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

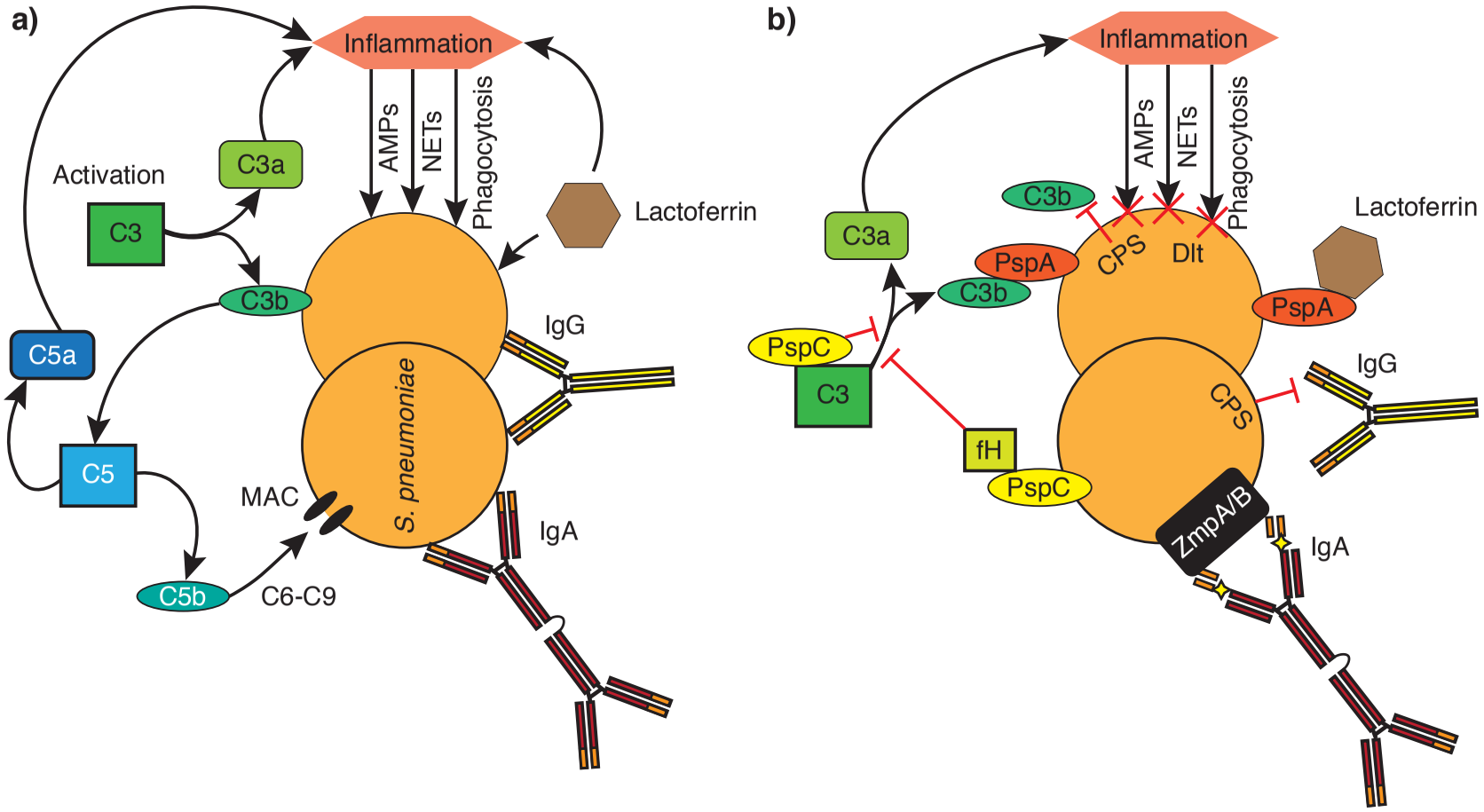

Figure 1.2: Interactions between the immune system and *S. pneumoniae*. **a)** Immune response to infection. Activation of the complement pathway by *S. pneumoniae* causes C3 to be cleaved into C3a and C3b. C3a increases inflammation; C3b binds to the bacterial surface, and leads C5 to be cleaved into 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

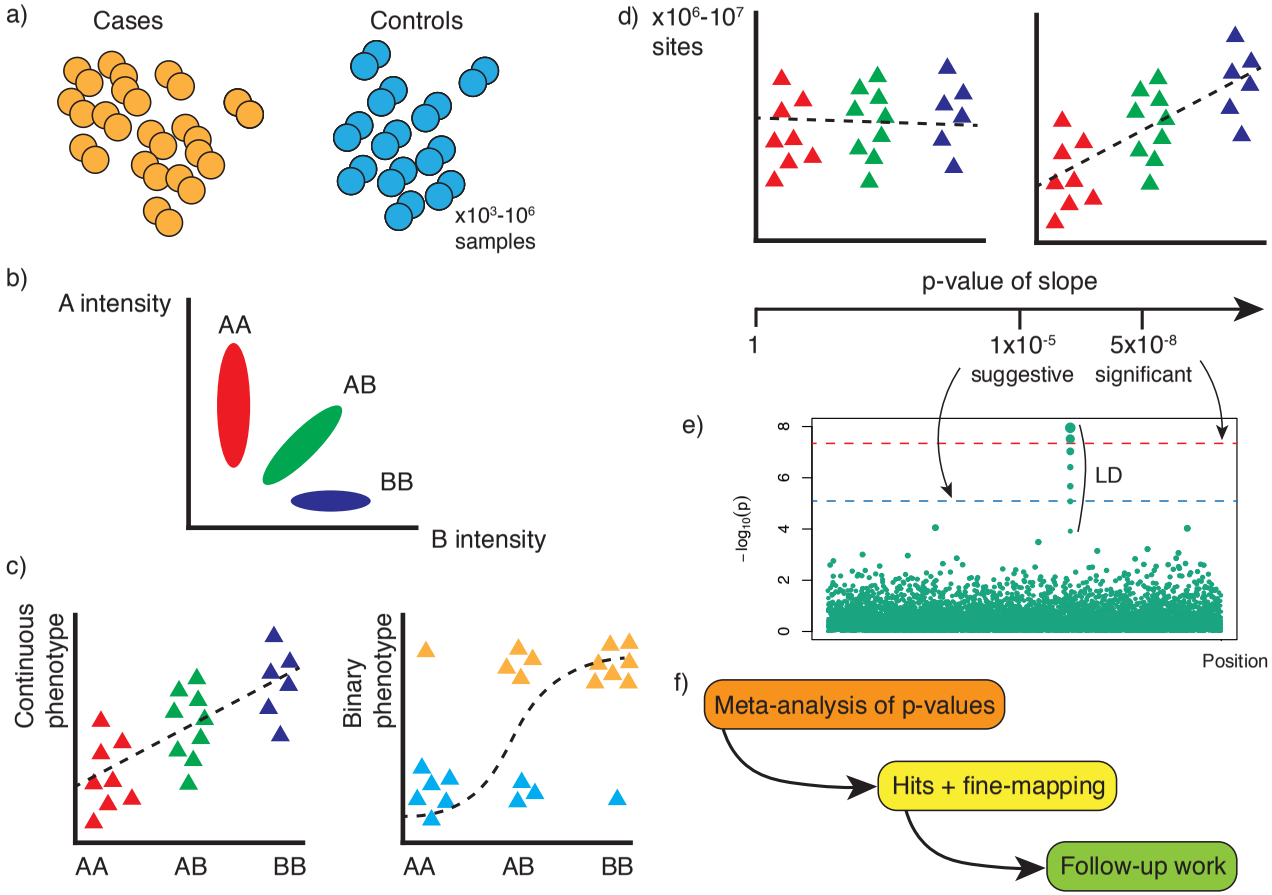

Figure 1.3: An overview of GWAS study design. **a)** Collect cases and control samples; discovery power increases with number of samples. Most successful GWAS studies in humans use at least 10^3 samples, with more recent studies reaching of the order of 10^6 samples. **b)** Pool samples together and 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).

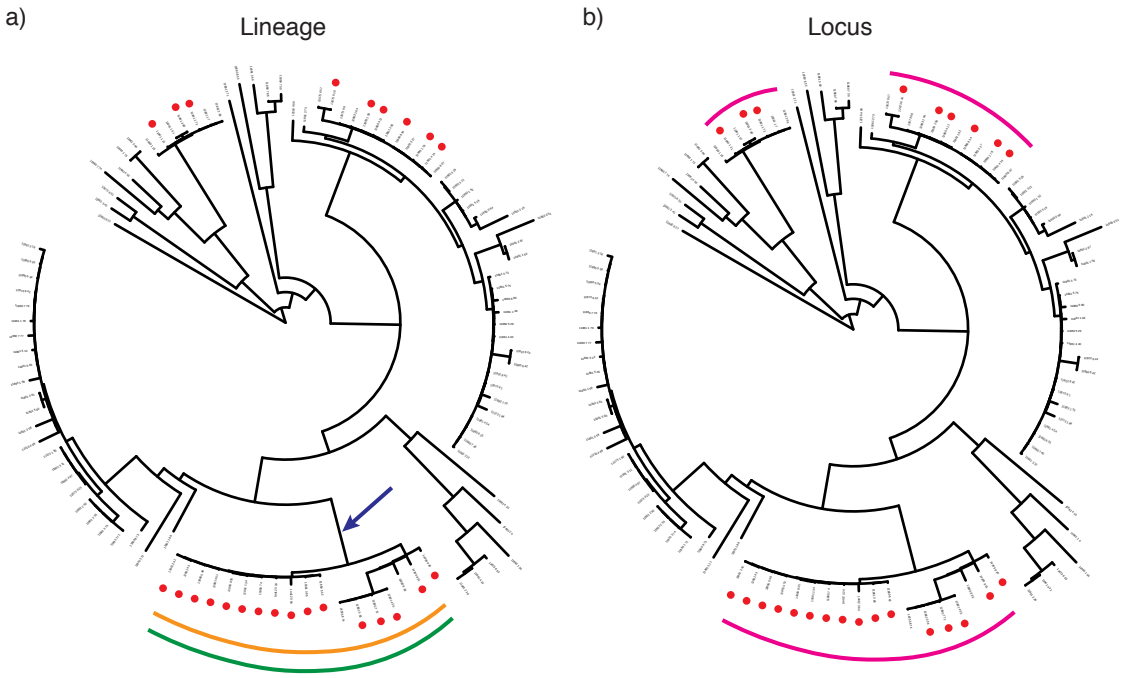

Figure 1.4: Phylogenetic illustration of lineage and locus variants. Depicted is an example phylogeny, with cases identified by red dots at the tips, and controls without dots. Variant presence is shown as coloured arcs. **a)**: The yellow variant is a causal lineage variant, and will be associated with 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.