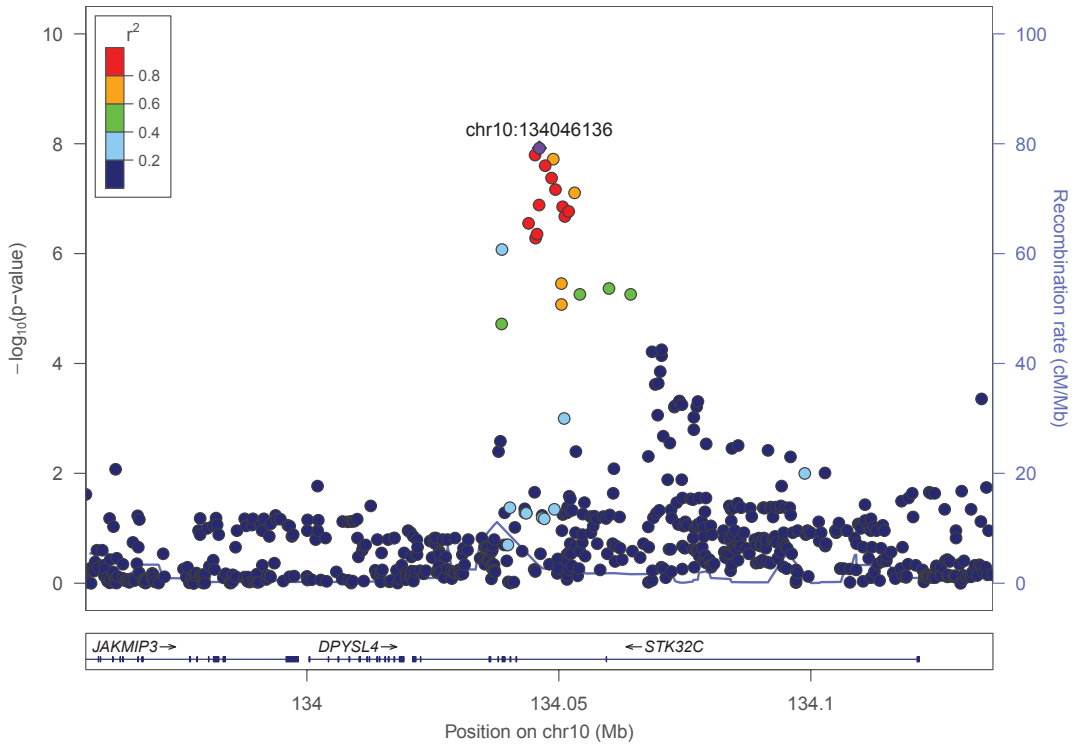


Figure 5.13: Locuszoom plot of association on chromosome 10 with sequence cluster 8, as in fig. 5.4.

higher power to be detected in an association analysis, and the higher rate of variation is potentially a sign of diversifying selection, which may mean the variation is more likely to be associated with specific interactions with the human immune system. Using a combined mapping and assembly approach, followed by a supervised machine learning, in section 4.3.1 I have classified the *pspA*, *pspC* and *zmpA* allele of every sample in the Meningene collection.

pspA is known to bind to C3b, preventing decomposition on the pneumococcal surface and blocking the complement pathway response to infection (Tu et al., 1999). The *LTF* gene encodes lactoferrin, an iron-binding protein found in the granules of neutrophils. This protein is bacteriocidal, and forms part of the innate immune response against pneumococci. It has been found that *pspA* binds lactoferrin to the surface of the pneumococcus, thus reducing their killing by this protein (Shaper et al., 2004).

Like *pspA*, *pspC* has been shown to bind C3 and prevent opsonic decomposition on the pneumococcal surface (Q. Cheng et al., 2000). In addition, some forms of *pspC* have been shown to bind factor H (Janulczyk et al., 2000; Dave et al., 2001). Factor H inhibits complement activation by preventing C3b degrading and activating the next step in the complement pathway. By binding this protein to the surface, the pneumococcus further prevents activation of C3. This locus in the human genome is also known to be involved in susceptibility to invasive meningococcal disease (Davila et al., 2010).

Finally I tested allelic variation of *zmpA*, which is a protease known to bind IgA

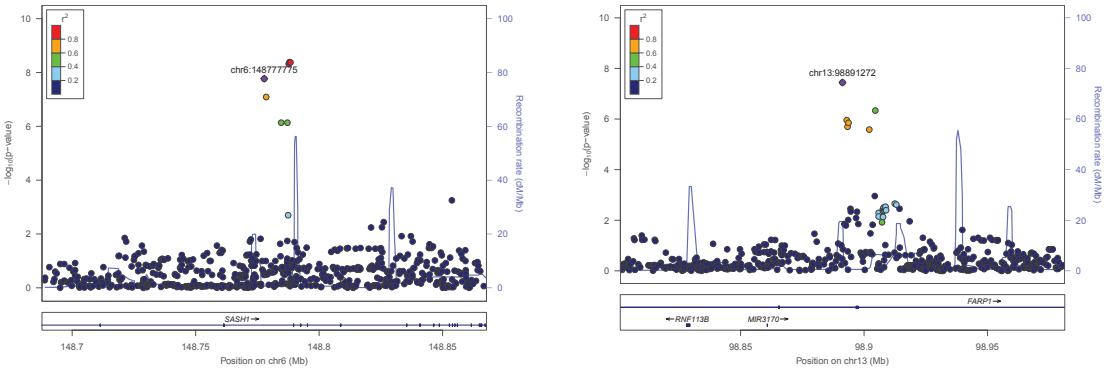
(Wani et al., 1996). This is the most abundant antibody in the nasopharynx, and is an important part of the immune response to pneumococcal infection (Cerutti & Rescigno, 2008). However, it is not produced by simple translation from a single gene and instead involves a pathway covering the HLA along with other regions of the genome (Fagarasan & Honjo, 2003; Ferreira et al., 2010).

For all of the antigen alleles with enough observations (fig. 5.14) I performed an association against all imputed human variants as in section 5.2.2. I used a more accurate imputation of the *CFH* region due to its potential relevance in these interactions. For each test I produced a genome-wide Manhattan plot, and a locuszoom plot for the known interaction partner.

Antigen	Allele	Samples	Tested
pspA	1	214	✓
	2	231	✓
	3	1	-
	4	1	-
cbpA	0	44	✓
	1	6	-
	2	17	-
	3	84	✓
	4	45	✓
	5	60	✓
	6	191	✓
pspC	0	347	✓
	7	7	-
	8	39	✓
	9	45	✓
	10	6	-
	11	3	-
zmpA	1	26	-
	2	236	✓
	3	185	✓

Figure 5.14: Antigen classification of *pspA*, *pspC* and *zmpA*. The total number of samples in the genome-to-genome analysis with each allele is shown, and those where an association test performed are noted.

None of the bacterial antigen alleles were significantly correlated with variants in their human interacting counterparts at the suggestive level ($p < 10^{-5}$). However, there were two associations of *pspC* allele reaching genome-wide significance elsewhere in the genome. Figure 5.15 shows a locuszoom plot of each of these associations. The first is between *pspC*-8 and position 148788006 on chromosome 6 (MAF = 0.08; OR = 9.20; p = 4.1×10^{-9}). This is in *SASH1*, which has previously been found to have decreased expression during meningococcal meningitis (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-11755/) . The second is between *pspC*-9 and position 98891272 on chromosome 13 (MAF = 0.16; OR = 6.30; p = 3.6×10^{-8}), in *FARP1*.



(a) *pspC* allele 8, and chromosome 6 (*SASH1*) **(b)** *pspC* allele 9, and chromosome 13 (*FARP1*)

Figure 5.15: Locuszoom plot of association between *pspC* allele and imputed human SNPs, as in fig. 5.4.

5.4 Conclusions

This chapter has considered the effect of host variation on susceptibility to and severity of pneumococcal meningitis. By using two relatively large well-phenotyped cohorts from the Netherlands and Denmark, I have estimated h^2_{SNP} to be around 30-40% for susceptibility, and around 25% for severity. This suggests that human genetics plays a role in determining how likely invasive disease is, given that a bacteria which is capable of invasion has colonised the individual (chapter 4). Additionally, I have shown that host genetics explains some of the variability in disease outcome after invasion has happened, which may occur by variation in immune response.

I then attempted to use GWAS to find specific variants which contribute to these traits, and while I found signals reaching significance in the Dutch population, none of them have replicated when meta-analysed with summary statistics from other similar studies. No data from other studies is currently available associating human variation with disease outcome, so any planned future confirmation of the association with *UBE2U* may have to use an *in vivo* model of pneumococcal meningitis.

It is difficult to collect bacterial meningitis cases due to: 1) their rarity, and 2) the difficulty to confirm the causative organism by culture. It is even more difficult to determine which of those cases resulted in a poor clinical outcome, as this requires a study design with patient follow-up potentially months after discharge from hospital. The number of cases collected by the collaborators for this analysis is impressive, and this has allowed the first heritability estimates of these traits to be made. These estimates suggest that continuation of the Meningene cohort is warranted, as is the meta-analysis with other well phenotyped studies. With enough cases, specific associations replicating in multiple cohorts will be found. The attempt at meta-analysis I performed here did not find any hits, perhaps due to heterogeneity of phenotype between cohorts. Additionally, a previously reported association in an intron of *CA10* could not be confirmed.

The only previously known genetic association with meningitis is the *CFH* region, which the minor allele is protective for susceptibility to invasive meningococcal disease in children (Davila et al., 2010). I did not find this association with pneumococcal meningitis, though when I restricted analysis to adult meningococcal cases, meta-analysis with the Dutch cohort did not refute its existence. This may suggest a difference in the host response based on invading pathogen, with *CFH* binding being less important for pneumococcal infection.

In the genome-to-genome analysis I was able to put a limit on the strength of interactions that could be detected. Despite being underpowered given the large combined complexity of the host and pathogen populations, I was able to find possible correlations between lineage and host variants. Additionally, some antigen alleles showed possible correlation with variants in the host, though not in regions they are known to interact with. The lack of association may point at the variability of antigen binding of host-proteins being uninvolved in disease course, or may just be limited by a combination of small sample size and high antigenic diversity. The other possible hits from this single study will need replication before biological meaning can confidently be inferred, but the method here shows how such analysis might be done for a bacterial infection, and the results can be used in any future meta-analysis with similar studies.

Chapter 6

Conclusions

6.1 Summary of findings

S. pneumoniae is a human commensal, which in rare cases can invade a usually sterile niche. If the blood or CSF is invaded this usually leads to serious disease, called bacteremia and meningitis respectively. While virulence factors of the pathogen necessary for invasive disease have been identified from bottom-up lab based approaches (often relying on a mouse model of infection), the role of naturally occurring sequence variation in the pathogen genome in invasive disease is generally unknown.

I have used a large cohort of *S. pneumoniae* genomes isolated from invasive disease and asymptomatic carriage to determine the importance of sequence variation in disease susceptibility and severity, and to find the specific regions of the genome which contribute to these variations in phenotype. My main approach was to use GWAS, which is a hypothesis-free way of testing all genomic variants for association with a given phenotype. This approach does not require prior assumptions about which genes may affect the phenotype and does not rely on large-effect size gene knock-outs or animal models of disease.

In the context of bacterial populations, GWAS faces difficulties caused by strong population structure and highly plastic genomes. I developed a piece of software to help overcome these issues by finding an appropriate adjustment for population stratification, and using sequence elements (k-mers) to test for variation of the pan-genome. After testing this method using antibiotic resistance as a positive control, I then applied it to the phenotype of pneumococcal carriage duration, where I also developed a model to estimate carriage duration from longitudinal swab data. By adapting methods derived from human genetics, I was able to calculate the heritability caused by the pathogen genome, and identify which variants explained variation in this important epidemiological parameter.

Using a range of bioinformatic approaches I catalogued variation of the population of pneumococcal genomes sampled from the Netherlands, from both carriage and dis-

ease. I then performed associations between all of these variants and three phenotypes: invasive disease potential, severity and mortality. This analysis showed the importance of pneumococcal variation beyond serotype for invasive potential, but not in disease outcome, and identified many putative genes and regions associated with increased or decreased invasiveness. I also performed an analysis of within-host variation between blood and CSF isolates, and while I didn't find adaptation specific to either niche I did find evidence of selection on genes post-invasion.

Finally I performed a GWAS of host variation with susceptibility and severity of meningitis. I found these traits to be heritable, but despite attempts at meta-analysis with other studies the relatively low sample size and possible prototypic heterogeneity hasn't yet led to a confirmed association in either case. I also attempted a genome-to-genome analysis using both host and pathogen variation. I calculated the limit of detection given the small sample size, and using dimensionality reduction and biological hypotheses found possible interaction effects.

In summary, I have made the following advances. I have developed one of the first methods to overcome the challenges of bacterial GWAS, and showed that it works better than existing approaches. Using this technique, and others, I have quantified the effect of pneumococcal variation on variation in carriage duration beyond the resolution of serotype, and found some of the specific variants which affect it. I also used this top-down approach of assessing the genetics of pneumococcal meningitis, both in host and pathogen. This was not based around known required virulence factors, and used variation occurring in the natural population. Analysis of within-host diversity during meningitis found selection acting on additional genes. I calculated the heritability of host susceptibility to pneumococcal meningitis, and performed an association study using human genetic data. I also attempted the first genome-to-genome analysis with bacterial genomes and human genotypes.

6.1.1 Bacterial genome-wide association studies

Bacterial GWAS approaches have faced three main difficulties: lack of large sample collections, strong population structure confounding results and extensive pan-genomic variation. With the first restriction starting to be lifted, there is a need for scalable GWAS methods directly applicable to large populations of bacterial genomes. Such methods must account for population structure, and ideally assay variation in both the core and accessory genome without relying on a reference alignment.

The use of k-mers to assess pan-genomic variation had previously proven successful, so I wished to implement an approach which could efficiently perform associations using these as sequence variants. As the application of phylogeny based approaches are restricted due to their heavy computational burden and the need for an accurate recombination-

free tree, I opted to adapt regression-based methods used in human genetics to apply to bacterial GWAS. I wrote code to maximise the likelihoods of these regressions in C++, using efficient optimisation techniques as a first try, and more robust methods as a second pass.

To work out how to deal with population structure I compared various approaches in terms of accuracy and computational burden for phylogeny reconstruction. Knowing that I would be using k-mers in the association, I found that a method using the Jaccard distance between subsets of overall k-mers was sufficient to control for population structure in my simulations, and for antibiotic resistance in *S. pneumoniae*. Since writing this code the minHash distance has been adapted for distances between genetic sequences (Ondov et al., 2016), and can now be used as a more efficient replacement for Jaccard distance. I used the eigenvectors with the three largest eigenvalues calculated from this pairwise distance matrices in a fixed effect logistic or linear regression, in analogy with the standard method used in human genetics. To deal with possible very large effect sizes in these regressions I used the LRT for significance, and Firth regression for when data was nearly separable (as in trimethoprim resistance).

This approach proved to be broadly successful for antibiotic resistance in *S. pneumoniae*, worked with simulated data, and found a potential virulence factor in *S. pyogenes*. However, in all of these cases the predicted effect size was very strong, and population structure was generally not strongly associated with the phenotypes tested. I did not test whether the population structure correction I applied here was more broadly applicable, and would be sufficient in other species or phenotypes where these conditions no longer hold. The use of more eigenvectors should improve the trade-off between false positive rate and power, but it may be the case that including them as random-effects under a linear mixed model may offer the best option. When used for carriage duration, I found that a LMM had slightly higher power for detecting homoplasic low frequency effects when compared to using fixed effects while controlling for false positive rate. However, it was not as useful as the fixed effects model for including possible lineage associated variants for follow-up elsewhere. For invasiveness of *S. pneumoniae*, the fixed effect model using ten population structure components appeared to have a high false positive rate, where the LMM offered better population structure control and was still powered to find associations.

I have therefore already observed situations in which different methods would be the best to use. A comparison of these possible methods based on a range of population structures, phenotype distribution, recombination rate/homoplasy, effect sizes, lineage and locus associations would be useful, and is not something I attempted here. It is difficult to simulate realistic bacterial phylogenies, and synthetic associations introduced as part of this kind of simulation may be easier to find than associations in real populations. To perform this comparison I would take observed sequence alignments from real populations, and introduce synthetic associations using eq. (2.11) over the range of parameters of interest.

A comparison of power and false positive rate of fixed and random effect regressions as well as phylogeny based methods would be useful for future applications.

The population structure correction I used in SEER is a reasonable start, and works well for strong effects such as antibiotic resistance. A comparison with other possibilities with positive and negative controls (either simulated or known associations) will help inform future development. I mostly tested methods on locus effects, and have ignored or controlled for lineage effects in the output. In the future, a ranking of lineage effects in the output would be useful in case lab-based follow-up of these sites is possible. Clear assignment of sites as either lineage or locus effects would be helpful too, and ancestral state reconstruction combined with a comparison between adjusted and unadjusted test statistics may help classify variants into one of these two classes.

A difficulty in both cases is picking a significance threshold. In my first attempts I reasoned that every possible site in the genome multiplied by all three possible mutations should be used as the number of tests, and backed this up with permutation testing. However, as samples are not independent and identically distributed due to their genetic relatedness then permuting phenotype labels may not be appropriate, as it assumes any switch of label has the same effect ignoring any covariance between samples. Permuting labels within population clusters may be better, but likely too conservative. Monte Carlo permutation using the a covariance structure calculated from the phylogeny is also possible, though with the usual caveats of computational burden and reliance on a high-quality tree. Inspection of Q-Q plots is useful and can visually allow for the identification of a breakpoint between population structure effects and a significant signal, and how much the former is affecting the association model overall. While this is not a consistent way of choosing a threshold, it can help with ranking the top hits. For the LMM, where population structure is well controlled at the lower end of the p-value spectrum, a conservative Bonferroni correction based on number of patterns seems appropriate based on the Q-Q plots tested here. For fixed effects models picking this threshold remains a challenge, though Q-Q plots can help.

The use of k-mers worked well in the applications tested, and managed to find associations SNPs would not. They enjoyed the expected advantages from not requiring an alignment or clustering of orthologous genes. In cases where nearby SNPs independently affect the phenotype, which occurs in some antibiotic resistance genes, k-mers may be split up into lower frequency sequence units, lowering their power. In later chapters I therefore assessed variation through k-mers, SNPs and COGs where possible. The interpretation of k-mers has proved more challenging, due to the difficulty of mapping to the correct place (particularly with smaller k-mers) in a well annotated genome. Ideally k-mers would further be annotated by labelling SNPs and their predicted functional change in the k-mers, using the ancestral state as reference. However, to map an associated region, especially mediated by gene presence/absence and not fine-map the function, k-mers have been

successful.

6.1.2 Epidemiological variation of *S. pneumoniae*

Duration of carriage of *S. pneumoniae* is an important measure of strain fitness in epidemiological models, and its variation has been proposed as a mechanism by which antibiotic sensitive and resistant strains can coexist. Previous analysis of the source of this variation have been limited to serotype resolution, so using genome sequences from a longitudinal study cohort offered the opportunity to refine the analysis of variance.

I first developed HMMs for longitudinal swabbing data per serotype, to allow carriage acquisition and clearance rates and false negative swabbing rates to be estimated from the whole data rather than from a set of assumptions. The only model that converged for the most common serotypes was the simplest: two states for carrying and not-carrying. These parameters could then be applied to individual carriage episodes to infer the most likely durations based on the observed data. Using these durations as a continuous phenotype, I used a LMM to investigate and quantify the variance components caused by serotype and resistance, and GWAS to identify possible specific genetic variation which further contributed to variation in carriage duration.

I found that bacterial genomic variation had a significant effect on carriage duration, and that serotype was the largest lineage effect. However, only serotype 19F appeared to have a contribution independent of the genetic background. I also identified prophage k-mers which were associated with a lowered carriage duration, and evidence that this may work through interruption of the competence mechanism (by inserting into the *comYC* gene). These findings support the existence of duration and fitness modifying alleles in the natural population, which can be used to explain coexistence of antibiotic resistant and sensitive strains despite strong fitness differences depending on whether treatment is currently being applied (Lehtinen et al., 2017). The increased precision of the carriage estimates, per carriage episode rather than per serotype, along with provision of useful covariates such as *comYC* status, host age and previous carriage will also be useful data for models of coexistence and transmission.

However, one of the main limitations of this analysis was the monthly swabbing resolution. While clearly a large and well-sampled collection, the design of swabs spaced linearly in time to probe carriage durations which appear to be exponentially distributed is suboptimal. A design that would be better for this purpose is exponentially distributed sampling of cases that remain positive (Abdullahi et al., 2012a). Given the swabbing design available here, the estimates of effect sizes of the explanatory variables on carriage duration were therefore positively skewed.

As with all GWAS studies from a single population, results may be affected by batch effects in this population. Therefore meta-analysis of the results from this section with

another similar study would be useful before being generalised to the entire pneumococcal species. However, as this amount of sequencing has previously been unfeasible and as children need to be followed for two years, these studies are difficult to set up and long-running from start to finish. There are no other studies currently combining carriage duration estimates with genomic data, so this meta-analysis is not presently possible. We are aware of a similar study starting collection in Cape Town, South Africa, so I have released our results to facilitate comparison when this study's sequencing has been completed.

The function of altered carriage duration through *comYC* is an association only, and does not prove causation through this mechanism. While it is possible to make evolutionary arguments to support this interpretation, isogenic strains (controlling perfectly for genetic background) in an *in vivo* model would be needed to bolster this claim.

6.1.3 Host and pathogen genetics of pneumococcal meningitis

In chapters 4 and 5 I have used genomic variation of infecting bacteria and human host respectively to determine the impact of genetics on susceptibility to and severity of pneumococcal meningitis. Heritability analysis showed that for susceptibility, host genetics played a role and the genome sequence of the infecting strain is very important in whether invasive disease can occur. For severity of disease a different picture emerged: pathogen variation is unimportant, and host genetics is likely to play a small role. Though the estimation of specific heritabilities with binary phenotypes can be problematic, the data and multiple models support this overall conclusion.

I was unable to find and validate specific host associations through meta-analysis with other studies given the current sample collection. This rules out the existence of common variants with large effect sizes, the fitness defect of which would be unlikely to exist evolutionarily. Whether the variation which contributes to this phenotype consists of low effect size common variants, or rarer large effect size variants is a question that will need to be answered by future studies with larger sample sizes and more sequencing covering the entire variant frequency spectrum.

I did not include the sex chromosomes in the present analysis due to difficulties with imputation, though I did perform an earlier analysis of the X chromosome when using `impute2` in the Dutch population that did not show any association with any of the phenotypes. Tools are being developed to deal with the sex chromosomes in the same way as the autosomes (Wise et al., 2013), and the imputation server and reference panel now allows the X chromosome to be included. Future analyses should therefore not ignore this variation.

Another issue was phenotype heterogeneity, as the cohorts differed in terms of participant age and the exact disease presentation. While these differences have not been

found to matter for many phenotypes, it is possible that differing effect sizes in the subtly different phenotypes here are making associations impossible to find given the model used. The sample size here may benefit from a specific model allowing for this heterogeneity and expected correlations between effect sizes (as evidenced by the lack of signal from subtest), though a simple first step would be to perform meta-analysis of only a subset of the available studies to test for this possibility.

For the bacterial genetic contribution to meningitis, using GWAS I found many regions of the genome to be associated with invasiveness. Reassuringly, positive controls such as capsule (which I separately estimated to account for half of the variation in invasiveness) and LoF mutations in virulence factors such as *zmpD* were found in this analysis. Some other genes had previously been reported to affect virulence in invasive disease models, and these results increase support for their importance in human disease too. The remaining regions were associated with virulence for the first time here, and may suggest new functions for these genes, or an impact on virulence through unknown interacting gene networks.

I used a simple burden test when testing the effect of rare variants, which would not be suitable if the variants included in the set had different directions of effect. While this is probably correct for LoF variants, a different test may increase power for rare missense variants affecting protein function. If there is still strong population structure at the tips of the tree the method I have used has not explicitly accounted for it. It would be possible to instead group variants manually, and perform the association using a LMM. A similar caveat exists with the Tajima's D analysis of differential selection, where permutation testing may be insufficient to correct for population structure. In this case, the confounding effect of different population histories or different effects of vaccine introduction may be impossible to disentangle from signatures of selection.

These GWAS results are particularly susceptible to batch effects, due to the difficulty of getting a perfectly matched sample of the population from carriage and invasive disease. When analysing binary traits, if a covariate (such as serotype) is perfectly correlated with the trait, then all the results will be confounded too. Therefore a crucial next step, before further interpretation, is replication and meta-analysis with another population where both carriage and disease have been sampled. Hits from both populations will then be much better supported as the confounders may cancel out if in random directions, and power will be raised for rarer and lower effect size variants. A project is underway in South Africa which has taken such a sample, so we intend to perform this meta-analysis using those sequences.

As mentioned in chapter 1 part of the power of GWAS over linkage studies comes from the simple study design, where as many samples as possible are used without necessarily worrying about matching for covariates or genetic background. These confounders can then be adjusted for in the downstream analysis instead, which maximises discovery power.

This is broadly true for bacterial genomes too, however the effect of population structure is a much stronger confounder, and for some phenotypes which are tightly correlated with genetic background (high heritability) this can make discovery of anything other than homoplasic variants impossible. An alternative study design is to instead compare variation from within the same bacterial population when it has divergent phenotypes. For example, sampling the diversity of the bacterial population within-host in the carriage niche and an invaded niche is not confounded by population structure (and also host covariates such as age and immune response) as the genetic background is the same. Performing a meta-analysis of the variation found to be associated with either niche across multiple samples will then find those variants which occur during infection which have allowed adaptation to the invaded niche.

I performed this analysis between blood and CSF isolates, as previous work on a single case of pneumococcal meningitis had found convincing evidence for evolution occurring during invasive disease. When I expanded to hundreds of cases, I found no evidence of any variation causing adaptation to either the blood or CSF niche during disease. The sample size was large enough to conclusively state that variation occurring after invasion is rarely important for the progression of meningitis. However, when comparing the variation present in populations from invasion to carriage reference sequences I did find signs that *dlt* loses function in carriage more frequently than would be expected, and that *pde1* is under selection in invasion. To refine this analysis of variation occurring within-host between carriage and disease I would need to use more samples than analysed here, and also deeper sequencing of samples to assay the background of variation that exists within the founding population that is then selected.

6.2 Future directions

6.2.1 Bacterial GWAS methods

Since its release, I have received feedback about SEER which, if implemented, would make it into a more broadly usable and applicable piece of software for microbiologists. In terms of software development and installation, inclusion of SEER in a common 'container' would make installation automatic for those without C/C++ development experience, deal with differences between platforms and ensure all users are working with the same version of the code base.

I designed SEER with k-mers in mind, and therefore concentrated on making a scalable piece of software with a single input source. As mentioned, k-mers may not be the ideal variant when close SNPs are associated with a phenotype as the resulting k-mers will be split up into words of smaller frequency, and therefore power. For some purposes it may be useful to allow other forms of input such as VCF for short variants (SNPs and INDELs) with respect to a reference, and a general presence/absence matrix for COGs and aligned intergenic regions. The interpretation of k-mers can be challenging, both in finding a suitable reference (even from the entire `nr/nt`) to map to and annotate them with, and to determine whether they represent presence/absence of a region or variation within the region. It has been recently argued that population variation is best represented by a pan-genome graph, with shared haplotypes of any length being the natural variant (Marschall et al., 2016; Paten et al., 2017). Though the counting of informative k-mers goes some way toward testing longer variants, testing haplotypes may improve association power and make interpretation easier. A method has been proposed using unitigs (high confidence contigs not requiring repeat resolution), though this is not likely to scale beyond hundreds of samples (Jaillard et al., 2017). Integrating a scalable approach such as `vg` (variant graph – https://github.com/vgteam/vg) would be a promising way to include haplotype association.

Section 4.4.2 considered rare variation in GWAS assuming population structure was not an issue, due to low frequency variants occurring at the tips of the phylogeny. Including a way to input pathways of variants in SEER would relax this assumption, and also allow both gene-based burden tests (in either direction) to be extended to operons and functional pathways. Adding a model such as `SKAT` (Wu et al., 2011) would also improve power when rare variants in a functional pathway do not all act in the same direction on the phenotype of interest.

I picked a single method to adjust for population structure in SEER, but many others could be used. For example, as shown in chapters 3 and 4, the fixed effect model of SEER is in some cases a poor control for population structure. In the current implementation, BAPS clusters could be used as a categorical covariate in the regression giving a similar test to the CMH. A LMM has generally shown good control of population structure, likely thanks to

using all SNPs in the population structure correction rather than a proportion through picking the top principal components. The LMM normally has complexity $\mathcal{O}(MN^3)$, which is infeasible for the GWAS problems considered here and as sample sizes grow in future. The model of FaST-LMM rotates the design ($\boldsymbol{X}$) and relatedness ($\boldsymbol{G}$) matrices so the regression becomes linear along the eigenvectors of $\boldsymbol{G}$ (first using a singular value decomposition of $\boldsymbol{G}$), which with correct selection of $\boldsymbol{G}$ has complexity $\mathcal{O}(MN)$ (Lippert et al., 2011; Kadie & Heckerman, 2017). In this case, $\boldsymbol{G}$ is a SNP-wise distance between samples. This is similar to the $\mathcal{O}(MN^2)$ phylogenetic regression method of Pagel (1997) which transforms correlated error terms (due to relatedness between samples) into uncorrelated errors by diagonalising the variance-covariance matrix $\boldsymbol{G}$. In this case, $\boldsymbol{G}$ is the distance between the root and MRCA of each pair of samples. These methods could be included as new association models in SEER to allow for population structure correction when the current fixed effect model is not appropriate.

The effect on GWAS power and false positive rate of these different population structure corrections is unknown, and will likely be different depending on variant penetrance, level of homoplasy and frequency. A simulation-based comparison between these methods over a range of situations would therefore be useful. Based on the simulations used in section 2.6.1, the best way to do this would be by adding in synthetic associations of different penetrance at various points of the phylogeny of a real population using eq. (2.11), which would allow varying homoplasy and frequency.

I used heritability and genomic partitioning to support the conclusions in chapters 3 and 4. While this is well-supported for continuous trait used in the former, the use of the liability scale for bacterial traits in the latter has not been properly explored. Extension to binary traits would be useful, and support of the applicability and robustness of the methods used from simulated data will be important for having faith in quantitative estimates. If this could be shown to work, the estimates of serotype importance may be better estimated in a framework where genetic background is separately accounted for.

The use of SEER has been exclusively to single traits, but with the increasing availability of high dimensional phenotypes as seen in genome-to-genome analysis (section 5.3), pheWAS (Bush et al., 2016) and eQTL studies (L. Franke & Jansen, 2009; Wang et al., 2009) the addition of a multitrait model could be considered. Transcriptomic data is now being produced for bacteria (Bruchmann et al., 2015), so improved association power of SEER for this purpose will be useful. Rather than associating every phenotype or transcript separately, necessitating a harsh multiple testing correction, the correlation structure of multiple traits can be exploited to find latent variables (biologically representing functional pathways) to test for association with genetic variation improving power (Marttinen & Corander, 2010; Marttinen et al., 2013; Marttinen et al., 2014). Recent implementations of non-negative matrix factorisation are fast, and a promising way to find latent variables in high dimensional phenotypes (Zhirong Yang et al., 2016) – so could be added as a further

module in SEER.

6.2.2 Genetics affecting pneumococcal meningitis

Further analysis using GWAS could further explain the biology of pneumococcal infection. A simple additional analysis would be adult versus child colonisation using the Dutch carriage population – I have already catalogued the variation, and host age is available for all samples. Any results may be informative of the differences in immune system evasion depending on host response, and could be important for vaccination which currently targets children.

In the carriage stage, bacteria will only persist in the population if they can be transmitted between hosts; 'transmissibility' of *S. pneumoniae* is therefore a measure of fitness. Alleles which affect transmissibility may also be a promising vaccine candidate, as compared to PCV they will reduce colonisation (and therefore disease) of all serotypes. Zafar et al. (2017) have shown that *ply* is necessary for transmission, as the host cell damage it causes increases shedding. A GWAS of *S. pneumoniae* transmissibility may be able to detect more subtle effects of alleles which occur in the natural population.

Nebenzahl-Guimaraes et al. (2016) performed a GWAS on transmissibility of *M. tuberculosis* by selecting low transmission strains from at-risk hosts with rare genotypes and high transmission strains from low-risk hosts with common genotypes. A similar way to perform this analysis would be to use the carriage durations I estimated in chapter 3 and assume equilibrium transmission in an susceptible-infected-recovered (SIR) model in the Maela population, which would then allow calculation of strain transmissibility from carriage duration divided by strain prevalence. However, evidence from infant mouse models suggests *S. pneumoniae* transmission may only occur shortly after colonisation, when inflammation is highest promoting increased shedding (Kono et al., 2016; Zafar et al., 2017). In this case a more complex transmission model using genetic similarity and infection times may be more appropriate, and model comparison between different functions of transmission intensity with respect to time would also be useful for inferring the biology of real-life transmission. Numminen et al. (2013) proposed a more flexible transmission model for the Maela population which was fitted with approximate Bayesian computation. Due to many proposals of the transmission tree being inconsistent with the observed infection times (and being assigned $\mathscr{L} = 0$) the fitting was computationally intensive; the use of the carriage durations estimated here rather than single time-points may ameliorate this problem. Inference of alleles affecting transmissiblity could then be jointly estimated in the process of inferring the transmission trees. Alternatively, if the dimension of genetic variation is too high, they could be inferred separately by first calculating strain-wise transmissibility from the transmission trees and then using these as a phenotype in GWAS. An alternative approach would be to sample within-host diversity

by deep sequencing of swabs, which allows finding the genotypes which make it through the transmission bottleneck in each case through ancestral state reconstruction (where the trait is the identity of the host). Averaged over many transmission chains, the variation shared by these genotypes would represent transmissibility alleles.

In the analysis of host genetics affecting bacterial meningitis, a better model of the shared architecture between the subtypes of meningitis analysed may help find associations (Pickrell et al., 2016). Rather than using subtest with underpowered genotype data, it may be better to use LD-score regression between summary statistics from all the studies available, which would allow estimation of coheritabilities between the different sub-phenotypes (Bulik-Sullivan et al., 2015). To aid in increasing power for detecting host genetics we have applied to access the UK biobank (http://www.ukbiobank.ac.uk/), which is about to release 500 000 genotypes of a richly phenotyped UK adult population. These phenotypes include ICD-10 codes, which show hospital diagnoses for bacterial meningitis, split up by causal species. Additionally, date of death is available, allowing inference of clinical outcome. The large size and well-defined phenotype of these samples will allow us to perform another GWAS, and meta-analyse the results with those of chapter 5 for both susceptibility and severity increasing discovery power.

The genome-to-genome analysis was limited by the small sample size when testing massive numbers of combinations of possible interactions. In future, the ~1 200 samples from the Danish cohort will also have the causal *S. pneumoniae* sequenced, allowing this analysis to be expanded. It may also be possible to model the effect of genome-to-genome interactions on severity as well as bacterial and host factors, by analysing a combined model of the form:

$$\text{severity} \sim X_{\text{bacteria}} + X_{\text{host}} + X_{\text{interaction}}$$

where the interaction term is $X_{\text{bacteria}} \times X_{\text{host}}$.

Finally, I would propose the following extensions to assessing with-host diversity during bacterial meningitis. As I have shown that selection does not occur between blood and CSF samples, but that it probably does occur between carriage and CSF, a greater number of carriage and invasive samples from the same patient should be taken: greater both in terms of the number of patients enrolled and in the depth of coverage of the within-host diversity. This is a difficult study to set up: in the MeninGene cohort recent attempts to swab bacteria from the nasopharynx of bacterial meningitis patients before treatment started yielded no positive cultures, likely due to the small carriage population (Wyllie et al., 2014; Wyllie et al., 2016). Alternative culture-free methods such as DNA pull-down may be helpful, or alternative a study in an alternative population with high rates of carriage may be able to achieve sufficient sample size.

The analysis of this data would benefit from an improved null model of mutation. In section 4.5 I assumed a simple model of equal mutation rate per base and Poisson dispersion

of number of mutations, which led to regions with higher mutation rates being found, and may have suppressed the discovery of genes with lower mutation rates. Improving this through a more refined model of mutation rates depending on sequence context and using observed dispersion of the number of mutations would be a useful extension (Samocha et al., 2014; Aggarwala & Voight, 2016). If more mutations were observed, using the observed number of synonymous changes, which are assumed to be neutral, as a basis for the null would also help (Ding et al., 2008). Finally, experimental evolution without selection pressure may give the most accurate null model (Tenaillon et al., 2016), though an experiment recreating the bottlenecks encountered in pneumococcal meningitis has not yet been performed.

6.2.3 Future of statistical genetics in bacterial diseases

Statistical genetics, and specifically GWAS, of host and pathogen genetics contributing bacterial diseases is still in its infancy. Looking at the boom in human genetics and given the large sample sizes becoming available, it is reasonable to expect the field to continue to expand. The near future is likely to consist of further methodological improvements and analysis of new phenotypes, going on to functional validation and eventually integration with host data. I hope that I have presented some reasonable early steps in this field in this thesis, and that others find elements of what we've done useful for future research.

Thanks a lot for reading all the way to the end! (unless you skipped straight here)

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65.

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Abdullahi, O., Karani, A., Tigoi, C. C., Mugo, D., Kungu, S., Wanjiru, E., … Scott, J. A. G. (2012a). Rates of acquisition and clearance of pneumococcal serotypes in the nasopharynges of children in kilifi district, kenya. *J. Infect. Dis. 206*(7), 1020–1029.

Abdullahi, O., Karani, A., Tigoi, C. C., Mugo, D., Kungu, S., Wanjiru, E., … Scott, J. A. G. (2012b). The prevalence and risk factors for pneumococcal colonization of the nasopharynx among children in kilifi district, kenya. *PLoS One*, *7*(2), e30787.

Abel, L. & Dessein, A. J. (1997). The impact of host genetics on susceptibility to human infectious diseases. *Curr. Opin. Immunol. 9*(4), 509–516.

Abeyta, M., Hardy, G. G. & Yother, J. (2003). Genetic alteration of capsule type but not PspA type affects accessibility of surface-bound complement and surface antigens of streptococcus pneumoniae. *Infect. Immun. 71*(1), 218–225.

Adriani, K. S., Brouwer, M. C. & Beek, D. V. D. (2015). Risk factors for community-acquired bacterial meningitis in adults. *Neth. J. Med. 73*(2), 53–60.

Afzal, M., Shafeeq, S., Henriques-Normark, B. & Kuipers, O. P. (2015). UlaR activates expression of the ula operon in streptococcus pneumoniae in the presence of ascorbic acid. *Microbiology*, *161*(Pt 1), 41–49.

Aggarwala, V. & Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet. 48*(4), 349–355.

Agresti, A. (2015). *Foundations of linear and generalized linear models (wiley series in probability and statistics)* (1 edition). Wiley-Blackwell.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr. 19*(6), 716–723.

Alam, M. T., Petit, R. A., 3rd, Crispell, E. K., Thornton, T. A., Conneely, K. N., Jiang, Y., … Read, T. D. (2014). Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome Biol. Evol. 6*(5), 1174–1185.

Alfonseca, M., Cebrián, M. & Ortega, A. (2005). Common pitfalls using the normalized compression distance: What to watch out for in a compressor. *Commun. Inf. Syst. 5*(4), 367–384.

Allegrucci, M., Hu, F., Shen, K., Hayes, J., Ehrlich, G. D., Post, J. C. & Sauer, K. (2006). Phenotypic characterization of streptococcus pneumoniae biofilm development. *J. Bacteriol. 188*(7), 2325.

AlonsoDeVelasco, E., Verheul, A. F., Verhoef, J. & Snippe, H. (1995). Streptococcus pneumoniae: Virulence factors, pathogenesis, and vaccines. *Microbiol. Rev. 59*(4), 591–603.

Altshuler, D., Daly, M. J. & Lander, E. (2008). Genetic mapping in human disease. *Science*, *322*(5903), 881–888.

Amos, W. & Hoffman, J. I. (2010). Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc. Biol. Sci. 277*(1678), 131–137.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nat. Protoc. 5*(9), 1564–1573.

Anderson, T. J. C., Williams, J. T., Nair, S., Sudimack, D., Barends, M., Jaidee, A., … Nosten, F. (2010). Inferred relatedness and heritability in malaria parasites. *Proc. Biol. Sci. 277*(1693), 2531–2540.

André, G. O., Politano, W. R., Mirza, S., Converso, T. R., Ferraz, L. F. C., Leite, L. C. C. & Darrieux, M. (2015). Combined effects of lactoferrin and lysozyme on streptococcus pneumoniae killing. *Microb. Pathog. 89*, 7–17.

Anttila, M., Voutilainen, M., Jäntti, V., Eskola, J. & Käyhty, H. (1999). Contribution of serotype-specific IgG concentration, IgG subclasses and relative antibody avidity to opsonophagocytic activity against streptococcus pneumoniae. *Clin. Exp. Immunol. 118*(3), 402–407.

Aronin, S. I., Peduzzi, P. & Quagliarello, V. J. (1998). Community-acquired bacterial meningitis: Risk stratification for adverse clinical outcome and effect of antibiotic timing. *Ann. Intern. Med. 129*(11), 862–869.

Attia, J., Hatala, R., Cook, D. J. & Wong, J. G. (1999). The rational clinical examination. does this adult patient have acute meningitis? *JAMA*, *282*(2), 175–181.

Auranen, K., Mehtälä, J., Tanskanen, A. & S Kaltoft, M. (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*(2), 169–176.

Azim Ansari, M., Pedergnana, V., Ip, C. L. C., Magri, A., Von Delft, A., Bonsall, D., … Spencer, C. C. A. (2017). Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.*

Bai, Y., Yang, J., Eisele, L. E., Underwood, A. J., Koestler, B. J., Waters, C. M., … Bai, G. (2013). Two DHH subfamily 1 proteins in streptococcus pneumoniae possess cyclic Di-AMP phosphodiesterase activity and affect bacterial growth and virulence. *J. Bacteriol. 195*(22), 5123–5132.

Bainbridge, T. W., DeAlmeida, V. I., Izrael-Tomasevic, A., Chalouni, C., Pan, B., Goldsmith, J., … Ernst, J. A. (2014). Evolutionary divergence in the catalytic activity of the CAM-1, ROR1 and ROR2 kinase domains. *PLoS One*, *9*(7), e102695.

Balachandran, P., Hollingshead, S. K., Paton, J. C. & Briles, D. E. (2001). The autolytic enzyme LytA of streptococcus pneumoniae is not responsible for releasing pneumolysin. *J. Bacteriol. 183*(10), 3108–3116.

Balmer, P., North, J., Baxter, D., Stanford, E., Melegaro, A., Kaczmarski, E. B., … Borrow, R. (2003). Measurement and interpretation of pneumococcal IgG levels for clinical management. *Clin. Exp. Immunol. 133*(3), 364–369.

Bamshad, M. & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nat. Rev. Genet. 4*(2), 99–111.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A. S., … Pevzner, P. a. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol. 19*(5), 455–477.

Barrett, J. C., Buxbaum, J., Cutler, D., Daly, M., Devlin, B., Gratten, J., … Wray, N. R. (2017). *New mutations, old statistical challenges*.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., … Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with escherichia coli. *Nature*, *461*(7268), 1243–1247.

Bartha, I., Carlson, J. M., Brumme, C. J., McLaren, P. J., Brumme, Z. L., John, M., … Fellay, J. (2013). A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife*, *2*, 1–16.

Bek-Thomsen, M., Poulsen, K. & Kilian, M. (2012). Occurrence and evolution of the paralogous zinc metalloproteases IgA1 protease, ZmpB, ZmpC, and ZmpD in streptococcus pneumoniae and related commensal species. *MBio*, *3*(5).

Bensing, B. A., Siboo, I. R. & Sullam, P. M. (2001). Proteins PblA and PblB of streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect. Immun. 69*(10), 6186–6192.

Bentley, S. D., Aanensen, D. M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., … Spratt, B. G. (2006). Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet. 2*(3), e31.

Bentley, S. D. & Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annu. Rev. Genet. 38*, 771–792.

Berry, A. M. & Paton, J. C. (2000). Additive attenuation of virulence ofstreptococcus pneumoniae by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. *Infect. Immun. 68*(1), 133–140.

Bijlsma, M. W., Brouwer, M. C., Kasanmoentalib, E. S., Kloek, A. T., Lucas, M. J., Tanck, M. W., … van de Beek, D. (2016). Community-acquired bacterial meningitis in adults in the netherlands, 2006–14: A prospective cohort study. *Lancet Infect. Dis. 16*(3), 339–347.

Bille, E., Ure, R., Gray, S. J., Kaczmarski, E. B., McCarthy, N. D., Nassif, X., … Tinsley, C. R. (2008). Association of a bacteriophage with meningococcal disease in young adults. *PLoS One*, *3*(12), e3885.

Bille, E., Zahar, J.-R., Perrin, A., Morelle, S., Kriz, P., Jolley, K. A., … Tinsley, C. R. (2005). A chromosomally integrated bacteriophage in invasive meningococci. *J. Exp. Med. 201*(12), 1905–1913.

Blanquart, F., Wymant, C., Cornelissen, M., Gall, A., Bakker, M., Bezemer, D., … BEEHIVE collaboration. (2017). Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in europe. *PLoS Biol. 15*(6), e2001855.

Bogaardt, C., van Tonder, A. J. & Brueggemann, A. B. (2015). Genomic analyses of pneumococci reveal a wide diversity of bacteriocins - including pneumocyclicin, a novel circular bacteriocin. *BMC Genomics*, *16*, 554.

Bogaert, D., van Belkum, A., Sluijter, M., Luijendijk, A., de Groot, R., Rümke, H. C., … Hermans, P. W. M. (2004). Colonisation by streptococcus pneumoniae and staphylococcus aureus in healthy children. *Lancet*, *363*(9424), 1871–1872.

Bohr, V., Rasmussen, N., Hansen, B., Kjersem, H., Jessen, O., Johnsen, N. & Kristensen, H. S. (1983). 875 cases of bacterial meningitis: Diagnostic procedures and the impact of preadmission antibiotic therapy. part III of a three-part series. *J. Infect. 7*(3), 193–202.

Borgström, E., Redin, D., Lundin, S., Berglund, E., Andersson, A. F. & Ahmadian, A. (2015). Phasing of single DNA molecules by massively parallel barcoding. *Nat. Commun. 6*, 7173.

Bosch, A. A. T. M., van Houten, M. A., Bruin, J. P., Wijmenga-Monsuur, A. J., Trzciński, K., Bogaert, D., … Sanders, E. A. M. (2016). Nasopharyngeal carriage of streptococcus pneumoniae and other bacteria in the 7th year after implementation of the pneumococcal conjugate vaccine in the netherlands. *Vaccine*, *34*(4), 531–539.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet. 32*(3), 314–331.

Brandtzaeg, P. (1993). Meningitis and septic shock as acute, fatal conditions. *Tidsskr. Nor. Laegeforen. 113*(16), 1994–1997.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., … Jansen, R. C. (2008). Genetical genomics: Spotlight on QTL hotspots. *PLoS Genet. 4*(10), e1000232.

Brittan, J. L., Buckeridge, T. J., Finn, A., Kadioglu, A. & Jenkinson, H. F. (2012). Pneumococcal neuraminidase a: An essential upper airway colonization factor for streptococcus pneumoniae. *Mol. Oral Microbiol. 27*(4), 270–283.

Brogden, K. A. (2005). Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol. 3*(3), 238–250.

Brooks-Walter, A., Briles, D. E. & Hollingshead, S. K. (1999). The pspc gene of streptococcus pneumoniae encodes a polymorphic protein, PspC, which elicits cross-reactive antibodies to PspA and provides immunity to pneumococcal bacteremia. *Infect. Immun. 67*(12), 6533–6542.

Brouwer, M. C., de Gans, J., Heckenberg, S. G. B., Zwinderman, A. H., van der Poll, T. & van de Beek, D. (2009). Host genetic susceptibility to pneumococcal and meningococcal disease: A systematic review and meta-analysis. *Lancet Infect. Dis. 9*(1), 31–44.

Brouwer, M. C., Heckenberg, S. G. B., de Gans, J., Spanjaard, L., Reitsma, J. B. & van de Beek, D. (2010). Nationwide implementation of adjunctive dexamethasone therapy for pneumococcal meningitis. *Neurology*, *75*(17), 1533–1539.

Brouwer, M. C., McIntyre, P., Prasad, K., van de Beek, D., Mc, B. & D, V. D. B. (2013). Corticosteroids for acute bacterial meningitis. *Cochrane Database Syst. Rev. 6*(6), CD004405.

Brouwer, M. C., Tunkel, A. R. & van de Beek, D. (2010). Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis. *Clin. Microbiol. Rev. 23*(3), 467–492.

Brown, P. D., Davies, S. L., Speake, T. & Millar, I. D. (2004). Molecular mechanisms of cerebrospinal fluid production. *Neuroscience*, *129*(4), 957–970.

Bruchmann, S., Muthukumarasamy, U., Pohl, S., Preusse, M., Bielecka, A., Nicolai, T., … Häussler, S. (2015). Deep transcriptome profiling of clinical klebsiella pneumoniae isolates reveals strain and sequence type-specific adaptation. *Environ. Microbiol. 17*(11), 4690–4710.

Brueggemann, A. B., Griffiths, D. T., Meats, E., Peto, T., Crook, D. W. & Spratt, B. G. (2003). Clonal relationships between invasive and carriage streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. *J. Infect. Dis. 187*(9), 1424–1432.

Bryan, J. P., de Silva, H. R., Tavares, A., Rocha, H. & Scheld, W. M. (1990). Etiology and mortality of bacterial meningitis in northeastern brazil. *Rev. Infect. Dis. 12*(1), 128–135.

Bryant, J. M., Grogono, D. M., Greaves, D., Foweraker, J., Roddick, I., Inns, T., … Floto, R. A. (2013). Whole-genome sequencing to identify transmission of mycobacterium abscessus between patients with cystic fibrosis: A retrospective cohort study. *Lancet*, *381*(9877), 1551–1560.

Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with scoary. *Genome Biol. 17*(1), 238.

Bucci, C., Lavitola, A., Salvatore, P., Del Giudice, L., Massardo, D. R., Bruni, C. B. & Alifano, P. (1999). Hypermutation in pathogenic bacteria. *Mol. Cell*, *3*(4), 435–445.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, … Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet. 47*(3), 291–295.

Burgner, D., Jamieson, S. E. & Blackwell, J. M. (2006). Genetic susceptibility to infectious diseases: Big is beautiful, but will bigger be even better? *Lancet Infect. Dis. 6*(10), 653–663.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., … Compston, A. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678.

Bush, W. S., Oetjens, M. T. & Crawford, D. C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet. 17*(3), 129–145.

Cartwright, R. a. (2005). DNA assembly with gaps (dawg): Simulating sequence evolution. *Bioinformatics*, *21*(SUPPL. 3), 31–38.

Casanova, J.-L. (2015). Severe infectious diseases of childhood as monogenic inborn errors of immunity. *Proc. Natl. Acad. Sci. U. S. A. 112*(51), E7128–37.

Caugant, D. A., Hoiby, E. A., Magnus, P., Scheel, O., Hoel, T., Bjune, G., … Froholm, L. O. (1994). Asymptomatic carriage of neisseria meningitidis in a randomly sampled population. *J. Clin. Microbiol. 32*(2), 323–330.

Cerutti, A. & Rescigno, M. (2008). The biology of intestinal immunoglobulin a responses. *Immunity*, *28*(6), 740–750.

Chaguza, C., Andam, C. P., Harris, S. R., Cornick, J. E., Yang, M., Bricio-Moreno, L., … Hanage, W. P. (2016). Recombination in streptococcus pneumoniae lineages increase with carriage duration and size of the polysaccharide capsule. *MBio*, *7*(5).

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), 7.

Chapman, S. J. & Hill, A. V. S. (2012). Human genetic susceptibility to infectious disease. *Nat. Rev. Genet. 13*(3), 175–188.

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M. & Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol. 26*(7), 1523–1531.

Chen, P. E. & Shapiro, B. J. (2015). The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol. 25*, 17–24.

Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol. 30*(5), 1224–1228.

Cheng, Q., Finkel, D. & Hostetter, M. K. (2000). Novel purification scheme and functions for a c3-binding protein from streptococcus pneumoniae. *Biochemistry*, *39*(18), 5450–5457.

Chengsong, Z. & Jianming, Y. (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics*, *182*(3), 875–888.

Chewapreecha, C., Harris, S. R., Croucher, N. J., Turner, C., Marttinen, P., Cheng, L., … Bentley, S. D. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet. 46*(3), 305–309.

Chewapreecha, C., Marttinen, P., Croucher, N. J., Salter, S. J., Harris, S. R., Mather, A. E., … Parkhill, J. (2014). Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet. 10*(8), e1004547.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, *7*(10), e46688.

Christensen, H., Trotter, C. L., Hickman, M. & John Edmunds, W. (2014). Re-evaluating cost effectiveness of universal meningitis vaccination (bexsero) in england: Modelling study. *BMJ*, *349*, g5725.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Lu, X. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118 ; iso-2; iso-3. *Fly*, *6*(2), 1–13.

Cleary, D. W., Devine, V. T., Jefferies, J. M. C., Webb, J. S., Bentley, S. D., Gladstone, R. A., … Clarke, S. C. (2016). Comparative genomics of carriage and disease isolates of *Streptococcus pneumoniae* serotype 22F reveals Lineage-Specific divergence and niche adaptation. *Genome Biol. Evol. 8*(4), 1243–1251.

Cleverley, R. M., Barrett, J. R., Baslé, A., Bui, N. K., Hewitt, L., Solovyova, A., … Lewis, R. J. (2014). Structure and function of a spectrin-like regulator of bacterial cytokinesis. *Nat. Commun. 5*, 5421.

Cobey, S., Baskerville, E. B., Colijn, C., Hanage, W., Fraser, C. & Lipsitch, M. (2017). *Host population structure and treatment frequency maintain balancing selection on drug resistance*.

Cobey, S. & Lipsitch, M. (2012). Niche and neutral effects of acquired immunity permit coexistence of pneumococcal serotypes. *Science*, *335*(6074), 1376–1380.

Cohen, C., Moyes, J., Tempia, S., Groom, M., Walaza, S., Pretorius, M., … Madhi, S. A. (2013). Severe influenza-associated respiratory infection in high HIV prevalence setting, south africa, 2009–2011. *Emerging Infectious Disease journal*, *19*(11), 1766.

Collins, C. & Didelot, X. (2017). *A phylogenetic method to perform Genome-Wide association studies in microbes that accounts for population structure and recombination*.

Compeau, P. E. C., Pevzner, P. a. & Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nat. Biotechnol. 29*(11), 987–991.

Cortes, A., Dendrou, C., Motyer, A., Jostins, L., Vukcevic, D., Dilthey, A., … McVean, G. (2017). *Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK biobank*.

Craig, A., Mai, J., Cai, S. & Jeyaseelan, S. (2009). Neutrophil recruitment to the lungs during bacterial pneumonia. *Infect. Immun. 77*(2), 568–575.

Crain, M. J., Waltman, W. D., Turner, J. S., Yother, J., Talkington, D. F., McDaniel, L. S., … Briles, D. E. (1990). Pneumococcal surface protein a (PspA) is serologically highly variable and is expressed by all clinically important capsular serotypes of streptococcus pneumoniae. *Infect. Immun. 58*(10), 3293–3299.

Cremers, A. J., Zomer, A. L., Gritzfeld, J. F., Ferwerda, G., van Hijum, S. A., Ferreira, D. M., … Hermans, P. W. (2014). The adult nasopharyngeal microbiome as a determinant of pneumococcal acquisition. *Microbiome*, *2*(1), 44.

Cron, L. E., Stol, K., Burghout, P., van Selm, S., Simonetti, E. R., Bootsma, H. J. & Hermans, P. W. M. (2011). Two DHH subfamily 1 proteins contribute to pneumococcal virulence and confer protection against pneumococcal disease. *Infect. Immun. 79*(9), 3697–3710.

Croucher, N. J., Campo, J. J., Le, T. Q., Liang, X., Bentley, S. D., Hanage, W. P. & Lipsitch, M. (2017). Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proc. Natl. Acad. Sci. U. S. A.*

Croucher, N. J., Coupland, P. G., Stevenson, A. E., Callendrello, A., Bentley, S. D. & Hanage, W. P. (2014). Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun. 5*, 5471.

Croucher, N. J., Finkelstein, J. a., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., … Lipsitch, M. (2013). Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet. 45*(6), 656–663.

Croucher, N. J., Hanage, W. P., Harris, S. R., McGee, L., van der Linden, M., de Lencastre, H., … Bentley, S. D. (2014). Variable recombination dynamics during the

emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. *BMC Biol. 12*(1), 49.

Croucher, N. J., Harris, S. R., Barquist, L., Parkhill, J. & Bentley, S. D. (2012). A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog. 8*(6).

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. a., Burton, J., van der Linden, M., … Bentley, S. D. (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science*, *331*(6016), 430–434.

Croucher, N. J., Kagedan, L., Thompson, C. M., Parkhill, J., Bentley, S. D., Finkelstein, J. A., … Hanage, W. P. (2015). Selective and genetic constraints on pneumococcal serotype switching. *PLoS Genet. 11*(3), 1–21.

Croucher, N. J., Mitchell, A. M., Gould, K. a., Inverarity, D., Barquist, L., Feltwell, T., … Bentley, S. D. (2013). Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. *PLoS Genet. 9*(10), e1003868.

Croucher, N. J., Mostowy, R., Wymant, C., Turner, P., Bentley, S. D. & Fraser, C. (2016). Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol. 14*(3), e1002394.

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., … Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic Acids Res. 43*(3), e15–e15.

Croucher, N. J., Vernikos, G. S., Parkhill, J. & Bentley, S. D. (2011). Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genomics*, *12*, 120.

Croucher, N. J., Walker, D., Romero, P., Lennard, N., Paterson, G. K., Bason, N. C., … Mitchell, T. J. (2009). Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone streptococcus pneumoniaeSpain23F ST81. *J. Bacteriol. 191*(5), 1480–1489.

Curtis, J., Luo, Y., Zenner, H. L., Cuchet-Lourenço, D., Wu, C., Lo, K., … Nejentsev, S. (2015). Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat. Genet. 47*(5), 523–527.

Dagan, R., Givon-Lavi, N., Zamir, O., Sikuler-Cohen, M., Guy, L., Janco, J., … Fraser, D. (2002). Reduction of nasopharyngeal carriage of streptococcus pneumoniae after administration of a 9-valent pneumococcal conjugate vaccine to toddlers attending day care centers. *J. Infect. Dis. 185*(7), 927–936.

Dalquen, D. a., Anisimova, M., Gonnet, G. H. & Dessimoz, C. (2012). ALF–a simulation framework for genome evolution. *Mol. Biol. Evol. 29*(4), 1115–1123.

Das, S. [Sayantan], Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat. Genet. 48*(10), 1284–1287.

Das, S. [Sudip], Lindemann, C., Young, B. C., Muller, J., Österreich, B., Ternette, N., … Fraunholz, M. J. (2016). Natural mutations in a staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. *Proceedings of the National Academy of Sciences*, *113*(22), E3101–E3110.

Dave, S., Brooks-Walter, A., Pangburn, M. K. & McDaniel, L. S. (2001). PspC, a pneumococcal surface protein, binds human factor H. *Infect. Immun. 69*(5), 3435–3437.

Davenport, E. E., Burnham, K. L., Radhakrishnan, J., Humburg, P., Hutton, P., Mills, T. C., … Knight, J. C. (2016). Genomic landscape of the individual host response and outcomes in sepsis: A prospective cohort study. *Lancet Respir Med*, *4*(4), 259–271.

Davey Smith, G. & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet. 23*(R1), R89–98.

Davila, S., Wright, V. J., Khor, C. C., Sim, K. S., Binder, A., Breunis, W. B., … Hibberd, M. L. (2010). Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nat. Genet. 42*(9), 772–776.

Dawid, S., Roche, A. M. & Weiser, J. N. (2007). The blp bacteriocins of streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo. *Infect. Immun. 75*(1), 443–451.

De Chiara, M., Hood, D., Muzzi, A., Pickard, D. J., Perkins, T., Pizza, M., … Donati, C. (2014). Genome sequencing of disease and carriage isolates of nontypeable haemophilus influenzae identifies discrete population structure. *Proc. Natl. Acad. Sci. U. S. A. 111*(14), 5439–5444.

de Gans, J., van de Beek, D. & European Dexamethasone in Adulthood Bacterial Meningitis Study Investigators. (2002). Dexamethasone in adults with bacterial meningitis. *N. Engl. J. Med. 347*(20), 1549–1556.

de Lange, K. M. & Barrett, J. C. (2015). Understanding inflammatory bowel disease via immunogenetics. *J. Autoimmun.*

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet. 49*(2), 256–261.

DeBardeleben, H. K., Lysenko, E. S., Dalia, A. B. & Weiser, J. N. (2014). Tolerance of a phage element by streptococcus pneumoniae leads to a fitness defect during colonization. *J. Bacteriol. 196*(14), 2670–2680.

Deininger, S., Figueroa-Perez, I., Sigel, S., Stadelmaier, A., Schmidt, R. R., Hartung, T. & von Aulock, S. (2007). Use of synthetic derivatives to determine the minimal active structure of cytokine-inducing lipoteichoic acid. *Clin. Vaccine Immunol. 14*(12), 1629–1633.

del Amo, E., Selva, L., de Sevilla, M. F., Ciruela, P., Brotons, P., Triviño, M., … Muñoz-Almagro, C. (2015). Estimation of the invasive disease potential of streptococcus pneumoniae in children by the use of direct capsular typing in clinical specimens. *Eur. J. Clin. Microbiol. Infect. Dis. 34*(4), 705–711.

Delaneau, O., Zagury, J.-F. & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, *10*(1), 5–6.

Delany, I., Grifantini, R., Bartolini, E., Rappuoli, R. & Scarlato, V. (2006). Effect of neisseria meningitidis fur mutations on global control of gene transcription. *J. Bacteriol. 188*(7), 2483–2492.

Denapaite, D., Brückner, R., Nuhn, M., Reichmann, P., Henrich, B., Maurer, P., … Hakenbeck, R. (2010). The genome of streptococcus mitis B6 - what is a commensal? *PLoS One*, *5*(2).

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., … Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol. 31*(12), 1102–1110.

Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., … Pym, A. S. (2016). Genomic and functional analyses of mycobacterium tuberculosis strains implicate ald in d-cycloserine resistance. *Nat. Genet. 48*(5).

Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. (2016). Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.*

Didelot, X. & Wilson, D. J. (2015). ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol. 11*(2), e1004041.

Dillard, J. P., Vandersea, M. W. & Yother, J. (1995). Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in streptococcus pneumoniae. *J. Exp. Med. 181*(3), 973–983.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., … Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, *455*(7216), 1069–1075.

Domenech, M., Garcıa, E. & Moscoso, M. (2012). Biofilm formation in streptococcus pneumoniae. *Microb. Biotechnol. 5*(4), 455–465.

Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N. J., Angiuoli, S. V., … Masignani, V. (2010). Structure and dynamics of the pan-genome of streptococcus pneumoniae and closely related species. *Genome Biol. 11*(10), R107.

Dubnau, D. (1999). DNA uptake in bacteria. *Annu. Rev. Microbiol. 53*(1), 217–244.

Duplessis, M. & Moineau, S. (2001). Identification of a genetic determinant responsible for host specificity in streptococcus thermophilus bacteriophages. *Mol. Microbiol. 41*(2), 325–336.

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*, *30*(9), 1266–1272.

Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., … Wilson, D. J. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, (April), 16041.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res. 32*(5), 1792–1797.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Stat. 32*(2), 407–499.

Ehrlich, G. D., Ahmed, A., Earl, J., Hiller, N. L., Costerton, J. W., Stoodley, P., … Hu, F. Z. (2010). The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes. *FEMS Immunol. Med. Microbiol. 59*(3), 269–279.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet. 11*(6), 446–450.

Enright, M. C. & Spratt, B. G. (1998). A multilocus sequence typing scheme for streptococcus pneumoniae: Identification of clones associated with serious invasive disease. *Microbiology*, *144*(11), 3049–3060.

Enright, M. C. & Spratt, B. G. (1999). Extensive variation in the ddl gene of penicillin-resistant streptococcus pneumoniae results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. *Mol. Biol. Evol. 16*(12), 1687–1695.

Evans, L., Tahmasbi, R., Vrieze, S., Abecasis, G., Das, S., Bjelland, D., … Keller, M. (2017). *Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits*.

Eyre, D. W., Cule, M. L., Wilson, D. J., Griffiths, D., Vaughan, A., O'Connor, L., … Walker, A. S. (2013). Diverse sources of c. difficile infection identified on Whole-Genome sequencing. *N. Engl. J. Med. 369*(13), 1195–1205.

Fagarasan, S. & Honjo, T. (2003). Intestinal IgA synthesis: Regulation of front-line body defences. *Nat. Rev. Immunol. 3*(1), 63–72.

Falush, D. & Bowden, R. (2006). Genome-wide association mapping in bacteria? *Trends Microbiol. 14*(8), 353–355.

Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., … Murray, M. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis. *Nat. Genet. 45*(10), 1183–1189.

Farhat, M. R., Shapiro, B., Sheppard, S. K., Colijn, C. & Murray, M. (2014). A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome Med. 6*(11), 101.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., … Goldstein, D. B. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, *317*(5840), 944–947.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat. 125*(1), 1–15.

Ferrándiz, M. J., Ardanuy, C., Liñares, J., Garcıa-Arenzana, J. M., Cercenado, E., Fleites, A., … Spanish Pneumococcal Infection Study Network. (2005). New mutations and horizontal transfer of rpob among rifampin-resistant streptococcus pneumoniae from four spanish hospitals. *Antimicrob. Agents Chemother. 49*(6), 2237–2245.

Ferreira, R. C., Pan-Hammarström, Q., Graham, R. R., Gateva, V., Fontán, G., Lee, A. T., … Hammarström, L. (2010). Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat. Genet. 42*(9), 777–780.

Fisher, R. A. (1919). XV.—The correlation between relatives on the supposition of mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb. 52*(2), 399–433.

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., … Fortune, S. M. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet. 45*(7), 784–790.

Forney, G. D. (1973). The viterbi algorithm. *Proc. IEEE*, *61*(3), 268–278.

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat. Genet. 42*(12), 1118–1125.

Franke, L. & Jansen, R. C. (2009). eQTL analysis in humans. *Methods Mol. Biol. 573*, 311–328.

Fransen, F., Heckenberg, S. G. B., Hamstra, H. J., Feller, M., Boog, C. J. P., van Putten, J. P. M., … van der Ley, P. (2009). Naturally occurring lipid a mutants in neisseria meningitidis from patients with invasive meningococcal disease are associated with reduced coagulopathy. *PLoS Pathog. 5*(4), e1000396.

Fraser, C., Lythgoe, K., Leventhal, G. E., Shirreff, G., Hollingsworth, T. D., Alizon, S. & Bonhoeffer, S. (2014). Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective. *Science*, *343*(6177), 1243727.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw. 33*(1), 1–22.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., … McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes. *Nature*, *536*(7614), 41–47.

Fusi, N., Lippert, C., Lawrence, N. D. & Stegle, O. (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat. Commun. 5*(May), 4890.

Gang, T. B., Hanley, G. A. & Agrawal, A. (2015). C-reactive protein protects mice against pneumococcal infection via both phosphocholine-dependent and phosphocholine-independent mechanisms. *Infect. Immun. 83*(5), 1845–1852.

Ganna, A., Genovese, G., Howrigan, D. P., Byrnes, A., Kurki, M. I., Zekavat, S. M., … Neale, B. M. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci. 19*(12), 1563–1565.

Gardner, S. N. & Hall, B. G. (2013). When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS One*, *8*(12), e81760.

Garland & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat. 155*(3), 346–364.

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv: 1207.3907 [q-bio.GN]

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol. 14*(7), 685–695.

Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. (2017). Phenome-wide heritability analysis of the UK biobank. *PLoS Genet. 13*(4), e1006711.

Gerlini, A., Colomba, L., Furi, L., Braccini, T., Manso, A. S., Pammolli, A., … Oggioni, M. R. (2014). The role of host and microbial factors in the pathogenesis of pneumococcal bacteraemia arising from a single bacterial cell bottleneck. *PLoS Pathog. 10*(3).

Gillberg, J., Marttinen, P., Pirinen, M., Kangas, A. J., Soininen, P., Ali, M., … Kaski, S. (2016). Multiple output regression with latent noise. *J. Mach. Learn. Res. 17*(122), 1–35.

Ginsberg, L. (2004). Difficult and recurrent meningitis. *J. Neurol. Neurosurg. Psychiatry*, *75*(suppl 1), i16–i21.

Glover, D. T., Hollingshead, S. K. & Briles, D. E. (2008). Streptococcus pneumoniae surface protein PcpA elicits protection against lung infection and fatal sepsis. *Infect. Immun. 76*(6), 2767–2776.

Gog, S., Beller, T., Moffat, A. & Petri, M. (2014). From theory to practice: Plug and play with succinct data structures. In J. Gudmundsson & J. Katajainen (Eds.), *Experimental algorithms SE - 28* (Vol. 8504, pp. 326–337). Lecture Notes in Computer Science. Springer International Publishing.

Gouy, M., Guindon, S. & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol. 27*(2), 221–224.

Gregory, S. G., Barlow, K. F., McLay, K. E., Kaul, R., Swarbreck, D., Dunham, A., … Prigmore, E. (2006). The DNA sequence and biological annotation of human chromosome 1. *Nature*, *441*(7091), 315–321.

Griffiths, N. J., Hill, D. J., Borodina, E., Sessions, R. B., Devos, N. I., Feron, C. M., … Virji, M. (2011). Meningococcal surface fibril (msf) binds to activated vitronectin and inhibits the terminal complement pathway to increase serum resistance. *Mol. Microbiol. 82*(5), 1129–1149.

Gripenland, J., Netterling, S., Loh, E., Tiensuu, T., Toledo-Arana, A. & Johansson, J. (2010). RNAs: Regulators of bacterial virulence. *Nat. Rev. Microbiol. 8*(12), 857–866.

Gritzfeld, J. F., Cremers, A. J. H., Ferwerda, G., Ferreira, D. M., Kadioglu, A., Hermans, P. W. M. & Gordon, S. B. (2014). Density and duration of experimental human pneumococcal carriage. *Clin. Microbiol. Infect. 20*(12), O1145–51.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., … Sakaguchi, A. Y. (1983). A polymorphic DNA marker genetically linked to huntington's disease. *Nature*, *306*(5940), 234–238.

Habets, M. G. J. L., Rozen, D. E. & Brockhurst, M. a. (2012). Variation in streptococcus pneumoniae susceptibility to human antimicrobial peptides may mediate intraspecific competition. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1743), 3803–3811.

Hammitt, L. L., Bruden, D. L., Butler, J. C., Baggett, H. C., Hurlburt, D. A., Reasonover, A. & Hennessy, T. W. (2006). Indirect effect of conjugate vaccine on adult carriage of streptococcus pneumoniae: An explanation of trends in invasive pneumococcal disease. *J. Infect. Dis. 193*(11), 1487–1494.

Hanage, W. P., Fraser, C., Tang, J., Connor, T. R. & Corander, J. (2009). Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science*, *324*(5933), 1454–1457.

Hanage, W. P., Kaijalainen, T., Herva, E., Saukkoriipi, A., Syrjänen, R. & Spratt, B. G. (2005). Using multilocus sequence data to define the pneumococcus. *J. Bacteriol. 187*(17), 6223–6230.

Hardin, G. (1960). The competitive exclusion principle. *Science*, *131*(3409), 1292–1297.

Harpaz, R., Dahl, R. & Dooling, K. (2016). The prevalence of immunocompromised adults: United states, 2013. *Open Forum Infect Dis*, *3*(suppl_1).

Harvey, R. M., Ogunniyi, A. D., Chen, A. Y. & Paton, J. C. (2011). Pneumolysin with low hemolytic activity confers an early growth advantage to streptococcus pneumoniae in the blood. *Infect. Immun. 79*(10), 4122–4130.

Hasbun, R., Abrahams, J., Jekel, J. & Quagliarello, V. J. (2001). Computed tomography of the head before lumbar puncture in adults with suspected meningitis. *N. Engl. J. Med. 345*(24), 1727–1733.

Hathaway, L. J., Brugger, S. D., Morand, B., Bangert, M., Rotzetter, J. U., Hauser, C., … Mühlemann, K. (2012). Capsule type of streptococcus pneumoniae determines growth phenotype. *PLoS Pathog. 8*(3), e1002574.

Haubold, B., Klötzl, F. & Pfaffelhuber, P. (2015). Andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, *31*(8), 1169–1175.

Hausdorff, W. P., Bryant, J., Paradiso, P. R. & Siber, G. R. (2000). Which pneumococcal serogroups cause the most invasive disease: Implications for conjugate vaccine formulation and use, part I. *Clin. Infect. Dis. 30*(1), 100–121.

Hava, D. L. & Camilli, A. (2002). Large-scale identification of serotype 4 streptococcus pneumoniae virulence factors. *Mol. Microbiol. 45*(5), 1389–1406.

Hebiri, M. & Lederer, J. C. (2012). How correlations influence lasso prediction. arXiv: 1204.1605 [math.ST]

Heckenberg, S. G. B., Brouwer, M. C., van der Ende, A. & van de Beek, D. (2012). Adjunctive dexamethasone in adults with meningococcal meningitis. *Neurology*, *79*(15), 1563–1569.

Heinze, G. & Ploner, M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Comput. Methods Programs Biomed. 71*(2), 181–187.

Herbert, A. P., Makou, E., Chen, Z. A., Kerr, H., Richards, A., Rappsilber, J. & Barlow, P. N. (2015). Complement evasion mediated by enhancement of captured factor h: Implications for protection of Self-Surfaces from complement. *J. Immunol. 195*(10), 4986–4998.

Hill, A. (2012). Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Philos. Trans. R. Soc. Lond. B Biol. Sci. 367*(1590), 840–849.

Hill, P. C., Townend, J., Antonio, M., Akisanya, B., Ebruke, C., Lahai, G., … Adegbola, R. A. (2010). Transmission of streptococcus pneumoniae in rural gambian villages: A longitudinal study. *Clin. Infect. Dis. 50*(11), 1468–1476.

Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R., Yu, S., Powell, E., … Hu, F. Z. (2007). Comparative genomic analyses of seventeen streptococcus pneumoniae strains: Insights into the pneumococcal supragenome. *J. Bacteriol. 189*(22), 8186–8195.

Hirschhorn, J. N. & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet. 6*(2), 95–108.

Hirst, R. A., Kadioglu, A., O'Callaghan, C. & Andrew, P. W. (2004). The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clin. Exp. Immunol. 138*(2), 195–201.

Högberg, L., Geli, P., Ringberg, H., Melander, E., Lipsitch, M. & Ekdahl, K. (2007). Age- and serogroup-related differences in observed durations of nasopharyngeal carriage of penicillin-resistant pneumococci. *J. Clin. Microbiol. 45*(3), 948–952.

Hollingshead, S. K., Becker, R. & Briles, D. E. (2000). Diversity of PspA: Mosaic genes and evidence for past recombination in streptococcus pneumoniae. *Infect. Immun. 68*(10), 5889–5900.

Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., Dance, D., … Thomson, N. R. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A. 112*(27), E3574–81.

Hoskins, J., Alborn, W. E., Arnold, J., Blaszczak, L. C., Burgett, S., DeHoff, B. S., … States, U. (2001). Genome of the bacterium streptococcus pneumoniae strain R6. *J. Bacteriol. 183*(19), 5709–5717.

Hosmer, D. W., Jr., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Howden, B. P., Gladman, S. L., Stinear, T. P., Tobias, N. J., Monk, I. R., Seemann, T. & Gao, W. (2015). Large tandem chromosome expansions facilitate niche adaptation during persistent infection with drug-resistant staphylococcus aureus. *Microbial Genomics*, *1*(1), 1–13.

Howie, B. N., Donnelly, P. & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet. 5*(6), e1000529.

Howie, B., Marchini, J. & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3*, *1*(6), 457–470.

Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., … Fan, W. (2012). pIRS: Profile-based illumina pair-end reads simulator. *Bioinformatics*, *28*(11), 1533–1535.

Hung, M.-C. & Christodoulides, M. (2013). The biology of neisseria adhesins. *Biology*, *2*(3), 1054–1109.

Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A. & Harris, S. R. (2017). *ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads*.

Hyams, C., Camberlein, E., Cohen, J. M., Bax, K. & Brown, J. S. (2010). The streptococcus pneumoniae capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. *Infect. Immun. 78*(2), 704–715.

Hyams, C., Trzcinski, K., Camberlein, E., Weinberger, D. M., Chimalapati, S., Noursadeghi, M., … Brown, J. S. (2013). Streptococcus pneumoniae capsular serotype invasiveness correlates with the degree of factor H binding and opsonization with C3b/iC3b. *Infect. Immun. 81*(1), 354–363.

Iannelli, F., Oggioni, M. R. & Pozzi, G. (2002). Allelic variation in the highly polymorphic locus pspc of streptococcus pneumoniae. *Gene*, *284*(1-2), 63–71.

Imöhl, M., Reinert, R. R., Ocklenburg, C. & van der Linden, M. (2010). Association of serotypes of streptococcus pneumoniae with age in invasive pneumococcal disease. *J. Clin. Microbiol. 48*(4), 1291–1296.

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., … Holt, K. E. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med. 6*(11), 90.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320.

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., … Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861.

International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., … Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752.

Iovino, F., Engelen-Lee, J.-Y., Brouwer, M., van de Beek, D., van der Ende, A., Valls Seron, M., … Henriques-Normark, B. (2017). pIgR and PECAM-1 bind to pneumococcal adhesins RrgA and PspC mediating bacterial brain invasion. *J. Exp. Med.*

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. (2012). De novo assembly and genotyping of variants using colored de bruijn graphs. *Nat. Genet. 44*(2), 226–232.

Jackson, C. H. (2011). Multi-State models for panel data: The msm package for R. *J. Stat. Softw. 38*(8), 1–28.

Jafri, R. Z., Ali, A., Messonnier, N. E., Tevi-Benissan, C., Durrheim, D., Eskola, J., … Abramson, J. (2013). Global epidemiology of invasive meningococcal disease. *Popul. Health Metr. 11*(1), 17.

Jaillard, M., Tournoud, M., Lima, L., Lacroix, V., Veyrieras, J.-B. & Jacob, L. (2017). *Representing genetic determinants in bacterial GWAS with compacted de bruijn graphs*.

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., … Malaria Genomic Epidemiology Network. (2009). Genome-wide and fine-resolution association analysis of malaria in west africa. *Nat. Genet. 41*(6), 657–665.

Janoff, E. N., Fasching, C., Orenstein, J. M., Rubins, J. B., Opstad, N. L. & Dalmasso, A. P. (1999). Killing of streptococcus pneumoniae by capsular polysaccharide-specific polymeric IgA, complement, and phagocytes. *J. Clin. Invest. 104*(8), 1139–1147.

Janulczyk, R., Iannelli, F., Sjoholm, A. G., Pozzi, G. & Bjorck, L. (2000). Hic, a novel surface protein of streptococcus pneumoniae that interferes with complement function. *J. Biol. Chem. 275*(47), 37257–37263.

Jedrzejas, M. J., Lamani, E. & Becker, R. S. (2001). Characterization of selected strains of pneumococcal surface protein a. *J. Biol. Chem. 276*(35), 33121–33128.

Jennett, B. & Bond, M. (1975). Assessment of outcome after severe brain damage: A practical scale. *Lancet*, *305*(7905), 480–484.

Jepson, A. (1998). Twin studies for the analysis of heritability of infectious diseases. *Bull. Inst. Pasteur*, *96*(2), 71–81.

Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M. & Cossart, P. (2002). An RNA thermosensor controls expression of virulence genes in listeria monocytogenes. *Cell*, *110*(5), 551–561.

Jolley, K. A. & Maiden, M. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*, 595.

Jombart, T., Devillard, S. & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet. 11*(1), 94.

Jones, M. R., Simms, B. T., Lupa, M. M., Kogan, M. S. & Mizgerd, J. P. (2005). Lung NF-kappaB activation and neutrophil recruitment require IL-1 and TNF receptor signaling during pneumococcal pneumonia. *J. Immunol. 175*(11), 7530–7535.

Jorth, P., Staudinger, B. J., Wu, X., Hisert, K. B., Hayden, H., Garudathri, J., … Singh, P. K. (2015). Regional isolation drives bacterial diversification within cystic fibrosis lungs. *Cell Host Microbe*, *18*(3), 307–319.

Kadie, C. M. & Heckerman, D. (2017). *Ludicrous speed linear mixed models for Genome-Wide association studies*.

Kadioglu, A., Taylor, S., Iannelli, F., Pozzi, G., Mitchell, T. J. & Andrew, P. W. (2002). Upper and lower respiratory tract infection by streptococcus pneumoniae is affected by pneumolysin deficiency and differences in capsule type. *Infect. Immun. 70*(6), 2886–2890.

Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. (2008). The role of streptococcus pneumoniae virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol. 6*(4), 288–301.

Kapatai, G., Sheppard, C. L., Troxler, L. J., Litt, D. J., Furrer, J., Hilty, M. & Fry, N. K. (2017). Pneumococcal 23B molecular subtype identified using whole genome sequencing. *Genome Biol. Evol.*

Kelley, D. R., Schatz, M. C. & Salzberg, S. L. (2010). Quake: Quality-aware detection and correction of sequencing errors. *Genome Biol. 11*(11), R116.

Kendall, M. & Colijn, C. (2015). A tree metric using structure and length to capture distinct phylogenetic signals.

Kendall, M. & Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.* (1200), 1–5.

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., … Suerbaum, S. (2011). Helicobacter pylori genome evolution during human infection. *Proceedings of the National Academy of Sciences*, *108*(12), 5033–5038.

Kent, W. J. (2002). BLAT—The BLAST-Like alignment tool. *Genome Res. 12*(4), 656–664.

Kett, K., Brandtzaeg, P., Radl, J. & Haaijman, J. J. (1986). Different subclass distribution of IgA-producing cells in human lymphoid organs and various secretory tissues. *J. Immunol. 136*(10), 3631–3635.

Khan, M. N. & Pichichero, M. E. (2012). Vaccine candidates PhtD and PhtE of streptococcus pneumoniae are adhesins that elicit functional antibodies in humans. *Vaccine*, *30*(18), 2900–2907.

Khatib, U., van de Beek, D., Lees, J. A. & Brouwer, M. C. (2016). Adults with suspected central nervous system infection: A prospective study of diagnostic accuracy. *J. Infect.*

Khor, C. C., Chapman, S. J., Vannberg, F. O., Dunne, A., Murphy, C., Ling, E. Y., … Hill, A. V. S. (2007). A mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nat. Genet. 39*(4), 523–528.

Khor, C. C., Chau, T. N. B., Pang, J., Davila, S., Long, H. T., Ong, R. T. H., … Simmons, C. P. (2011). Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nat. Genet. 43*(11), 1139–1141.

King, D. E. (2009). Dlib-ml : A machine learning toolkit. *J. Mach. Learn. Res. 10*, 1755–1758.

King, S. J., Hippe, K. R., Gould, J. M., Bae, D., Peterson, S., Cline, R. T., … Weiser, J. N. (2004). Phase variable desialylation of host proteins that bind to streptococcus pneumoniae in vivo and protect the airway. *Mol. Microbiol. 54*(1), 159–171.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet. 46*(3), 310–315.

Kjos, M., Miller, E., Slager, J., Lake, F. B., Gericke, O., Roberts, I. S., … Veening, J.-W. (2016). Expression of streptococcus pneumoniae bacteriocins is induced by antibiotics via regulatory interplay with the competence system. *PLoS Pathog. 12*(2), e1005422.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U. & Hochreiter, S. (2012). cn.MOPS: Mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res. 40*(9), e69.

Kleckner, N. (1981). Transposable elements in prokaryotes. *Annu. Rev. Genet. 15*(1), 341–404.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., … Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389.

Klugman, K. P. (2001). Efficacy of pneumococcal conjugate vaccines and their effect on carriage and antimicrobial resistance. *Lancet Infect. Dis. 1*(2), 85–91.

Knol, M. J., Wagenvoort, G. H. J., Sanders, E. A. M., Elberse, K., Vlaminckx, B. J., Melker, H. E. d. & Ende, A. v. d. (2015). Invasive pneumococcal disease 3 years after introduction of 10-valent pneumococcal conjugate vaccine, the netherlands. *Emerging Infectious Disease journal*, *21*(11), 2040.

Knoll, M. D., Moısi, J. C., Muhib, F. B., Wonodi, C. B., Lee, E. H., Grant, L., … PneumoADIP-Sponsored Surveillance Investigators. (2009). Standardizing surveillance of pneumococcal disease. *Clin. Infect. Dis. 48 Suppl 2*, S37–48.

Ko, D. C. & Urban, T. J. (2013). Understanding human variation in infectious disease susceptibility through clinical and cellular GWAS. *PLoS Pathog. 9*(8), e1003424.

Kolaczkowska, E. & Kubes, P. (2013). Neutrophil recruitment and function in health and inflammation. *Nat. Rev. Immunol. 13*(3), 159–175.

Kono, M., Zafar, M. A., Zuniga, M., Roche, A. M., Hamaguchi, S. & Weiser, J. N. (2016). Single cell bottlenecks in the pathogenesis of streptococcus pneumoniae. *PLoS Pathog. 12*(10), e1005887.

Koopmans, M. M., Bijlsma, M. W., Brouwer, M. C., van de Beek, D. & van der Ende, A. (2017). Listeria monocytogenes meningitis in the netherlands, 1985-2014: A nationwide surveillance study. *J. Infect. 75*(1), 12–19.

Koppe, U., Suttorp, N. & Opitz, B. (2012). Recognition of streptococcus pneumoniae by the innate immune system. *Cell. Microbiol. 14*(4), 460–466.

Kosiol, C., Holmes, I. & Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol. Biol. Evol. 24*(7), 1464–1479.

Kovács, M., Halfmann, A., Fedtke, I., Heintz, M., Peschel, A., Vollmer, W., … Brückner, R. (2006). A functional dlt operon, encoding proteins required for incorporation of d-alanine in teichoic acids in gram-positive bacteria, confers resistance to cationic antimicrobial peptides in streptococcus pneumoniae. *J. Bacteriol. 188*(16), 5797–5805.

Kremer, P. H. C., Lees, J. A., Koopmans, M. M., Ferwerda, B., Arends, A. W. M., Feller, M. M., … Bentley, S. D. (2017). Benzalkonium tolerance genes and outcome in listeria monocytogenes meningitis. *Clin. Microbiol. Infect. 23*(4), 265.e1–265.e7.

Kruger, P., Saffarzadeh, M., Weber, A. N. R., Rieber, N., Radsak, M., von Bernuth, H., … Hartl, D. (2015). Neutrophils: Between host defence, immune modulation, and tissue injury. *PLoS Pathog. 11*(3), e1004651.

Kuipers, K., Gallay, C., Martınek, V., Rohde, M., Martınková, M., van der Beek, S. L., … de Jonge, M. I. (2016). Highly conserved nucleotide phosphatase essential for

membrane lipid homeostasis in streptococcus pneumoniae. *Mol. Microbiol. 101*(1), 12–26.

Kulohoma, B. W., Cornick, J. E., Chaguza, C., Yalcin, F., Harris, S. R., Gray, K. J., … Heyderman, R. S. (2015). Comparative genomic analysis of meningitis and bacteremia causing pneumococci identifies a common core genome. *Infect. Immun.* (August), IAI.00814–15.

La Scolea, L. J. & Dryja, D. (1984). Quantitation of bacteria in cerebrospinal fluid and blood of children with meningitis and its diagnostic significance. *J. Clin. Microbiol. 19*(2), 187–190.

Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., … Massey, R. C. (2014). Predicting the virulence of MRSA from its genome sequence. *Genome Res. 24*(5), 839–849.

Lambris, J. D., Ricklin, D. & Geisbrecht, B. V. (2008). Complement evasion by human pathogens. *Nat. Rev. Microbiol. 6*(2), 132–142.

Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat. Genet. 11*(3), 241–247.

Lander, E., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Lanie, J. A., Ng, W.-L., Kazmierczak, K. M., Andrzejewski, T. M., Davidsen, T. M., Wayne, K. J., … Winkler, M. E. (2007). Genome sequence of avery's virulent serotype 2 strain D39 of streptococcus pneumoniae and comparison with that of unencapsulated laboratory strain R6. *J. Bacteriol. 189*(1), 38–51.

Lee, C. J., Banks, S. D. & Li, J. P. (1991). Virulence, immunity, and vaccine related to streptococcus pneumoniae. *Crit. Rev. Microbiol. 18*(2), 89–114.

Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet. 88*(3), 294–305.

Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet. 95*(1), 5–23.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. a., … Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet. 91*, 224–237.

Lefébure, T. & Stanhope, M. J. (2007). Evolution of the core and pan-genome of streptococcus: Positive selection, recombination, and genome composition. *Genome Biol. 8*(5), R71.

Lehtinen, S., Blanquart, F., Croucher, N. J., Turner, P., Lipsitch, M. & Fraser, C. (2017). Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc. Natl. Acad. Sci. U. S. A. 114*(5), 1075–1080.

Levandowsky, M. & Winter, D. (1971). Distance between sets. *Nature*, *234*(5323), 34–35.

Levin, H. L. & Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet. 12*(9), 615–627.

Levine, O. S., Cherian, T., Hajjeh, R. & Deloria Knoll, M. (2009). Progress and future challenges in coordinated surveillance and detection of pneumococcal and hib disease in developing countries. *Clin. Infect. Dis. 48*(Supplement_2), S33–S36.

Li, B. & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet. 83*(3), 311–321.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, 3.

Li, J., Li, J.-W., Feng, Z., Wang, J., An, H., Liu, Y., … Zhang, J.-R. (2016). Epigenetic switch driven by DNA inversions dictates phase variation in streptococcus pneumoniae. *PLoS Pathog. 12*(7), e1005762.

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., … Wang, J. (2010). Building the sequence map of the human pan-genome. *Nat. Biotechnol. 28*(1), 57–63.

Li, Y., Thompson, C. M., Trzciński, K. & Lipsitch, M. (2013). Within-host selection is limited by an effective population of streptococcus pneumoniae during nasopharyngeal colonization. *Infect. Immun. 81*(12), 4534–4543.

Li, Y., Weinberger, D. M., Thompson, C. M., Trzciński, K. & Lipsitch, M. (2013). Surface charge of streptococcus pneumoniae predicts serotype distribution. *Infect. Immun. 81*(12), 4519–4524.

Liley, J., Todd, J. A. & Wallace, C. (2017). A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. *Nat. Genet. 49*(2), 310–316.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods*, *8*(10), 833–835.

Lipsitch, M. (2001). Measuring and interpreting associations between antibiotic use and penicillin resistance in streptococcus pneumoniae. *Clin. Infect. Dis. 32*, 1044–1054.

Lipsitch, M., Abdullahi, O., D'Amour, A., Xie, W., Weinberger, D. M., Tchetgen Tchetgen, E. & Scott, J. A. G. (2012). Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in kenya with a markov transition model. *Epidemiology*, *23*(4), 510–519.

Lipsitch, M., Colijn, C., Cohen, T., Hanage, W. P. & Fraser, C. (2009). No coexistence for free: Neutral null models for multistrain pathogens. *Epidemics*, *1*(1), 2–13.

Lipsitch, M. & O'Hagan, J. J. (2007). Patterns of antigenic diversity and the mechanisms that maintain them. *J. R. Soc. Interface*, *4*(16), 787–802.

Liu, J. Z. & Anderson, C. A. (2014). Genetic studies of crohn's disease: Past, present and future. *Best Practice and Research: Clinical Gastroenterology*, *28*(3), 373–386.

Llull, D., Muñoz, R., López, R. & Garcıa, E. (1999). A single gene (tts) located outside the cap locus directs the formation of streptococcus pneumoniae type 37 capsular polysaccharide. *J. Exp. Med. 190*(2), 241–252.

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014). A significance test for the lasso. *Ann. Stat. 42*(2), 413–468.

Loeffler, J. M. & Fischetti, V. A. (2006). Lysogeny of streptococcus pneumoniae with MM1 phage: Improved adherence and other phenotypic changes. *Infect. Immun. 74*(8), 4486–4495.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., … L Price, A. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet. 48*(11), 1443–1448.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., … Price, A. L. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet. 47*(3), 284–290.

Lund, E. & Henrichsen, J. (1978). Chapter XI laboratory diagnosis, serology and epidemiology of streptococcus pneumoniae. *Methods in Microbiology*, *12*, 241–262.

Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., … Anderson, C. A. (2017). Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet. 49*(2), 186–192.

Lutz, S. M., Cho, M. H., Young, K., Hersh, C. P., Castaldi, P. J., McDonald, M.-L., … COPDGene Investigators. (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of european and african ancestry. *BMC Genet. 16*, 138.

Lynch, M. & Walsh, B. (1998). *Genetics and analysis of quantitative traits* (1998 edition). Sinauer.

Lysenko, E. S., Ratner, A. J., Nelson, A. L. & Weiser, J. N. (2005). The role of innate immune responses in the outcome of interspecies competition for colonization of mucosal surfaces. *PLoS Pathog. 1*(1), e1.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., … Parkinson, H. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res. 45*(D1), D896–D901.

Magnusson, M., Tobes, R., Sancho, J. & Pareja, E. (2007). Cutting edge: Natural DNA repetitive extragenic sequences from Gram-Negative pathogens strongly stimulate TLR9. *The Journal of Immunology*, *179*(1), 31–35.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., … Salzberg, S. L. (2013). GAGE-B: An evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, *29*(14), 1718–1725.

Mahdi, L. K., Van der Hoek, M. B., Ebrahimie, E., Paton, J. C. & Ogunniyi, A. D. (2015). Characterization of pneumococcal genes involved in bloodstream invasion in a mouse model. *PLoS One*, *10*(11), e0141816.

Mai, N. T. H., Chau, T. T. H., Thwaites, G., Chuong, L. V., Sinh, D. X., Nghia, H. D. T., … Farrar, J. J. (2007). Dexamethasone in vietnamese adolescents and adults with bacterial meningitis. *N. Engl. J. Med. 357*(24), 2431–2440.

Maiden, M., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., … Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A. 95*(6), 3140–3145.

Manco, S., Hernon, F., Yesilkaya, H., Paton, J. C., Andrew, P. W. & Kadioglu, A. (2006). Pneumococcal neuraminidases a and B both have essential roles during infection of the respiratory tract and sepsis. *Infect. Immun. 74*(7), 4014–4020.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M. & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

Manso, A. S., Chai, M. H., Atack, J. M., Furi, L., De Ste Croix, M., Haigh, R., … Oggioni, M. R. (2014). A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun. 5*, 5055.

Maródi, L. (2006). Neonatal innate immunity to infectious agents. *Infect. Immun. 74*(4), 1999–2006.

Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., … Schoenhuth, A. (2016). *Computational Pan-Genomics: Status, promises and challenges*.

Martins, E. P. & Garland, T. (1991). Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. *Evolution*, *45*(3), 534–557.

Marttinen, P. & Corander, J. (2010). Efficient bayesian approach for multilocus association mapping including gene-gene interactions. *BMC Bioinformatics*, *11*, 443.

Marttinen, P., Gillberg, J., Havulinna, A., Corander, J. & Kaski, S. (2013). Genome-wide association studies with high-dimensional phenotypes. *Stat. Appl. Genet. Mol. Biol. 12*(4), 413–431.

Marttinen, P., Pirinen, M., Sarin, A.-P., Gillberg, J., Kettunen, J., Surakka, I., … Kaski, S. (2014). Assessing multivariate gene-metabolome associations with rare variants using bayesian reduced rank regression. *Bioinformatics*, *30*(14), 2026–2034.

Marvig, R. L., Sommer, L. M., Molin, S. & Johansen, H. K. (2014). Convergent evolution and adaptation of pseudomonas aeruginosa within patients with cystic fibrosis. *Nat. Genet. 47*(1), 57–64.

Maskell, J. P., Sefton, a. M. & Hall, L. M. C. (2001). Multiple mutations modulate the function of dihydrofolate reductase in trimethoprim-resistant streptococcus pneumoniae. *Antimicrob. Agents Chemother. 45*(4), 1104–1108.

Maury, M., Tsai, Y.-H., Charlier, C., Touchon, M., Chenal-Francisque, V., Leclercq, A., … Lecuit, M. (2016). Uncovering listeria monocytogenes hypervirulence by harnessing its biodiversity. *Nat. Genet. 48*(3), 308–313.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*

McCool, T. L., Cate, T. R., Moy, G. & Weiser, J. N. (2002). The immune response to pneumococcal proteins during experimental human carriage. *J. Exp. Med. 195*(3), 359–365.

McCullers, J. A. (2006). Insights into the interaction between influenza virus and pneumococcus. *Clin. Microbiol. Rev. 19*(3), 571–582.

McCulloch, C. E. (2003). Chapter 4: Generalized linear mixed models (GLMMs). In *Generalized linear mixed models* (pp. 28–33). IMS and ASA.

McInerney, J. O., McNally, A. & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, *2*, 17040.

McIntyre, P. B., O'Brien, K. L., Greenwood, B. & van de Beek, D. (2012). Effect of vaccines on bacterial meningitis worldwide. *Lancet*, *380*(9854), 1703–1711.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. (2010). Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics*, *26*(16), 2069–2070.

McNeil, L. K., Zagursky, R. J., Lin, S. L., Murphy, E., Zlotnick, G. W., Hoiseth, S. K., … Anderson, A. S. (2013). Role of factor H binding protein in neisseria meningitidis virulence and its potential as a vaccine candidate to broadly protect against meningococcal disease. *Microbiol. Mol. Biol. Rev. 77*(2), 234–252.

Melegaro, A., Choi, Y., Pebody, R. & Gay, N. (2007). Pneumococcal carriage in united kingdom families: Estimating serotype-specific transmission parameters from longitudinal data. *Am. J. Epidemiol. 166*(2), 228–235.

Melin, M., Trzciński, K., Meri, S., Käyhty, H. & Väkeväinen, M. (2010). The capsular serotype of streptococcus pneumoniae is more important than the genetic background for resistance to complement. *Infect. Immun. 78*(12), 5262–5270.

Miller, E., Kjos, M., Abrudan, M., Roberts, I. S., Veening, J.-W. & Rozen, D. (2017). *Crosstalk and eavesdropping among quorum sensing peptide signals that regulate bacteriocin production in streptococcus pneumoniae*.

Mitov, V. & Stadler, T. (2016). The heritability of pathogen traits - definitions and estimators. *biorxiv*, 1–46.

Mohedano, M. L., Overweg, K., de la Fuente, A., Reuter, M., Altabe, S., Mulholland, F., … Wells, J. M. (2005). Evidence that the essential response regulator YycF in streptococcus pneumoniae modulates expression of fatty acid biosynthesis genes and alters membrane composition. *J. Bacteriol. 187*(7), 2357–2367.

Moll, G., Ubbink-Kok, T., Hildeng-Hauge, H., Nissen-Meyer, J., Nes, I. F., Konings, W. N. & Driessen, A. J. (1996). Lactococcin G is a potassium ion-conducting, two-component bacteriocin. *J. Bacteriol. 178*(3), 600–605.

Molyneux, E., Walsh, A., Forsyth, H., Tembo, M., Mwenechanya, J., Kayira, K., … Malenga, G. (2002). Dexamethasone treatment in childhood bacterial meningitis in malawi: A randomised controlled trial. *Lancet*, *360*(9328), 211–218.

Molzen, T. E., Burghout, P., Bootsma, H. J., Brandt, C. T., Der Gaast-De Jongh, C. E. V., Eleveld, M. J., … Hermans, P. W. M. (2011). Genome-wide identification of streptococcus pneumoniae genes essential for bacterial replication during experimental meningitis. *Infect. Immun. 79*(1), 288–297.

Molzen, T. E., Burghout, P., Bootsma, H. J., Brandt, C. T., van der Gaast-de Jongh, C. E., Eleveld, M. J., … Hermans, P. W. M. (2011). Genome-wide identification of streptococcus pneumoniae genes essential for bacterial replication during experimental meningitis. *Infect. Immun. 79*(1), 288–297.

Monitoring Reports SHM. (2013). https://www.hiv-monitoring.nl/english/research/monitoringreports/. Accessed: 2017-6-1.

Mook-Kanamori, B. B., Geldhoff, M., van der Poll, T. & van de Beek, D. (2011). Pathogenesis and pathophysiology of pneumococcal meningitis. *Clin. Microbiol. Rev. 24*(3), 557–591.

Morelli, G., Didelot, X., Kusecek, B., Schwarz, S., Bahlawane, C., Falush, D., … Achtman, M. (2010). Microevolution of helicobacter pylori during prolonged infection of single hosts and within families. *PLoS Genet. 6*(7), e1001036.

Morris, A. P. & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol. 34*(2), 188–193.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet. 7*(3), 277–318.

Mostowy, R., Croucher, N. J., Andam, C. P., Corander, J., Hanage, W. P. & Marttinen, P. (2017). Efficient inference of recent and ancestral recombination within bacterial populations. *Mol. Biol. Evol. 34*(5), 1167–1182.

Moxon, E. R. & Murphy, P. A. (1978). Haemophilus influenzae bacteremia and meningitis resulting from survival of a single organism. *Proc. Natl. Acad. Sci. U. S. A. 75*(3), 1534–1536.

Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol. 4*(1), 24–33.

Musher, D. M. (1992). Infections caused by streptococcus pneumoniae: Clinical spectrum, pathogenesis, immunity, and treatment. *Clin. Infect. Dis. 14*(4), 801–807.

Mwangi, M. M., Wu, S. W., Zhou, Y., Sieradzki, K., de Lencastre, H., Richardson, P., … Tomasz, A. (2007). Tracking the in vivo evolution of multidrug resistance in staphylococcus aureus by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A. 104*(22), 9451–9456.

Nadeem Khan, M., Coleman, J. R., Vernatter, J., Varshney, A. K., Dufaud, C. & Pirofski, L.-A. (2014). An ahemolytic pneumolysin of streptococcus pneumoniae manipulates human innate and CD4+ t-cell responses and reduces resistance to colonization in mice in a serotype-independent manner. *J. Infect. Dis. 210*(10), 1658–1669.

Nebenzahl-Guimaraes, H., van Laarhoven, A., Farhat, M. R., Koeken, V. A., Mandemakers, J. J., Zomer, A., … van Soolingen, D. (2016). Transmissible mycobacterium tuberculosis strains share genetic markers and immune phenotypes. *Am. J. Respir. Crit. Care Med.*

Newman, S. C. (2003). Appendix d: Quadratic equation for the odds ratio. In *Biostatistical methods in epidemiology* (pp. 329–330). John Wiley & Sons, Inc.

Newport, M. J. & Finan, C. (2011). Genome-wide association studies and susceptibility to infectious diseases. *Brief. Funct. Genomics*, *10*(2), 98–107.

Ng, P. C. & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res. 31*(13), 3812–3814.

Nigrovic, L. E., Malley, R., Macias, C. G., Kanegaye, J. T., Moro-Sutherland, D. M., Schremmer, R. D., … American Academy of Pediatrics, Pediatric Emergency Medicine Collaborative Research Committee. (2008). Effect of antibiotic pretreatment on cerebrospinal fluid profiles of children with bacterial meningitis. *Pediatrics*, *122*(4), 726–730.

Numminen, E., Cheng, L., Gyllenberg, M. & Corander, J. (2013). Estimating the transmission dynamics of streptococcus pneumoniae from strain prevalence data. *Biometrics*, *69*(3), 748–757.

Obaro, S. K., Adegbola, R. A., Banya, W. & Greenwood, B. M. (1996). Carriage of pneumococci after pneumococcal vaccination. *Lancet*, *348*(9022), 271–272.

Obert, C., Sublett, J., Kaushal, D., Hinojosa, E., Barton, T., Tuomanen, E. I. & Orihuela, C. J. (2006). Identification of a candidate streptococcus pneumoniae core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect. Immun. 74*(8), 4766–4777.

Ochman, H., Elwyn, S. & Moran, N. A. (1999). Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences*, *96*(22), 12638–12643.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., … Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet. 10*(4), e1004234.

Ogunniyi, A. D., LeMessurier, K. S., Graham, R. M. A., Watt, J. M., Briles, D. E., Stroeher, U. H. & Paton, J. C. (2007). Contributions of pneumolysin, pneumococcal surface protein a (PspA), and PspC to pathogenicity of streptococcus pneumoniae D39 in a mouse model. *Infect. Immun. 75*(4), 1843–1851.

Oliver, W. J., Shope, T. C. & Kuhns, L. R. (2003). Fatal lumbar puncture: Fact versus fiction–an approach to a clinical dilemma. *Pediatrics*, *112*(3 Pt 1), e174–6.

Omer, H., Rose, G., Jolley, K. a., Frapy, E., Zahar, J. R., Maiden, M., … Bille, E. (2011). Genotypic and phenotypic modifications of neisseria meningitidis after an accidental human passage. *PLoS One*, *6*(2).

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol. 17*(1), 1–14.

Ott, J., Wang, J. & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet. 16*(5), 275–284.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., … Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(July), btv421.

Page, A. J., De Silva, N., Hunt, M., Quail, M. A., Parkhill, J., Harris, S. R., … Keane, J. A. (2016). Robust high-throughput prokaryote de novo assembly and improvement pipeline for illumina data. *Microbial Genomics*, *2*(8).

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zool. Scr. 26*(4), 331–348.

Park, I. H., Kim, K.-H., Andrade, A. L., Briles, D. E., McDaniel, L. S. & Nahm, M. H. (2012). Nontypeable pneumococci can be divided into multiple cps types, including one type expressing the novel gene pspk. *MBio*, *3*(3).

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D. & Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. *Genome Res. 21*(9), 1512–1528.

Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res. 27*(5), 665–676.

Paternoster, L., Evans, D. M., Aagaard Nohr, E., Holst, C., Gaborieau, V., Brennan, P., … Sørensen, T. I. A. (2011). Genome-Wide Population-Based association study of extremely overweight young adults – the GOYA study. *PLoS One*, *6*(9), e24303.

Paterson, G. K. & Mitchell, T. J. (2006). Innate immunity and the pneumococcus. *Microbiology*, *152*(Pt 2), 285–293.

Paterson, G. K., Nieminen, L., Jefferies, J. M. C. & Mitchell, T. J. (2008). PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol. Lett. 285*(2), 170–176.

Paterson, G. K. & Orihuela, C. J. (2010). Pneumococci: Immunology of the innate host response. *Respirology*, *15*(7), 1057–1063.

Pathan, N., Faust, S. N. & Levin, M. (2003). Pathophysiology of meningococcal meningitis and septicaemia. *Arch. Dis. Child. 88*(7), 601–607.

Patwa, Z. & Wahl, L. M. (2008). The fixation probability of beneficial mutations. *J. R. Soc. Interface*, *5*(28), 1279–1289.

Pericone, C. D., Overweg, K., Hermans, P. W. M. & Weiser, J. N. (2000). Inhibitory and bactericidal effects of hydrogen peroxide production by streptococcus pneumoniae on other inhabitants of the upper respiratory tract. *Infect. Immun. 68*(7), 3990–3997.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y. & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet. 48*(7), 709–717.

Piet, J. R., Geldhoff, M., Van Schaik, B. D. C., Brouwer, M. C., Valls Seron, M., Jakobs, M. E., … Van De Beek, D. (2014). Streptococcus pneumoniae arginine synthesis genes promote growth and virulence in pneumococcal meningitis. *J. Infect. Dis. 209*(11), 1781–1791.

Pletz, M. W. R., Fugit, R. V., McGee, L., Glasheen, J. J., Keller, D. L., Welte, T. & Klugman, K. P. (2006). Fluoroquinolone-resistant streptococcus pneumoniae. *Emerg. Infect. Dis. 12*(9), 1462–1463.

Plumptre, C. D., Ogunniyi, A. D. & Paton, J. C. (2013). Surface association of pht proteins of streptococcus pneumoniae. *Infect. Immun. 81*(10), 3644–3651.

Poulsen, K., Reinholdt, J. & Kilian, M. (1996). Characterization of the streptococcus pneumoniae immunoglobulin A1 protease gene (iga) and its translation product. *Infect. Immun. 64*(10), 3957–3966.

Power, R. A., Parkhill, J. & de Oliveira, T. (2016). Microbial genome-wide association studies: Lessons from human GWAS. *Nat. Rev. Genet.*

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet. 38*(8), 904–909.

Price, A. L., Zaitlen, N. A. [N A], Reich, D. & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet. 11*(7), 459–463.

Price, A. L., Zaitlen, N. A. [Noah A], Reich, D. & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet. 11*(7), 459–463.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol. 26*(7), 1641–1650.

Proulx, N., Fréchette, D., Toye, B., Chan, J. & Kravcik, S. (2005). Delays in the administration of antibiotics are associated with mortality from adult acute bacterial meningitis. *QJM*, *98*(4), 291–298.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., … Willer, C. J. (2010). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics*, *26*(18), 2336–2337.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet. 81*(3), 559–575.

Quinton, L. J., Jones, M. R., Simms, B. T., Kogan, M. S., Robson, B. E., Skerrett, S. J. & Mizgerd, J. P. (2007). Functions and regulation of NF-kappaB RelA during pneumococcal pneumonia. *J. Immunol. 178*(3), 1896–1903.

Raeder, R. & Boyle, M. D. (1993). Association between expression of immunoglobulin g-binding proteins by group a streptococci and virulence in a mouse skin infection model. *Infect. Immun. 61*(4), 1378–1384.

Raeder, R. & Boyle, M. D. (1995). Analysis of immunoglobulin g-binding-protein expression by invasive isolates of streptococcus pyogenes. *Clin. Diagn. Lab. Immunol. 2*(4), 484–486.

Ragunathan, L., Ramsay, M., Borrow, R., Guiver, M., Gray, S. & Kaczmarski, E. B. (2000). Clinical features, laboratory findings and management of meningococcal meningitis in england and wales: Report of a 1997 survey. meningococcal meningitis: 1997 survey report. *J. Infect. 40*(1), 74–79.

Rau, M. H., Marvig, R. L., Ehrlich, G. D., Molin, S. & Jelsbak, L. (2012). Deletion and acquisition of genomic content during early stage adaptation of pseudomonas aeruginosa to a human host environment. *Environ. Microbiol. 14*(8), 2200–2211.

Rautanen, A., Pirinen, M., Mills, T. C., Rockett, K. A., Strange, A., Ndungu, A. W., … Spencer, C. C. A. (2016). Polymorphism in a lincRNA associates with a doubled risk of pneumococcal bacteremia in kenyan children. *Am. J. Hum. Genet. 2*, 1092–1100.

Read, A. F. & Nee, S. (1995). Inference from binary comparative data. *J. Theor. Biol. 173*(1), 99–108.

Read, T. D. & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: A new direction for bacteriology. *Genome Med. 6*(11), 109.

Reddy, U. R., Phatak, S. & Pleasure, D. (1996). Human neural tissues express a truncated ror1 receptor tyrosine kinase, lacking both extracellular and transmembrane domains. *Oncogene*, *13*(7), 1555–1559.

Regev-Yochay, G., Trzcinski, K., Thompson, C. M., Malley, R. & Lipsitch, M. (2006). Interference between streptococcus pneumoniae and staphylococcus aureus: In vitro hydrogen peroxide-mediated killing by streptococcus pneumoniae. *J. Bacteriol. 188*(13), 4996–5001.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., … Lander, E. (2001). Linkage disequilibrium in the human genome. *Nature*, *411*(6834), 199–204.

Revell, L. J. (2013). Two new graphical methods for mapping trait evolution on phylogenies. *Methods Ecol. Evol. 4*(8), 754–759.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*, 1516–1517.

Rizk, G., Lavenier, D. & Chikhi, R. (2013). DSK: K-mer counting with very low memory usage. *Bioinformatics*, *29*(5), 652–653.

Roberts, A. P. & Mullany, P. (2009). A modular master on the move: The tn916 family of mobile genetic elements. *Trends Microbiol. 17*(May), 251–258.

Robinson, D. A., Edwards, K. M., Waites, K. B., Briles, D. E., Crain, M. J. & Hollingshead, S. K. (2001). Clones of streptococcus pneumoniae isolated from nasopharyngeal carriage and invasive disease in young children in central tennessee. *J. Infect. Dis. 183*(10), 1501–1507.

Robinson, M. W., Buchtmann, K. A., Jenkins, C., Tacchi, J. L., Raymond, B. B. A., To, J., … Djordjevic, S. P. (2013). MHJ_0125 is an M42 glutamyl aminopeptidase that moonlights as a multifunctional adhesin on the surface of mycoplasma hyopneumoniae. *Open Biol. 3*(4), 130017.

Rodrigo, C., Bewick, T., Sheppard, C., Greenwood, S., Macgregor, V., Trotter, C., … Lim, W. S. (2014). Pneumococcal serotypes in adult non-invasive and invasive pneumonia in relation to child contact and child vaccination status. *Thorax*, *69*(2), 168–173.

Romero, P., Croucher, N. J., Hiller, N. L., Hu, F. Z., Ehrlich, G. D., Bentley, S. D., … Mitchell, T. J. (2009). Comparative genomic analysis of ten streptococcus pneumoniae temperate bacteriophages. *J. Bacteriol. 191*(15), 4854–4862.

Rosenthal, R. (1978). Combining results of independent studies. *Psychol. Bull. 85*(1), 185.

Rouphael, N. G. & Stephens, D. S. (2012). Neisseria meningitidis: Biology, microbiology, and epidemiology. *Methods Mol. Biol. 799*, 1–20.

Rubins, J. B., Paddock, A. H., Charboneau, D., Berry, A. M., Paton, J. C. & Janoff, E. N. (1998). Pneumolysin in pneumococcal adherence and colonization. *Microb. Pathog. 25*(6), 337–342.

Russell, J. E., Jolley, K. a., Feavers, I. M., Maiden, M. & Suker, J. (2004). PorA variable regions of neisseria meningitidis. *Emerg. Infect. Dis. 10*(4), 674–678.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., … International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), 928–933.

Salipante, S. J., Roach, D. J., Kitzman, J. O., Snyder, M. W., Stackhouse, B., Butler-Wu, S. M., … Shendure, J. (2015). Large-scale genomic sequencing of extraintestinal pathogenic escherichia coli strains. *Genome Res. 25*(1), 119–128.

Salter, S. J., Hinds, J., Gould, K. A., Lambertsen, L., Hanage, W. P., Antonio, M., … Bentley, S. D. (2012). Variation at the capsule locus, cps, of mistyped and non-typable streptococcus pneumoniae isolates. *Microbiology*, *158*(Pt 6), 1560–1569.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nat. Genet. 46*(9), 944–950.

Samore, M. H., Lipsitch, M., Alder, S. C., Haddadin, B., Stoddard, G., Williamson, J., … Sande, M. a. (2006). Mechanisms by which antibiotics promote dissemination of resistant pneumococci in human populations. *Am. J. Epidemiol. 163*(2), 160–170.

Sánchez-Beato, A. R., López, R. & Garcıa, J. L. (1998). Molecular characterization of PcpA: A novel choline-binding protein of streptococcus pneumoniae. *FEMS Microbiol. Lett. 164*(1), 207–214.

Sanderson, C. (2010). Armadillo: An open source c++ linear algebra library for fast prototyping and computationally intensive experiments. In *NICTA* (Vol. NICTA, pp. 1–16). Australia.

Sanderson, C. & Curtin, R. (2016). Armadillo: A template-based c++ library for linear algebra. *JOSS*, *1*(2).

Sarkari, J., Pandit, N., Moxon, E. R. & Achtman, M. (1994). Variable expression of the opc outer membrane protein in neisseria meningitidis is caused by size variation of a promoter containing poly-cytidine. *Mol. Microbiol. 13*(2), 207–217.

Schraiber, J. G. & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet. 16*(12), 727–740.

Schuchat, A., Robinson, K., Wenger, J. D., Harrison, L. H., Farley, M., Reingold, A. L., … Perkins, B. A. (1997). Bacterial meningitis in the united states in 1995. active surveillance team. *N. Engl. J. Med. 337*(14), 970–976.

Seale, A. C., Davies, M. R., Anampiu, K., Morpeth, S. C., Nyongesa, S., Mwarumba, S., … Berkley, J. A. (2016). Invasive group a streptococcus infection among children, rural kenya. *Emerging Infectious Disease journal*, *22*(2), 224.

Serruto, D., Rappuoli, R., Scarselli, M., Gros, P. & van Strijp, J. A. G. (2010). Molecular mechanisms of complement evasion: Learning from staphylococci and meningococci. *Nat. Rev. Microbiol. 8*(6), 393–399.

Seth, S., Välimäki, N., Kaski, S. & Honkela, A. (2014). Exploration and retrieval of whole-metagenome sequencing samples. *Bioinformatics*, *30*(17), 16.

Shah, T. S., Liu, J. Z., Floyd, J. a. B., Morris, J. a., Wirth, N., Barrett, J. C. & Anderson, C. a. (2012). Opticall: A robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics*, *28*(12), 1598–1603.

Shakhnovich, E. A., King, S. J. & Weiser, J. N. (2002). Neuraminidase expressed by streptococcus pneumoniae desialylates the lipopolysaccharide of neisseria meningitidis and haemophilus influenzae: A paradigm for interbacterial competition among pathogens of the human respiratory tract. *Infect. Immun. 70*(12), 7161–7164.

Shaper, M., Hollingshead, S. K., Benjamin, W. H., Jr & Briles, D. E. (2004). PspA protects streptococcus pneumoniae from killing by apolactoferrin, and antibody to PspA enhances killing of pneumococci by apolactoferrin. *Infect. Immun. 72*(9), 5031–5040.

Shapiro, E. D. & Austrian, R. (1994). Serotypes responsible for invasive streptococcus pneumoniae infections among children in connecticut. *J. Infect. Dis. 169*(1), 212–214.

Shea, P. R., Beres, S. B., Flores, A. R., Ewbank, A. L., Gonzalez-Lugo, J. H., Martagon-Rosado, A. J., … Musser, J. M. (2011). Distinct signatures of diversifying selection revealed by genome analysis of respiratory tract and invasive bacterial populations. *Proc. Natl. Acad. Sci. U. S. A. 108*(12), 5039–5044.

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., … Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in campylobacter. *Proceedings of the National Academy of Sciences*, *110*(29), 11923–11927.

Shivshankar, P., Sanchez, C., Rose, L. F. & Orihuela, C. J. (2009). The streptococcus pneumoniae adhesin PsrP binds to keratin 10 on lung cells. *Mol. Microbiol. 73*(4), 663–679.

Siddique, T., Figlewicz, D. A., Pericak-Vance, M. A., Haines, J. L., Rouleau, G., Jeffers, A. J., … McKenna-Yasek, D. (1991). Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *N. Engl. J. Med. 324*(20), 1381–1384.

Siegel, S. J., Roche, A. M. & Weiser, J. N. (2014). Influenza promotes pneumococcal growth during coinfection by providing host sialylated substrates as a nutrient source. *Cell Host Microbe*, *16*(1), 55–67.

Simpson, J. T. & Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res. 22*(3), 549–556.

Skwark, M. J., Croucher, N. J., Puranen, S., Chewapreecha, C., Pesonen, M., Xu, Y. Y., … Corander, J. (2017). Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet. 13*(2), e1006508.

Smith, E. E., Buckley, D. G., Wu, Z., Saenphimmachak, C., Hoffman, L. R., D'Argenio, D. A., … Olson, M. V. (2006). Genetic adaptation by pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences*, *103*(22), 8487–8492.

Smith, T. C., Sledjeski, D. D. & Boyle, M. D. P. (2003a). Regulation of protein H expression in M1 serotype isolates of streptococcus pyogenes. *FEMS Microbiol. Lett. 219*(1), 9–15.

Smith, T. C., Sledjeski, D. D. & Boyle, M. D. P. (2003b). Streptococcus pyogenes infection in mouse skin leads to a Time-Dependent Up-Regulation of protein H expression. *Infect. Immun. 71*(10), 6079–6082.

Snelson, E., Ghahramani, Z. & Rasmussen, C. E. (2004). Warped gaussian processes. In S. Thrun, L. K. Saul & P. B. Schölkopf (Eds.), *Advances in neural information processing systems 16* (pp. 337–344). MIT Press.

Snyder, L. A. S., Saunders, N. J. & Shafer, W. M. (2001). A putatively phase variable gene (dca) required for natural competence in neisseria gonorrhoeae but not neisseria meningitidis is located within the division cell wall (dcw) gene cluster. *J. Bacteriol. 183*(4), 1233–1241.

Snyder, L. A. S., Shafer, W. M. & Saunders, N. J. (2003). Divergence and transcriptional analysis of the division cell wall (dcw) gene cluster in neisseria spp. *Mol. Microbiol. 47*(2), 431–442.

Spain, S. L. & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Hum. Mol. Genet. 24*(R1), R111–R119.

Speed, D., Cai, N., UCLEB Consortium, Johnson, M. R., Nejentsev, S. & Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*

Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet. 91*(6), 1011–1021.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol. 64*(4), 583–639.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet. 52*(3), 506–516.

Spijkerman, J., van Gils, E. J. M., Veenhoven, R. H., Hak, E., Yzerman, E. P. F., van der Ende, A., … Sanders, E. A. M. (2011). Carriage of streptococcus pneumoniae 3 years after start of vaccination program, the netherlands. *Emerg. Infect. Dis. 17*(4), 584–591.

Spratt, B. G. (1994a). Chapter 25: Resistance to β-lactam antibiotics. *New Compr. Biochem. 27*, 517–534.

Spratt, B. G. (1994b). Resistance to antibiotics mediated by target alterations. *Science*, *264*(5157), 388–393.

Sreevatsan, S., Pan, X., Zhang, Y., Deretic, V. & Musser, J. M. (1997). Analysis of the oxyR-ahpC region in isoniazid-resistant and -susceptible mycobacterium tuberculosis complex organisms recovered from diseased humans and animals in diverse localities. *Antimicrob. Agents Chemother. 41*(3), 600–606.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Steer, A. C., Magor, G., Jenney, A. W. J., Kado, J., Good, M. F., McMillan, D., … Carapetis, J. R. (2009). Emm and c-repeat region molecular typing of beta-hemolytic streptococci in a tropical country: Implications for vaccine development. *J. Clin. Microbiol. 47*(8), 2502–2509.

Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc. 7*(3), 500–507.

Stessman, H. A. F., Xiong, B., Coe, B. P., Wang, T., Hoekzema, K., Fenckova, M., … Eichler, E. E. (2017). Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet. 49*(4), 515–526.

Stranger, B. E., Stahl, E. A. & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, *187*(2), 367–383.

Sveinbjornsson, G., Gudbjartsson, D. F., Halldorsson, B. V., Kristinsson, K. G., Gottfredsson, M., Barrett, J. C., … Stefansson, K. (2016). HLA class II sequence variants influence tuberculosis risk in populations of european ancestry. *Nat. Genet. 48*(3), 318–322.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Takahashi, H., Hirose, K. & Watanabe, H. (2004). Necessity of meningococcal gamma-glutamyl aminopeptidase for neisseria meningitidis growth in rat cerebrospinal fluid (CSF) and CSF-Like medium. *J. Bacteriol. 186*(1), 244–247.

Takashima, K., Tateda, K., Matsumoto, T., Iizawa, Y., Nakao, M. & Yamaguchi, K. (1997). Role of tumor necrosis factor alpha in pathogenesis of pneumococcal pneumonia in mice. *Infect. Immun. 65*(1), 257–260.

Tamayo, R., Pratt, J. T. & Camilli, A. (2007). Roles of cyclic diguanylate in the regulation of bacterial pathogenesis. *Annu. Rev. Microbiol. 61*, 131–148.

Tasoulis, S., Cheng, L., Valimaki, N., Croucher, N. J., Harris, S. R., Hanage, W. P., … Corander, J. (2014). Random projection based clustering for population genomics. In *2014 IEEE international conference on big data (big data)* (pp. 675–682).

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., … Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, *536*(7615), 165–170.

Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. a., Read, T. D., Peterson, S., … Fraser, C. M. (2001). Complete genome sequence of a virulent isolate of streptococcus pneumoniae. *Science*, *293*(5529), 498–506.

Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, a. C., Nelson, K. E., Eisen, J. a., … Venter, J. C. (2000). Complete genome sequence of neisseria meningitidis serogroup B strain MC58. *Science*, *287*(5459), 1809–1815.

The Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nat. Genet. 46*(8), 818–825.

Thorpe, H. A., Bayliss, S. C., Hurst, L. D. & Feil, E. J. (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, *206*(1), 363–376.

Tian, C., Hinds, D. A., Hromatka, B. S., Kiefer, A. K., Eriksson, N. & Tung, J. Y. (2016). *Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections*.

Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B Stat. Methodol. 63*(2), 411–423.

Trappetti, C., McAllister, L. J., Chen, A., Wang, H., Paton, A. W., Oggioni, M. R., … Paton, J. C. (2017). Autoinducer 2 signaling via the phosphotransferase FruA drives galactose utilization by streptococcus pneumoniae, resulting in hypervirulence. *MBio*, *8*(1).

Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol. 15*(11), 524.

Trzciński, K., Li, Y., Weinberger, D. M., Thompson, C. M., Cordy, D., Bessolo, A., … Lipsitch, M. (2015). Effect of serotype on pneumococcal competition in a mouse colonization model. *MBio*, *6*(5), e00902–15.

Tu, A. H., Fulgham, R. L., McCrory, M. A., Briles, D. E. & Szalai, A. J. (1999). Pneumococcal surface protein a inhibits complement activation by streptococcus pneumoniae. *Infect. Immun. 67*(9), 4720–4724.

Tunjungputri, R. N., Mobegi, F. M., Cremers, A. J., van der Gaast-de Jongh, C. E., Ferwerda, G., Meis, J. F., … de Jonge, M. I. (2017). Phage-Derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease. *MBio*, *8*(1).

Tunkel, A. R. & Scheld, W. M. (2002). Treatment of bacterial meningitis. *Curr. Infect. Dis. Rep. 4*(1), 7–16.

Turner, C., Turner, P., Carrara, V., Burgoine, K., Htoo, S. T. L., Watthanaworawit, W., … Nosten, F. (2013). High rates of pneumonia in children under two years of age in a south east asian refugee population. *PLoS One*, *8*(1), e54026.

Turner, P., Turner, C., Jankhot, A., Helen, N., Lee, S. J., Day, N. P., … Goldblatt, D. (2012). A longitudinal study of streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. *PLoS One*, *7*(5).

Turner, P., Turner, C., Jankhot, A., Phakaudom, K., Nosten, F. & Goldblatt, D. (2013). Field evaluation of culture plus latex sweep serotyping for detection of multiple pneumococcal serotype colonisation in infants and young children. *PLoS One*, *8*(7), 1–7.

Unemo, M. & Shafer, W. M. (2014). Antimicrobial resistance in neisseria gonorrhoeae in the 21st century: Past, evolution, and future. *Clin. Microbiol. Rev. 27*(3), 587–613.

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., … Peterlongo, P. (2014). Reference-free detection of isolated SNPs. *Nucleic Acids Res. 33*(0), 1–11.

Välimäki, N. & Puglisi, S. (2012). Distributed string mining for High-Throughput sequencing data. In B. Raphael & J. Tang (Eds.), *Algorithms in bioinformatics SE - 35* (Vol. 7534, pp. 441–452). Lecture Notes in Computer Science. Springer Berlin Heidelberg.

van de Beek, D., de Gans, J., Spanjaard, L., Weisfelt, M., Reitsma, J. B. & Vermeulen, M. (2004). Clinical features and prognostic factors in adults with bacterial meningitis. *N. Engl. J. Med. 351*(18), 1849–1859.

van de Beek, D., de Gans, J., Tunkel, A. R. & Wijdicks, E. F. M. (2006). Community-Acquired bacterial meningitis in adults. *N. Engl. J. Med. 354*(1), 44–53.

van de Beek, D., Farrar, J. J., de Gans, J., Mai, N. T. H., Molyneux, E. M., Peltola, H., … Zwinderman, A. H. (2010). Adjunctive dexamethasone in bacterial meningitis: A meta-analysis of individual patient data. *Lancet Neurol. 9*(3), 254–263.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2002). From FastQ data to High-Confidence variant calls: The genome analysis toolkit best practices pipeline. In *Current protocols in bioinformatics*. John Wiley & Sons, Inc.

van der Ende, A., Hopman, C. T., Zaat, S., Essink, B. B., Berkhout, B. & Dankert, J. (1995). Variable expression of class 1 outer membrane protein in neisseria meningitidis is caused by variation in the spacing between the -10 and -35 regions of the promoter. *J. Bacteriol. 177*(9), 2475–2480.

van der Ende, A., Hopman, C. T. P. & Dankert, J. (2000). Multiple mechanisms of phase variation of PorA in neisseria meningitidis. *Infect. Immun. 68*(12), 6685–6690.

van Es, M. a., Veldink, J. H., Saris, C. G. J., Blauw, H. M., van Vught, P. W. J., Birve, A., … van den Berg, L. H. (2009). Genome-wide association study identifies 19p13.3

(UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet. 41*(10), 1083–1087.

van Opijnen, T., Bodi, K. L. & Camilli, A. (2009). Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods*, *6*(10), 767–772.

van Veen, M. G., Presanis, A. M., Conti, S., Xiridou, M., Stengaard, A. R., Donoghoe, M. C., ... De Angelis, D. (2011). National estimate of HIV prevalence in the netherlands: Comparison and applicability of different estimation tools. *AIDS*, *25*(2), 229–237.

van Wijngaarden, J. P., Dhonukshe-Rutten, R. a. M., van Schoor, N. M., van der Velde, N., Swart, K. M. a., Enneman, A. W., ... de Groot, L. C. P. G. M. (2011). Rationale and design of the B-PROOF study, a randomized controlled trial on the effect of supplemental intake of vitamin B12 and folic acid on fracture incidence. *BMC Geriatr. 11*(1), 80.

Veyrier, F. J., Boneca, I. G., Cellier, M. F. & Taha, M. K. (2011). A novel metal transporter mediating manganese export (mntx) regulates the mn to fe intracellular ratio and neisseria meningitidis virulence. *PLoS Pathog. 7*(9).

Visscher, P. M., Hill, W. G. & Wray, N. R. (2008). Heritability in the genomics era–concepts and misconceptions. *Nat. Rev. Genet. 9*(4), 255–266.

Vitányi, P. M. B., Balbach, F. J., Cilibrasi, R. L. & Li, M. (2009). Normalized information distance. *Information Theory and Statistical Learning*, 45–82.

Walker, M. J., Barnett, T. C., McArthur, J. D., Cole, J. N., Gillen, C. M., Henningham, A., ... Nizet, V. (2014). Disease manifestations and pathogenic mechanisms of group a streptococcus. *Clin. Microbiol. Rev. 27*(2), 264–301.

Walport, M. J. (2001a). Complement. first of two parts. *N. Engl. J. Med. 344*(14), 1058–1066.

Walport, M. J. (2001b). Complement. second of two parts. *N. Engl. J. Med. 344*(15), 1140–1144.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet. 10*(1), 57–63.

Wani, J. H., Gilbert, J. V., Plaut, A. G. & Weiser, J. N. (1996). Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of streptococcus pneumoniae. *Infect. Immun. 64*(10), 3967–3974.

Ward, N. & Moreno-Hagelsieb, G. (2014). Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: How much do we miss? *PLoS One*, *9*(7), e101850.

Wartha, F., Beiter, K., Albiger, B., Fernebro, J., Zychlinsky, A., Normark, S. & Henriques-Normark, B. (2007). Capsule and d-alanylated lipoteichoic acids protect streptococ-

cus pneumoniae against neutrophil extracellular traps. *Cell. Microbiol. 9*(5), 1162–1171.

Weinberger, D. M., Dagan, R., Givon-Lavi, N., Regev-Yochay, G., Malley, R. & Lipsitch, M. (2008). Epidemiologic evidence for serotype-specific acquired immunity to pneumococcal carriage. *J. Infect. Dis. 197*(11), 1511–1518.

Weinberger, D. M., Harboe, Z. B., Flasche, S., Scott, J. A. & Lipsitch, M. (2011). Prediction of serotypes causing invasive pneumococcal disease in unvaccinated and vaccinated populations. *Epidemiology*, *22*(2), 199–207.

Weinberger, D. M., Malley, R. & Lipsitch, M. (2011). Serotype replacement in disease after pneumococcal vaccination. *Lancet*, *378*(9807), 1962–1973.

Weinberger, D. M., Trzciński, K., Lu, Y.-J., Bogaert, D., Brandes, A., Galagan, J., … Lipsitch, M. (2009). Pneumococcal capsular polysaccharide structure predicts serotype prevalence. *PLoS Pathog. 5*(6), e1000476.

Weinert, L. a., Chaudhuri, R. R., Wang, J., Peters, S. E., Corander, J., Jombart, T., … Terra, V. (2015). Genomic signatures of human and animal disease in the zoonotic pathogen streptococcus suis. *Nat. Commun. 6*, 6740.

Weisfelt, M., van de Beek, D., Spanjaard, L., Reitsma, J. B. & de Gans, J. (2006). Clinical features, complications, and outcome in adults with pneumococcal meningitis: A prospective case series. *Lancet Neurol. 5*(2), 123–129.

Willer, C. J., Li, Y. & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–2191.

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J. H., Cheesbrough, J., Gee, S., Bolton, E., … Fearnhead, P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen campylobacter jejuni. *Mol. Biol. Evol. 26*(2), 385–397.

Winantea, J., Hoang, M. N., Ohlraun, S., Rietschel, M., Cichon, S., Propping, P., … Freudenberg-Hua, Y. (2006). A summary statistic approach to sequence variation in noncoding regions of six schizophrenia-associated gene loci. *Eur. J. Hum. Genet. 14*(9), 1037–1043.

Winkler, F., Kastenbauer, S., Yousry, T. A., Maerz, U. & Pfister, H.-W. (2002). Discrepancies between brain CT imaging and severely raised intracranial pressure proven by ventriculostomy in adults with pneumococcal meningitis. *J. Neurol. 249*(9), 1292–1297.

Wise, A. L., Gyi, L. & Manolio, T. A. (2013). Exclusion: Toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet. 92*(5), 643–647.

Woehrl, B., Brouwer, M. C., Murr, C., Heckenberg, S. G. B., Baas, F., Pfister, H. W., … Van De Beek, D. (2011). Complement component 5 contributes to poor disease outcome in humans and mice with pneumococcal meningitis. *J. Clin. Invest. 121*(10), 3943–3953.

Wood, D. E. & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol. 15*(3), R46.

Wörmann, M. E., Horien, C. L., Bennett, J. S., Jolley, K. a., Maiden, M., Tang, C. M., … Exley, R. M. (2014). Sequence, distribution and chromosomal context of class I and class II pilin genes of neisseria meningitidis identified in whole genome sequences. *BMC Genomics*, *15*, 253.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of Guinea-Pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet. 89*(1), 82–93.

Wyllie, A. L., Chu, M. L. J. N., Schellens, M. H. B., van Engelsdorp Gastelaars, J., Jansen, M. D., van der Ende, A., … Trzciński, K. (2014). Streptococcus pneumoniae in saliva of dutch primary school children. *PLoS One*, *9*(7), e102045.

Wyllie, A. L., Wijmenga-Monsuur, A. J., van Houten, M. A., Bosch, A. A. T. M., Groot, J. A., van Engelsdorp Gastelaars, J., … Trzciński, K. (2016). Molecular surveillance of nasopharyngeal carriage of streptococcus pneumoniae in children vaccinated with conjugated polysaccharide pneumococcal vaccines. *Sci. Rep. 6*, 23809.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet. 42*(7), 565–569.

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet. 88*(1), 76–82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., … Visscher, P. M. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet. 43*(6), 519–525.

Yang, L., Jelsbak, L., Marvig, R. L., Damkiær, S., Workman, C. T., Rau, M. H., … Molin, S. (2011). Evolutionary dynamics of bacteria in a human host environment. *Proc. Natl. Acad. Sci. U. S. A. 108*(18), 7481–7486.

Yang, Z. [Zhirong], Corander, J. & Oja, E. (2016). Low-Rank doubly stochastic matrix decomposition for cluster analysis. *J. Mach. Learn. Res. 17*(187), 1–25.

Yang, Z. [Ziheng]. (2006). *Computational molecular evolution*. OUP Oxford.

Yesilkaya, H., Spissu, F., Carvalho, S. M., Terra, V. S., Homer, K. A., Benisty, R., … Andrew, P. W. (2009). Pyruvate formate lyase is required for pneumococcal fermentative metabolism and virulence. *Infect. Immun. 77*(12), 5418–5427.

Yother, J. (2011). Capsules of streptococcus pneumoniae and other bacteria: Paradigms for polysaccharide biosynthesis and regulation. *Annu. Rev. Microbiol. 65*, 563–581.

Young, B. C., Golubchik, T., Batty, E. M., Fung, R., Larner-Svensson, H., Votintseva, A. A., … Wilson, D. J. (2012). Evolutionary dynamics of staphylococcus aureus during progression from carriage to disease. *Proceedings of the National Academy of Sciences*, *109*(12), 4550–4555.

Zafar, M. A., Wang, Y., Hamaguchi, S. & Weiser, J. N. (2017). Host-to-Host transmission of streptococcus pneumoniae is driven by its inflammatory toxin, pneumolysin. *Cell Host Microbe*, *21*(1), 73–83.

Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D. A., Shenker, S., … Sittler, T. (2011). Faster and more accurate sequence alignment with SNAP. *CoRR*, *abs/1111.5*, 1–10.

Zerbino, D. R. & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res. 18*(5), 821–829.

Zhang, F.-R., Huang, W., Chen, S.-M., Sun, L.-D., Liu, H., Li, Y., … Liu, J.-J. (2009). Genomewide association study of leprosy. *N. Engl. J. Med. 361*(27), 2609–2618.

Zhou, X. & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet. 44*(7), 821–824.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L. & Yorke, J. a. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677.

Appendix A

Supplementary information

A.1 Data access and code availability

The following new data generated as part of this work is available publicly:

- *S. pyogenes* sequence reads from section 2.6.3 are available on the European Nucleotide Archive under study accession IDs PRJEB2839 (isolates from Fiji) and PRJEB3313 (isolates from Kilifi).
- From the paired blood and CSF isolates in section 4.5 read data, assembled and annotated contigs were deposited in the European Nucleotide Archive (ENA): study accession number ERP004245.
- Sample metadata used from these paired blood and CSF isolates has been deposited in Figshare (DOI: 10.6084/m9.figshare.4329809).

Relevant code for each section can be found on github:

- Testing of tree inference methods, section 2.3.1: https://github.com/johnlees/which_tree
- SEER, section 2.5: https://github.com/johnlees/seer
- Carriage duration analysis and results, chapter 3: https://github.com/johnlees/carriage-duration
- Paired sample analysis, section 4.5: https://github.com/johnlees/paired-samples
- Calculation of Tajima's D, section 4.4.2: https://github.com/johnlees/tajima-D
- Fix to subtest code, section 5.2.2: https://github.com/johnlees/subtest
- Code to perform all-by-all variant association in genome-to-genome analysis, section 5.3.1: https://github.com/johnlees/epistasis-code
- Miscellaneous code and scripts, referred to throughout: https://github.com/johnlees/bioinformatics

A.2 Supplementary figures

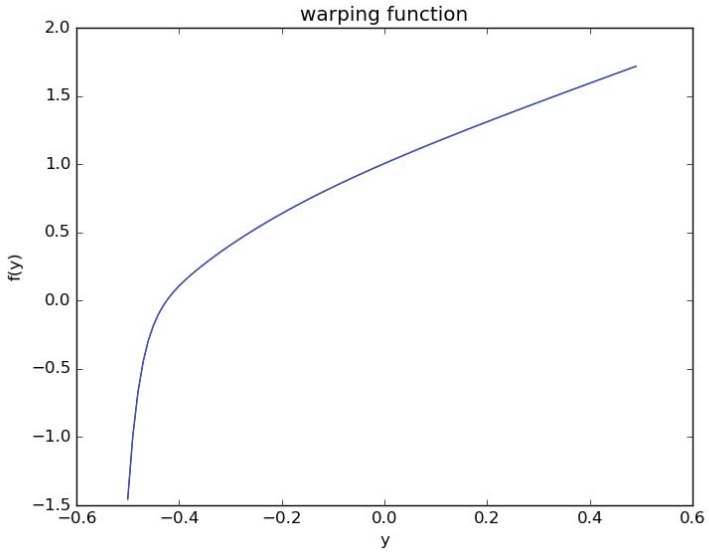


Figure A.1: Monotonic warping function from warped-lmm. x-axis shows the centred and normalised input phenotype; y-axis shows corresponding warped value.

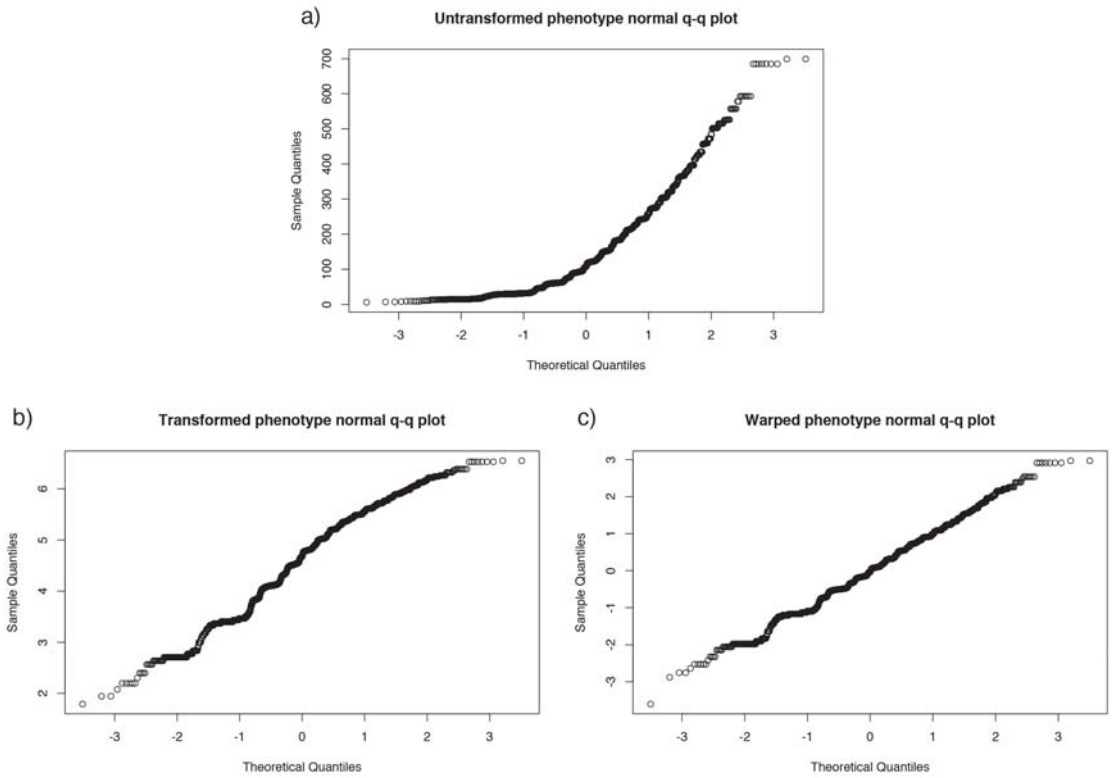


Figure A.2: Normal quantile-quantile plot of carriage length, and effect of monotonic transformation. Panel **a)** the inferred carriage duration, **b)** after the natural logarithm is taken, and **c)** after the warping function is applied.

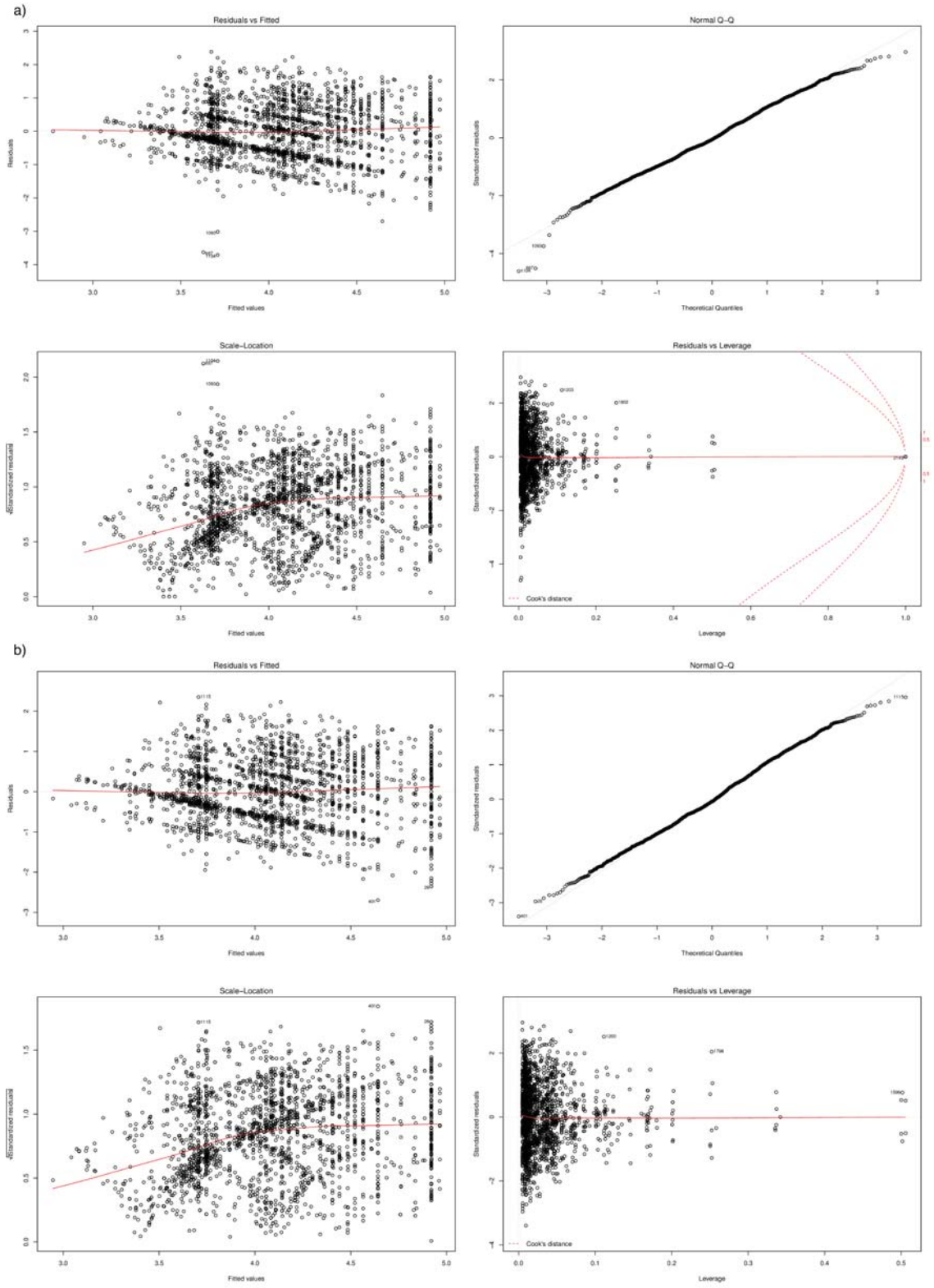


Figure A.3: Regression diagnostics and outlier removal. Panel **a)** shows prior to outlier removal, **b)** after outlier removal as produced by `plot.lm()` in R. Points deviating from normal residuals (top right plot), and at high leverage (bottom right plot) were removed. These observations appeared to be due to swabs not taken at the prescribed monthly intervals.

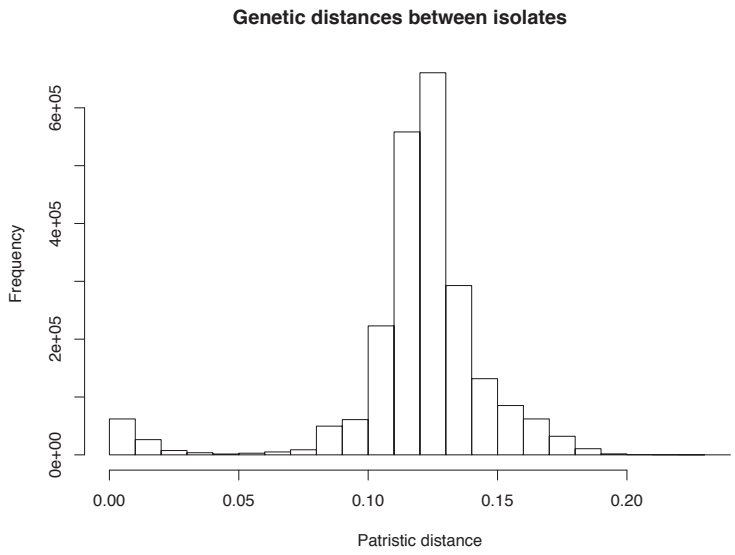


Figure A.4: Histogram of pairwise patristic distances on the inferred phylogeny. A cut-off for heritability estimation was chosen at 0.04, under which a clear second maxima corresponds to closely related isolates on the tree.

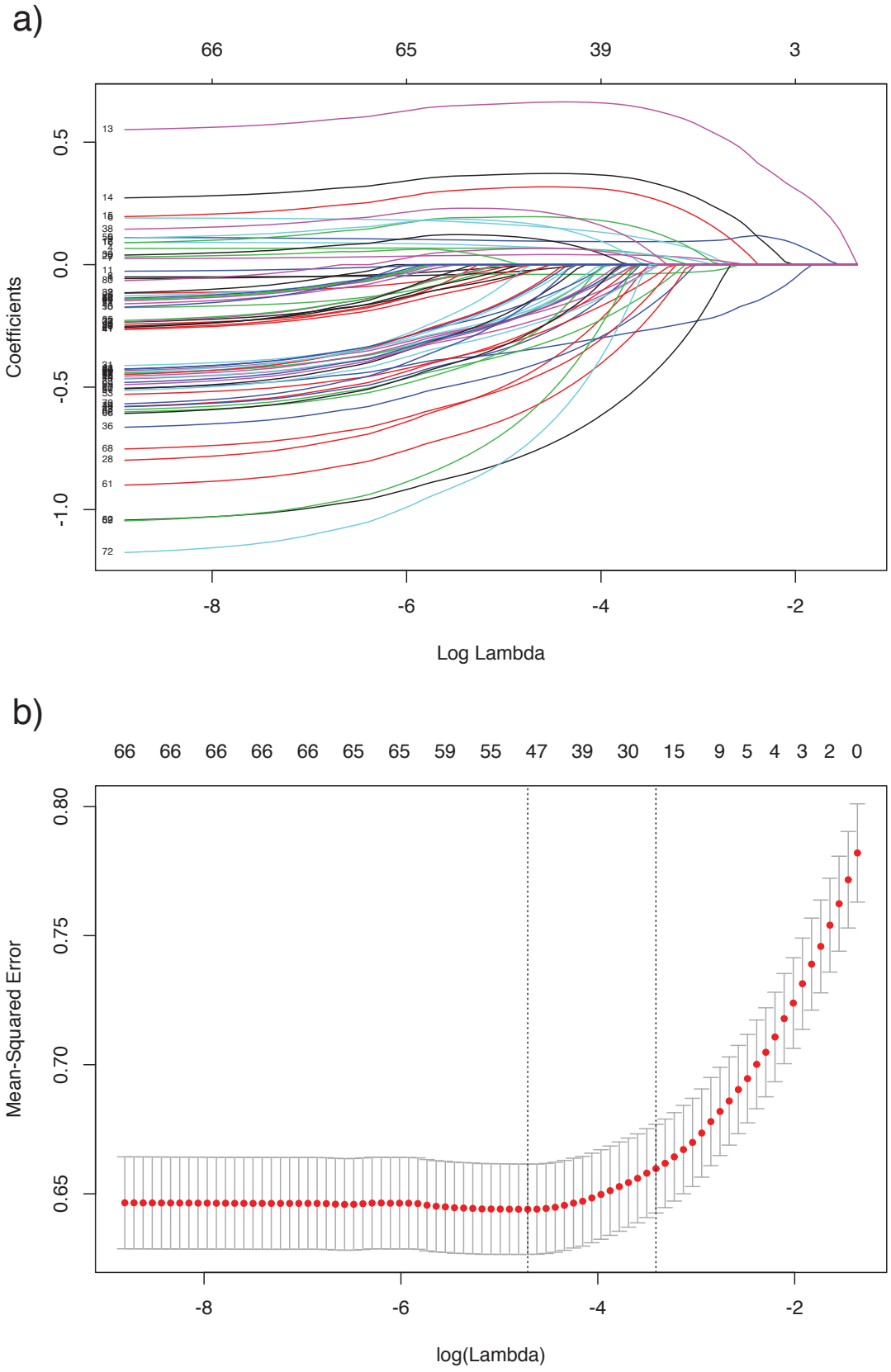


Figure A.5: Lasso regression plots for lineage effects. Panel **a)** shows the value of each predictor on the y-axis for different values of the ℓ_1 penalty λ on the x-axis, which increases from left to right. The labels along the top are the number of predictors remaining in the model for each λ. Panel **b)** shows the results of leave-one-out cross validation on the mean-squared error, along the same x-scale. The λ at minimum error is shown by the left dashed line, and the λ within one standard error is shown by the right dashed line.

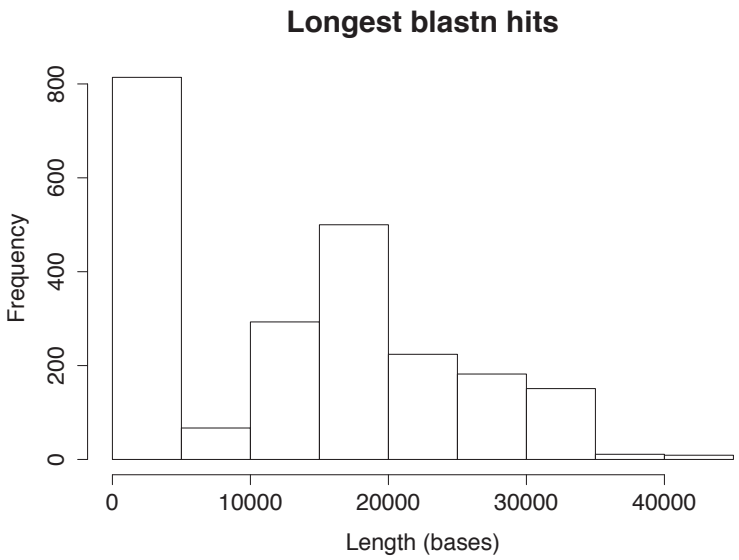


Figure A.6: Identification of phage in assemblies by `blastn` hit length. Histogram of the length of top hits against a database of phage sequence by `blastn`. Isolates with >5000bp hits were defined as having phage present.

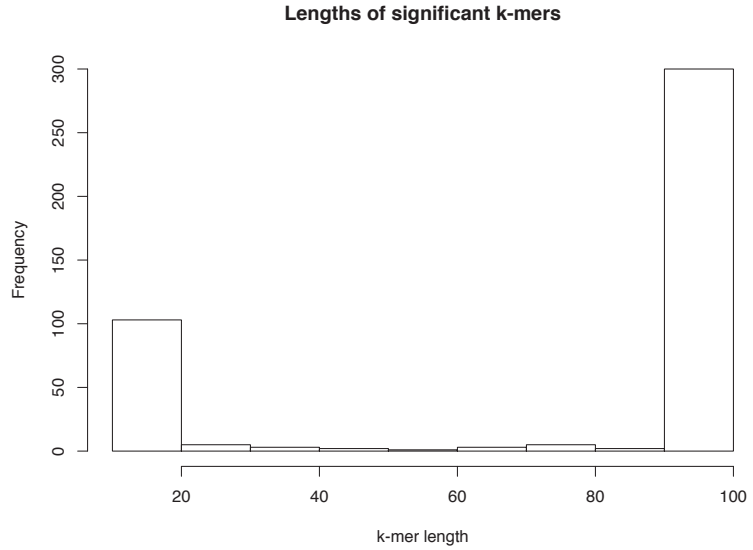


Figure A.7: The lengths of those k-mers reaching significance in the LMM analysis, binned by frequency. Lengths below 20 bases were filtered from downstream analysis, due to having low specificity.

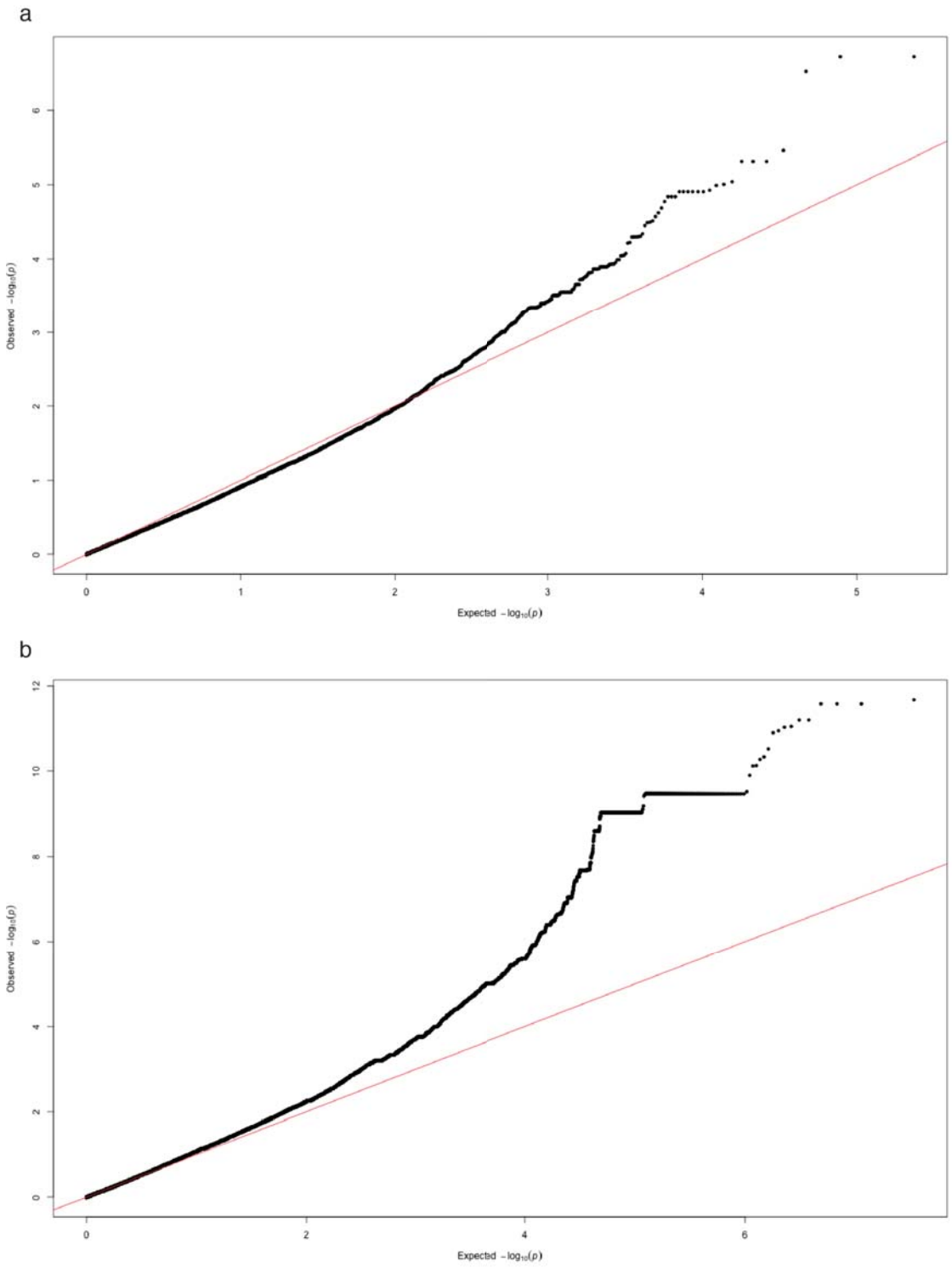


Figure A.8: Quantile-quantile plots of association p-values. For `fast-lmm` results on **a)** SNPs passing quality filters and **b)** k-mers of all lengths passing quality filters.

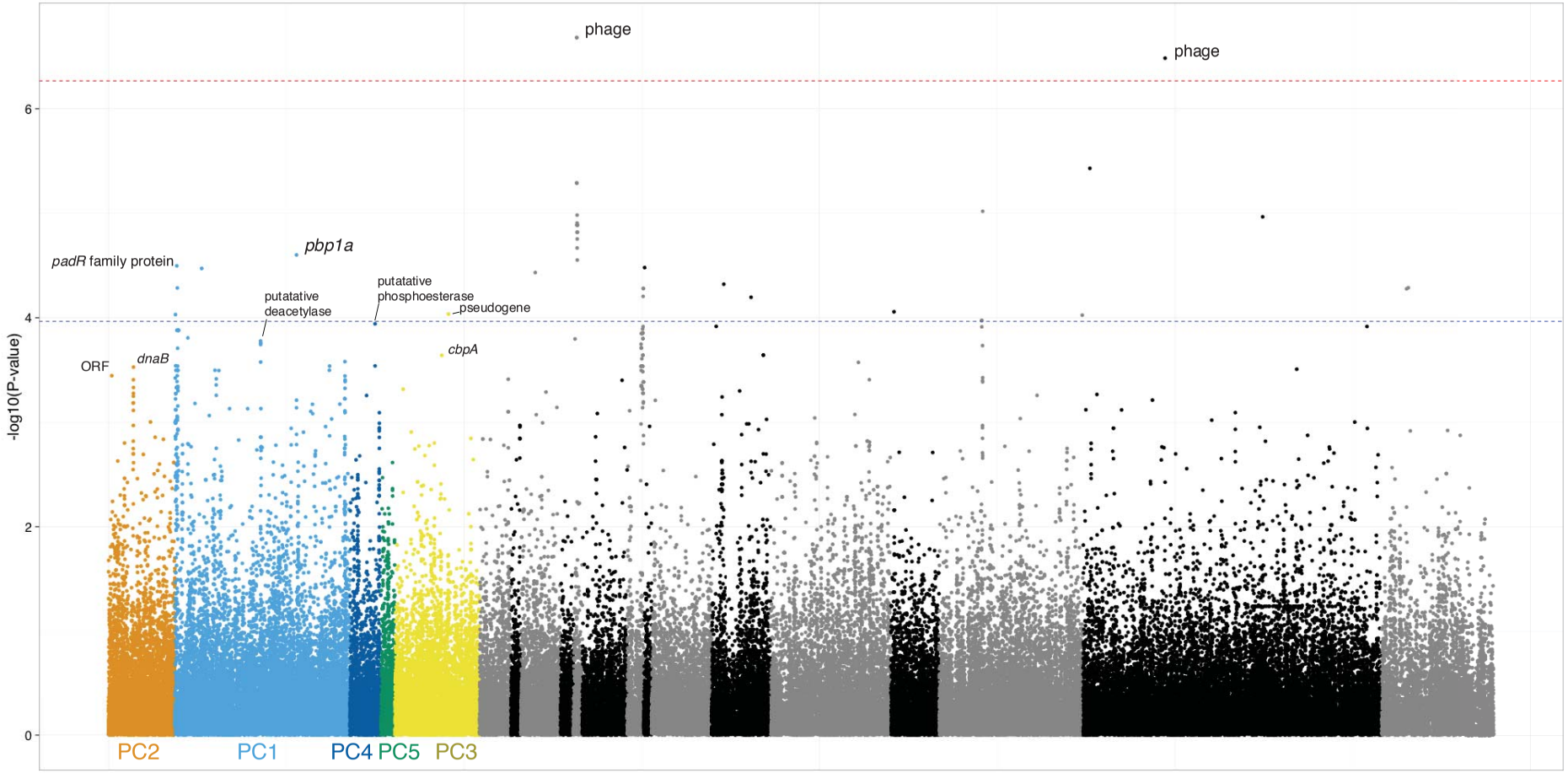


Figure A.9: Possible SNPs associated with lineage and carriage duration. The SNPs and p-values as shown in fig. 3.6, however the x-axis is now ordered by strength of association of lineage (defined by principal component) with carriage duration. The left most lineages are those most associated, those in black/grey were not significantly associated. SNPs are coloured by the lineage they are most associated with.

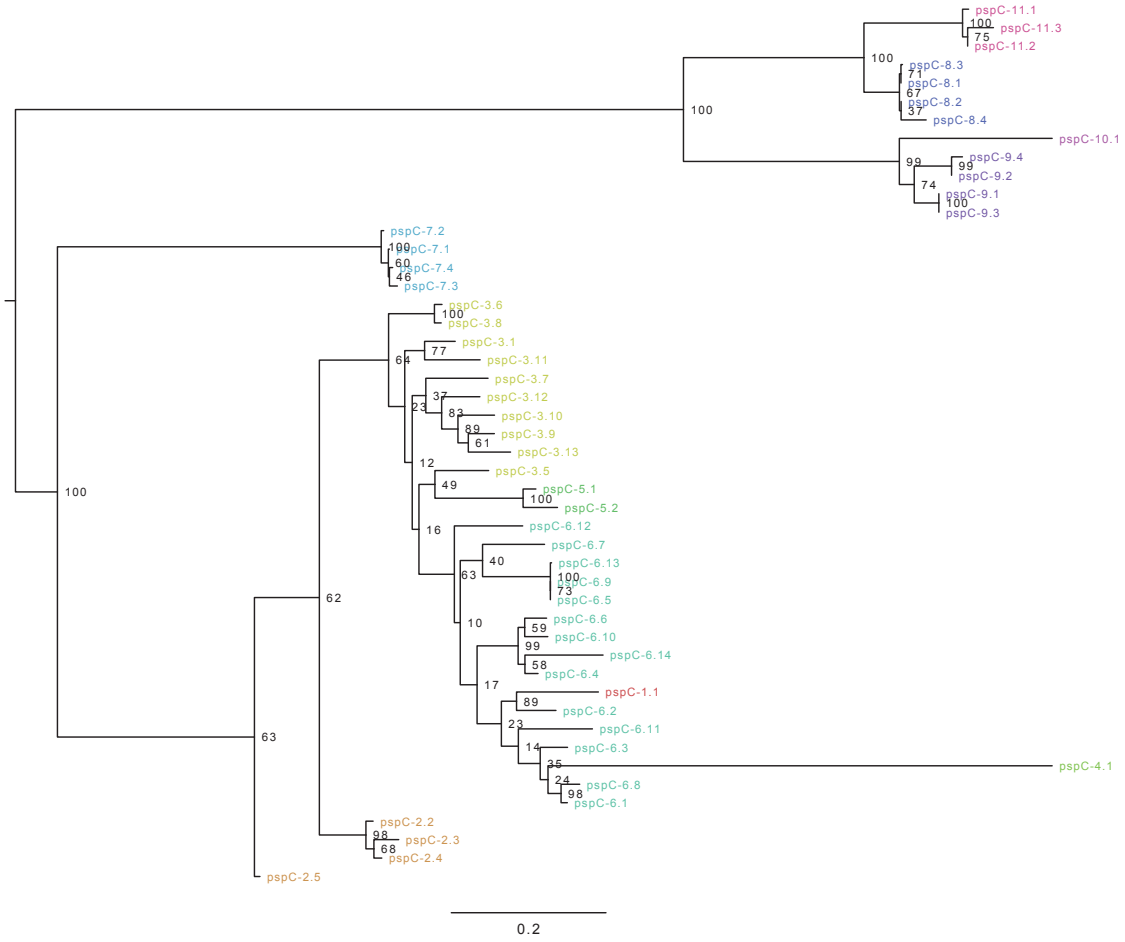


Figure A.10: Maximum likelihood tree of *pspC* protein alignment, with 100 bootstrap replicates (nodes are labelled with bootstrap supports). Tips are coloured by allele group.

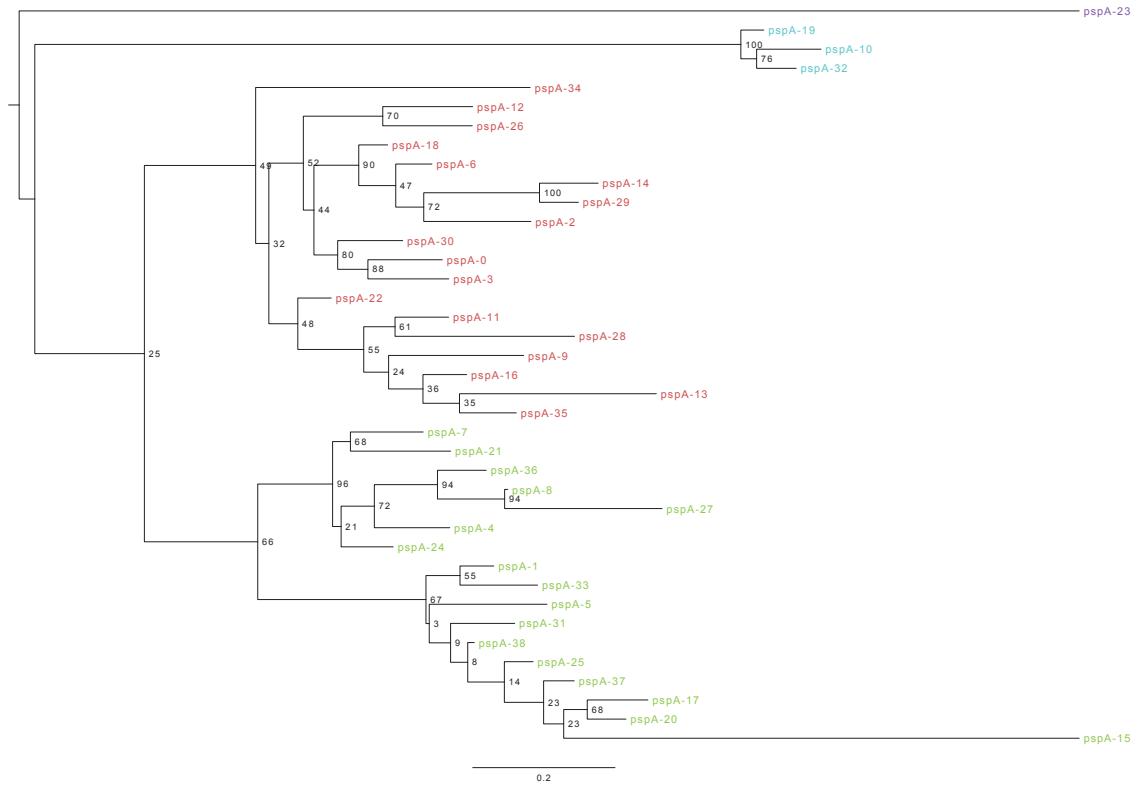


Figure A.11: Maximum likelihood tree of *pspA* protein alignment, with 100 bootstrap replicates (nodes are labelled with bootstrap supports). Tips are coloured by allele group.

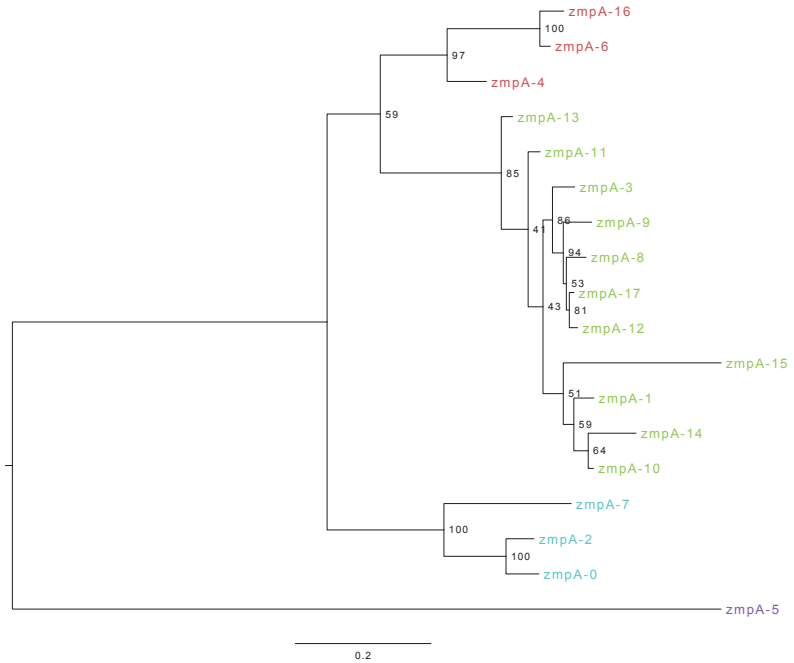


Figure A.12: Maximum likelihood tree of *zmpC* protein alignment, with 100 bootstrap replicates (nodes are labelled with bootstrap supports). Tips are coloured by allele group.

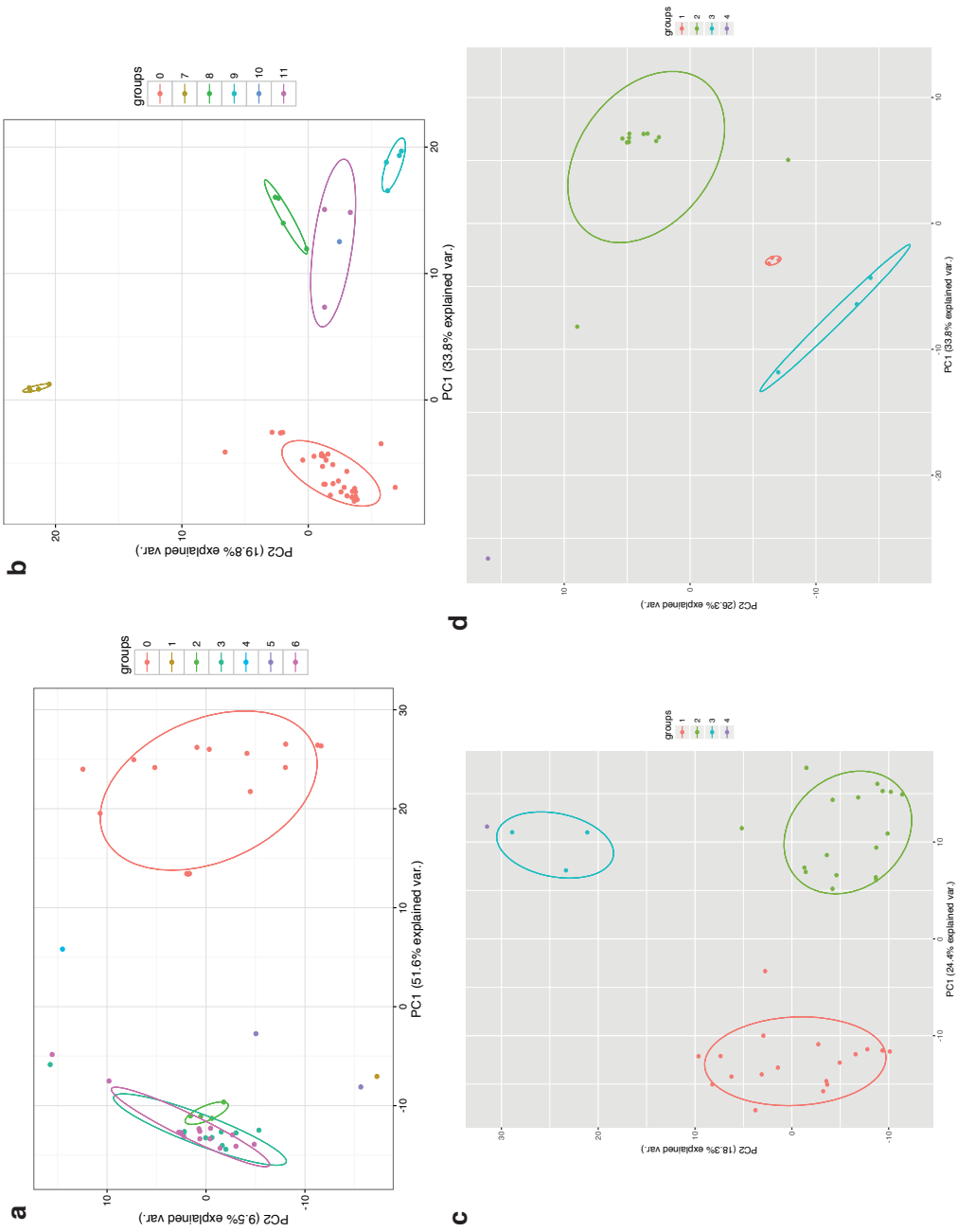


Figure A.13: PCA plots of classifiers used on antigen training data, first two principal components shown in each case. Points are coloured by the allele number, where 0 is a genome without the antigen. Where more than one point is available for a class, an ellipse has been drawn around its centroid. **a)** PspC, alleles 0–6; **b)** PspC, alleles 0,7–11; **c)** PspA, alleles 1–4; **d)** ZmpA, alleles 1–4

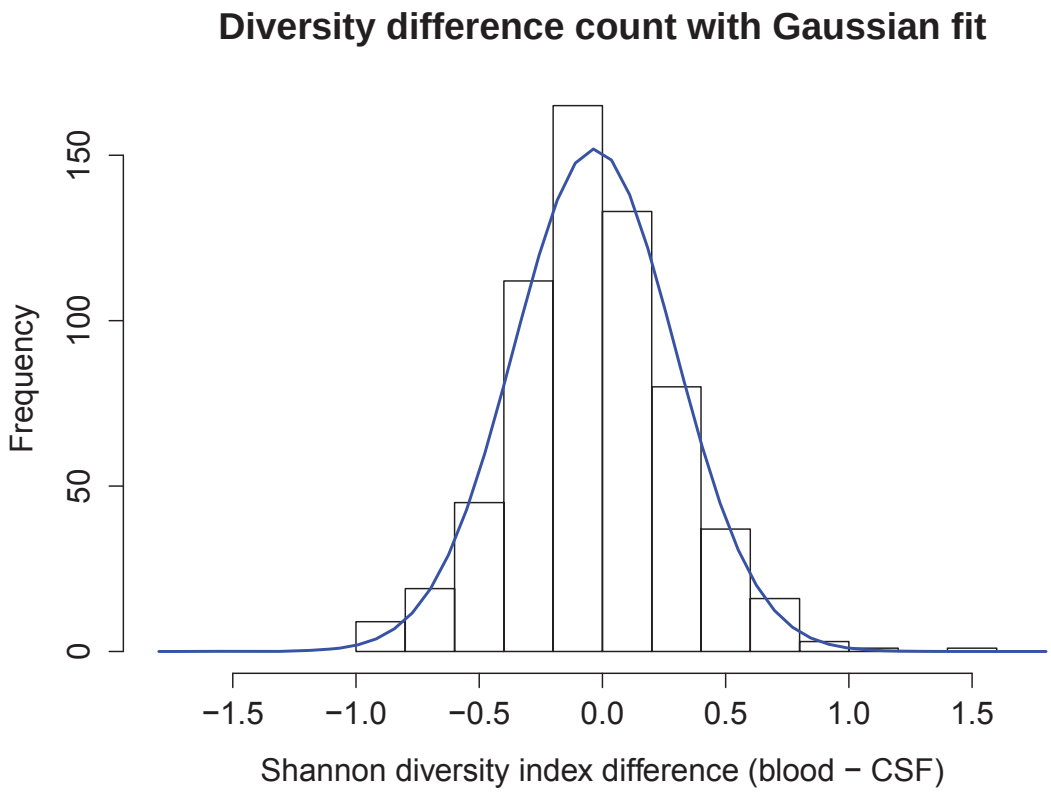


Figure A.14: Distribution of difference in Shannon diversity index between the *ivr* locus model π_{blood} and π_{CSF}. A Gaussian distribution was fitted to the data, which has a mean of roughly zero and little skew. The maximum possible Shannon diversity index (for equal amounts of each allele A-F) is 1.8.

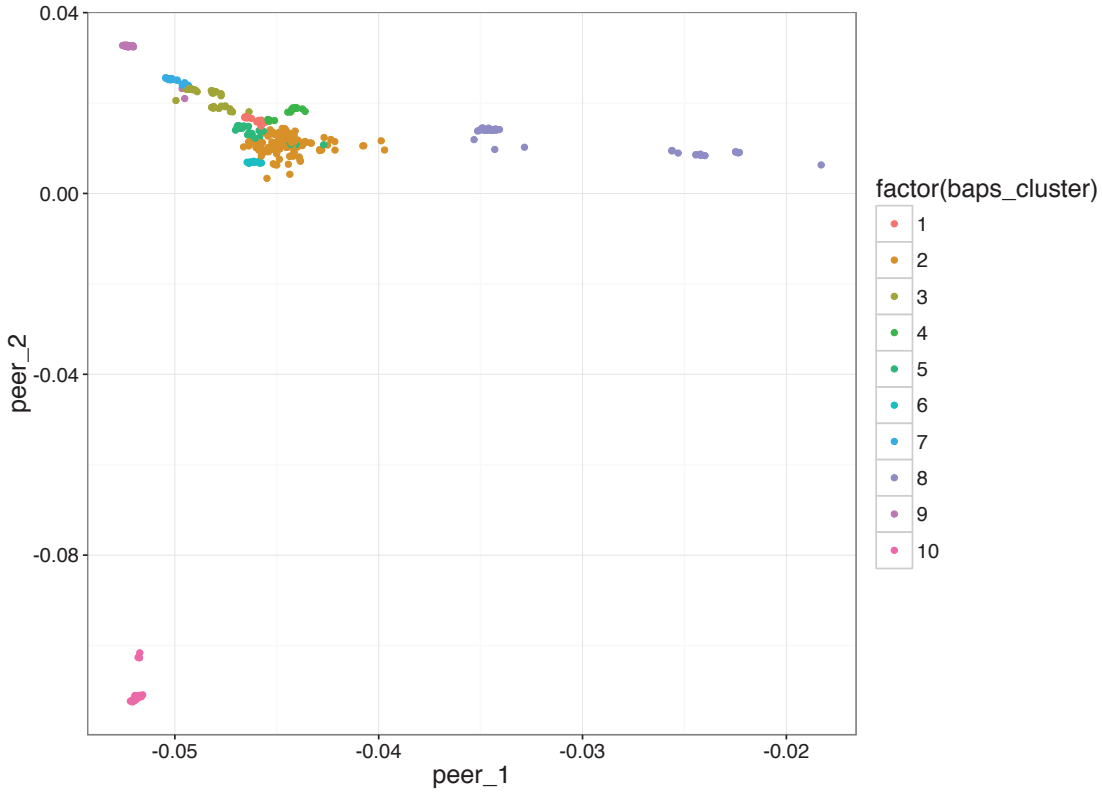


Figure A.15: Plot of the samples in the genome-to-genome analysis. x-axis is the first PEER factor loading, y-axis is the second PEER factor loading. Sample are coloured by the BAPS cluster they were assigned to.