Chapter 1

Epidemiology of *Neisseria meningitidis*

Neisseria meningitidis, also known as the meningococcus, is the bacterium responsible for meningococcal septicaemia and meningitis in humans. *N. meningitidis* has a global distribution, and the diseases it causes are fatal in around 11% of cases in the West (Goldacre *et al.* 2003). Meningococcal disease primarily affects children

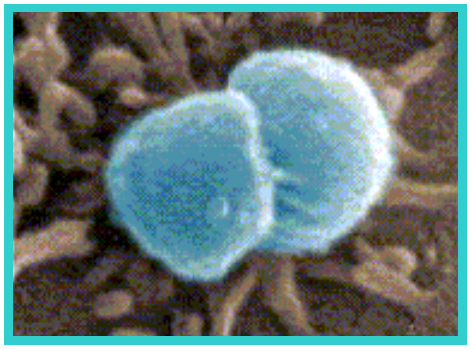


Figure 1 Neisseria meningitidis is a diplococcus ordinarily resident in the nasopharynx. Source: scanning electronic micrograph, Sanger Centre.

under 5 years of age, and is often characterized by a rapid deterioration from first symptoms to death. Cases of meningococcal disease tend to occur at a rate of about 1 case per 100,000 people throughout the world (Achtman 1995), but reach levels in excess of 500 per 100,000 people in severe epidemics that occur with some regularity in the Sahel, commonly known as the African meningitis belt, and China (Caugant 2001).

Meningococcal disease is principally controlled by mass vaccination of target groups. Population genetic studies can inform control and prevention strategies in two

important ways. Firstly, patterns of genetic diversity across multiple loci provide an, albeit corrupted, account of the epidemiological history and structure of the pathogen population. Secondly, patterns of genetic diversity can reveal the selection pressures exerted on meningococci at specific loci; of particular interest is natural selection imposed by interaction with the immune system. The role of population genetics is to model the epidemiological processes that give rise to the observed genetic diversity from an evolutionary perspective, in order to better understand those processes. In addition to informing control and prevention strategies, pathogens such as *N. meningitidis* provide interesting case studies in the study of evolution by virtue of their high levels of genetic diversity, short generation times and ongoing co-evolutionary arms race with the host immune system.

In this chapter I will begin by reviewing the field, and justifying the use of population genetics models, and the coalescent in particular, in modelling microparasites such as *N. meningitidis*. In Chapter 2 I use a modification to approximate Bayesian computation to assess the fit of the standard neutral model to populations of carried meningococci. The source of genetic structuring is investigated in Chapter 3, first by fitting the neutral microepidemic model using approximate Bayesian computation, and then with AMOVA and Mantel tests to quantify the differentiation between carriage and disease populations, and the extent to which geography and host age structure carriage populations. Together these results suggest ways in which the standard neutral model might be revised to provide a better fit to observed patterns of genetic diversity.

The role of natural selection in shaping meningococcal diversity is investigated in Chapter 4 using a novel method that utilises an approximation to the coalescent and reversible-jump Markov chain Monte Carlo to detect sites under selection in the presence of recombination. Having performed a simulation study to assess the statistical properties of the method, in Chapter 5 I apply it to the *porB* antigen locus and seven housekeeping loci in *N. meningitidis*. The differences in selection pressures experienced by these different types of loci reflect the function and exposure to the host immune system of their gene products. Finally in Chapter 6 I discuss the results and limitations of the methods covered in this thesis, and consider the future direction of population genetic approaches to understanding infectious disease.

This chapter is organised into four sections. I begin in section 1.1 with an overview of the biology of *N. meningitidis*, including the pathology, epidemiology of carriage and disease populations, methods used for meningococcal typing and public health strategies used in control and prevention. Next in section 1.2 I review how the understanding of meningococcal population biology has changed over time as typing technologies have developed and as larger-scale studies have been undertaken. Some mathematical models that have been used to describe meningococcal populations are discussed. In section 1.3 I discuss the application of population genetics techniques to infectious disease in general, and how the population genetics approach has helped understand pathogen evolution. Finally in section 1.4 I argue that the coalescent is the natural starting point for population genetic analysis of *N. meningitidis*, by showing how, in a simple population model of meningococcal infection in a host population, the dynamics of prevalence are described by a familiar SIRS epidemiological model

and the genealogy of a sample of the pathogen population is described by the coalescent.

1.1 Overview of *Neisseria meningitidis*

1.1.1 Epidemiology

Despite its notorious pathogenicity, *N. meningitidis* is a natural human commensal, normally residing in the nasopharynx (Figure 1). Whereas incidence of disease is of the order of one case per 100,000 people endemically, carriage of disease is very much more common, typically one carrier per 10 people. The meningococcus has several adaptations to life in the nasopharynx, including pili for cytoadhesion to the nasopharyngeal epithelium and human transferrin and lactoferrin binding receptors for sequestering iron (Cartwright 1995). Disease occurs only when the meningococcus crosses the nasopharyngeal epithelium and enters the blood stream.

1.1.1.1 Pathology

When meningococci ordinarily commensal to the nasopharyngeal epithelium invade the blood stream they can cause septicaemia (blood poisoning) and, if the bacteria cross the blood-brain barrier, meningitis, an inflammation of the brain lining (meninges). When treated, meningococcal disease has a fatality rate of 11% (Goldacre *et al.* 2003). Meningitis alone has a fatality rate of 5%; most deaths from meningococcal disease are caused by septicaemia. Patients presenting with septicaemia but not meningitis have a 20% mortality rate, but this is closer to 50% if the patient has already gone into shock. Of those infected with meningococci, 15%

Figure 2 Symptoms of meningococcal meningitis. Source: Meningitis Research Foundation (2005).

suffer meningitis alone, 30% septicaemia alone and 50% a combination (Meningitis Research Foundation 2005). The remainder suffer milder symptoms.

Meningitis can progress rapidly from first symptoms to death. The onset of meningitis is associated with sore throat, headache, drowsiness, fever, irritability and neck stiffness (Figure 2). Bacterial toxins in the brain cause inflammation and can result in coma. Septicaemia is manifest externally as a haemorrhagic skin rash (Figure 3) that does not fade when pressed, by a glass tumbler for example. For 35% of patients this septicaemia is fulminating, including disseminated coagulation in blood vessels, flooding of the circulatory system with bacterial endotoxins, shock and kidney failure. In the most severe cases bleeding can occur in the brain and adrenal glands (Mims 1998).

N. meningitis is a gram negative bacterium, and treatment proceeds by immediate administration of the antibiotic penicillin, ampicillin or chloramphenicol. In the

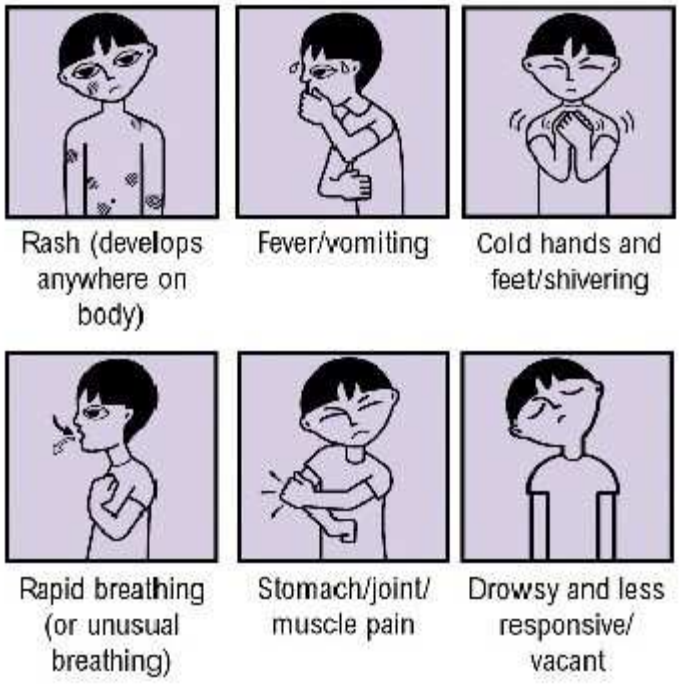


Figure 3 Symptoms of meningococcal septicaemia. Source: Meningitis Research Foundation (2005).

absence of treatment, the fatality rate for meningococcal disease approaches 100%. Following the acute phase of the infection the patient is treated with rifampin to clear nasopharyngeal carriage, and close contacts such as family are treated prophylactically with rifampin (Mims 1998). After-effects are rare, but include hearing damage, nerve palsies and epilepsy.

1.1.1.2 Epidemiology of meningococcal disease

Prevalence of meningococcal disease varies globally, seasonally, and with age of host. To some extent meningococcal disease epidemiology obeys national boundaries meaning that adjacent countries can experience quite different meningococcal epidemiology, yet historically meningococcal disease has been characterized by a

number of successive sweeps of global pandemics affecting several countries at any one time.

The meningitis belt of sub-Saharan Africa suffers semi-regular outbreaks of meningococcal disease with a period of some 8-12 years and attack rates of the order of 500 cases per 100,000 people (Lapeyssonie 1963; Schwartz *et al.* 1987). Outbreaks in developed countries have been rare since large-scale mobilisation of troops during the Second World War caused meningococcal pandemics in Europe and North America. During the 1970s an outbreak emerged in Norway with attack rates of the order of 10 cases per 100,000 people, which subsequently spread across Europe including the United Kingdom and reached countries as far away as Cuba, Chile and Brazil. In 1987 a virulent meningococcal outbreak during the annual Haj pilgrimage to Mecca was spread globally by pilgrims returning to their home countries (Schwartz *et al.* 1987). Meningococcal disease in developed countries is generally characterized by small sporadic outbreaks, with a background attack rate of 1 case per 100,000 people (Achtman 1995).

Disease outbreaks are sensitive to seasonal effects, but the exact relationship varies globally. In the African meningitis belt epidemics coincide abruptly with the harmattan (dry season). During this time climatological features such as humidity, airborne dust, rainfall and wind patterns undergo marked changes, and these in turn lead to changes in human behaviour. The harmattan ends with the arrival of the rains. By contrast in Europe and North America disease rates peak during winter months and steadily decline to low levels by autumn (Cartwright 1995). Many other bacterial and viral infections show a similar seasonality in incidence.

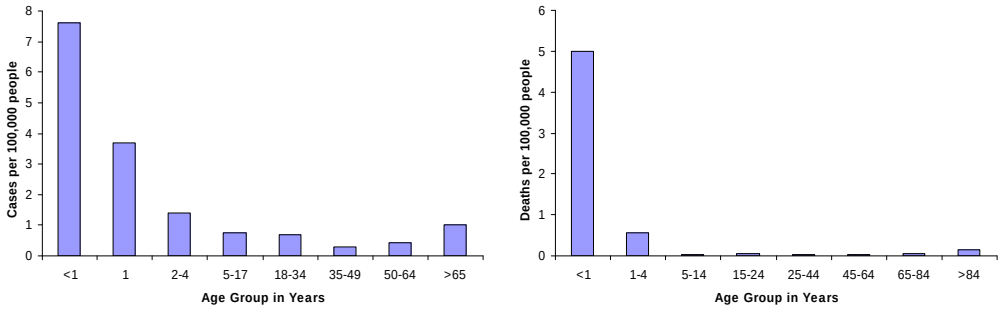


Figure 4 Left: rate of incidence of meningococcal disease with age in the United States, 2000. Right: fatality rate of meningococcal disease with age in the United Kingdom, 2002. Sources: Centers for Disease Control and Prevention (2000), National Office of Statistics (2002).

Age is an important factor in rates of incidence and recovery from meningococcal disease. Meningococcal disease primarily affects children under 5 years of age: incidence peaks in infants aged about 6 months and subsequently declines steadily (Cartwright 1995). Figure 4 (left) shows that in the United States, the rate of meningococcal disease in children halves by the age of 1 and halves again by the age of 4 (Centers for Disease Control and Prevention 2000). By comparison, Figure 4 (right) shows that the fatality rate is considerably worse in young children (National Office of Statistics 2002), assuming that the incidence rates are similar in the U.K. and U.S.

1.1.1.3 Epidemiology of carriage

Not only is meningococcal carriage vastly more prevalent than disease, but patterns of carriage differ markedly to patterns of disease. The carriage rate in the United States and Europe is about 10% (Broome *et al.* 1986; Caugant *et al.* 1994), some 10,000 times the rate of disease. However, institutions which house closed or partially-closed

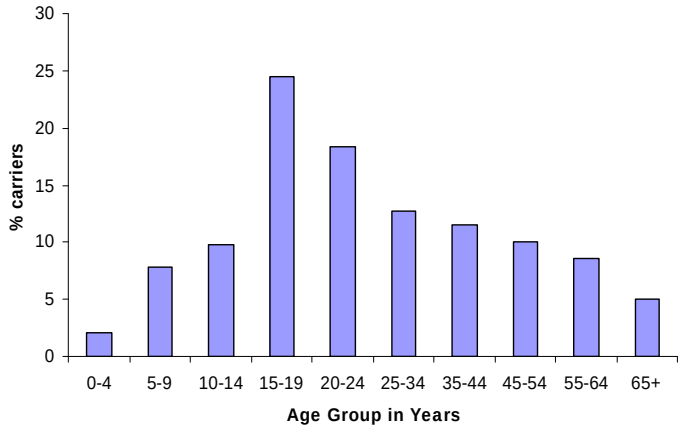


Figure 5 Percentage carriage with age in Gloucestershire, United Kingdom, 1986. Source: Cartwright *et al.* (1987).

communities traditionally exhibit elevated carriage rates. Military training camps, boarding schools and prisons commonly have carriage rates in excess of 50% (Cartwright 1995). Patterns of carriage differ from patterns of disease in their geographic distribution, sensitivity to seasonal effects and age profile of hosts.

Despite the dramatic seasonality of disease incidence in Asia and the African meningitis belt, meningococcal carriages rates are relatively insensitive to seasonal fluctuations. A distinct lack of seasonal variability has been reported in studies in Nigeria and India (Blakebrough *et al.* 1982; Ichhpujani *et al.* 1990). Similarly, carriage rates in temperate regions do not appear to mirror the seasonality of incidence rates according to studies in Belgium and the United States (De Wals *et al.* 1983; Aycock and Mueller 1950). Nevertheless, carriage rates are known to react to passing epidemics, with up to 70% carriage during severe disease outbreaks (Cartwright 1995).

The age distribution of meningococcal carriage differs substantially to the age distribution of infected hosts. Figure 5 shows the results of a study in Gloucestershire, United Kingdom (Cartwright *et al.* 1987). Carriage rates are lowest in infants and young children, and peak at about 25% in late teenage years and early twenties. This contrasts firstly with the observation that mortality is gravest in children under 5, and secondly with *N. lactamica* carriage rates, which peak in infancy and then steadily decline (Bennett *et al.* 2005).

The contrast in epidemiology between meningococcal carriage and disease raises several questions, in particular: Can carriage isolates cause disease or are disease-causing isolates genetically distinct? Are disease-causing isolates a subset of carriage isolates or do they circulate independently? Can disease-causing isolates persist long-term or do they emerge recurrently from carriage isolates? Do the incongruent age profiles of carriers and cases reflect different susceptibilities or different circulating forms? In order to address these problems it is necessary to genetically characterize the meningococci, and that is the role of typing.

1.1.2 Typing

In general, typing is useful if there is any association between genotype and a phenotype of interest such as propensity to cause disease or susceptibility to particular drugs. If closely-related groups of meningococci share epidemiological or pathological features in common, then typing provides information about the bacteria that may help in tracking and controlling the spread of disease-causing and non-disease-causing isolates and prescribing appropriate treatment to infected patients. Typing systems determine the genotype using a phenotypic marker or directly using

Table 1 Meningococcal outer membrane proteins

OMP Class	Protein	Function	Typing level
1	PorA	Porin	Serosubtyping
2 and 3	PorB	Porin	Serotyping
4	Rmp	Reduction modifiable protein	Not used
5	Opa/Opc	Opacity protein	Not used

sequencing. Three kinds of typing have been used widely in the study of *N. meningitidis*: immunological typing, electrophoretic typing and sequence typing. Immunological typing and electrophoretic typing use phenotypic markers. In these typing schemes it is not necessary to know the underlying genotype, but there must be a strong correspondence between variants of the marker and variants of the underlying genetic locus to make typing useful. However, as DNA sequencing technology has developed and become less costly, direct sequencing has become more important for typing. DNA sequencing has also allowed the genotypes underlying phenotypic typing schemes to be determined.

1.1.2.1 Immunological typing

Traditionally meningococci have been differentiated according to their immunogenic properties, which are determined primarily by the capsular polysaccharide and proteins that span the phospholipid outer membrane. Between the outer membrane and the cytoplasmic membrane lies a peptoglycan layer. Shedding of outer membrane vesicles known as blebbing plays an important role in immune evasion. Blebs contain outer membrane proteins (OMPs) and lipopolysaccharide that are highly immunogenic. Blebs bind antibodies that might otherwise bind to the whole

bacterium. Five principal classes of OMP have been identified (Table 1) and together with the capsular polysaccharide, these form the basis of immunological typing (Poolman *et al.* 1995).

Serogroup is the primary immunological type and is determined by the polysaccharide capsule. There are thirteen recognised serogroups (A, B, C, 29-E, H, I, K, L, W-135, X, Y, Z). Serogroups A, B and C are responsible for 90% of invasive disease; the remainder is accounted for chiefly by serogroups Y and W135 (Poolman *et al.* 1995). The capsule is a pre-requisite for invasive disease; many meningococci do not express a capsule and cannot be typed serologically. The capsule-synthesis (*cps*) cluster is the genetic determinant of meningococcal serogroup, and comprises five regions A-E. The capsules of serogroups B, C, Y and W135 all contain sialic acid, and are variants of the *siaD* locus. Serogroup A capsules do not contain sialic acid but do contain mannosaminephosphate, encoded at the *myn* locus. Both *siaD* and *myn* are situated in region A of *cps*. Meningococci that are serologically ungroupable occur either because mutation leads to the capsule not being expressed, or because the capsule-encoding loci are lacking (Vogel *et al.* 2001; Claus *et al.* 2002). In the former case, these meningococci can still be characterized at the *cps* cluster using sequencing (Claus *et al.* 2002).

Serotype is determined by variants of the PorB OMP, a porin encoded at the *porB* locus, and serosubtype by variants of the PorA OMP, another highly-expressed porin encoded at the *porA* locus. PorB and PorA are subcapsular proteins that allow the passage of ions across the phospholipid membrane. PorA and PorB show cation and anion selectivity respectively (Poolman *et al.* 1995). The PorA protein has two

hypervariable regions, VR1 and VR2, and combinations of variants at each region are possible. The full typing classification is denoted, for example, B:4:P1.16,7, meaning serogroup B, serotype 4, serosubtype 1 with VR1 16 and VR2 7.

1.1.2.2 Electrophoretic typing

Gel electrophoresis has a higher resolution than immunological typing because amino acid variants that are immunologically equivalent can be distinguished. Using gel electrophoresis, non-synonymous nucleotide variation at a locus can be detected and the frequencies of the different variants estimated, regardless of the immunogenic properties of those variants. Gel electrophoresis is generally applied to water-soluble cellular enzymes and works because amino acid variants have different electrophoretic properties. Amino acid polymorphism causes variation in the net electrostatic charge of the enzymes because different amino acids have different charges. This variation is detected by differential rates of migration across the gel when a current is applied.

Variants identified by gel electrophoresis are known as allozymes (i.e. allelomorphs, or variants, of the same enzyme), or electromorphs. Allozyme refers to variants of a particular orthologous locus, whereas the term isozyme can refer to paralogous variants. As a result of the inability of gel electrophoresis to detect synonymous nucleotide polymorphism, a particular allozyme may represent multiple nucleotide alleles. However, not all non-synonymous variants are distinguishable because some have equivalent electrophoretic mobility. Studies suggest that gel electrophoresis detects around 80-90% of non-synonymous variation (Selander *et al.* 1986).

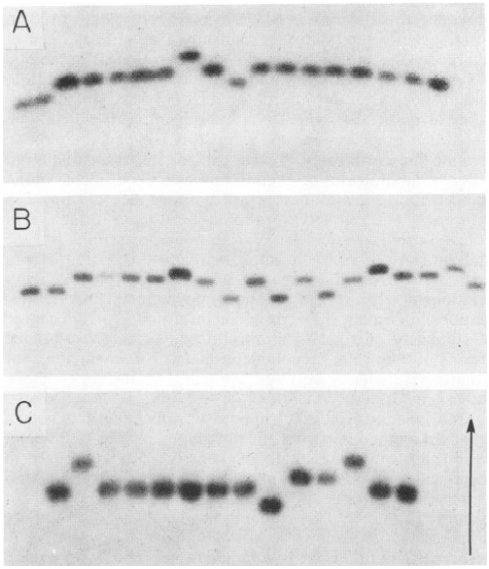


Figure 6 Gel illustrating electrophoresis of three enzymes of *Escherichia coli*. The arrow indicates the direction of migration of the enzymes across the gel. Source: Selander *et al.* (1986).

Figure 6 illustrates the results of gel electrophoresis on three enzymes of *Escherichia coli*. The columns correspond to bacterial isolates and the vertical height of the band reveals variation in electrophoretic mobility, corresponding to amino acid polymorphism. Results from several loci can be combined to give information about the frequencies of multilocus allelic combinations. This is the basis of the technique known as multilocus enzyme electrophoresis (MLEE), which has been widely applied to bacterial populations (Selander *et al.* 1986). A moderate number of loci, usually between 10 and 30, are usually analysed with MLEE. Combining loci in this way is useful because (i) it increases the information content that is limited by diversity at any one locus and (ii) highlights any differences between epidemiological processes influencing different loci. Each electromorph (allozyme) at a locus is given an arbitrary label, usually a number reflecting the order in which the electromorph was first discovered. Each combination of electromorph numbers across loci, the multilocus profile, is also designated by a number, and this is referred to as the electrophoretic type (ET). The number of observed ETs will typically be much fewer

than the sample size of the isolate collection, even when the sample size is much smaller than the number of possible ETs.

One difficulty with MLEE is that the nomenclature for labelling electromorphs and, hence, ETs is not readily portable between laboratories in the sense that gel electrophoresis gives only relative electrophoretic mobility. The relative electrophoretic mobility cannot be reliably converted into an absolute measurement. Therefore comparing results between laboratories requires standards to be shared between laboratories and included in every electrophoresis.

1.1.2.3 Sequence typing

For genetic characterisation the highest level of resolution is the nucleotide sequence itself. Sequence typing is able to distinguish between alleles that differ only by synonymous nucleotide substitutions, which would be invisible to immunological and electrophoretic typing, and allows non-coding loci to be typed. Multilocus sequence typing (MLST; Maiden *et al.* 1998) has several advantages over MLEE for epidemiological surveillance (Urwin and Maiden 2003). Whereas it has made MLEE redundant, MLST coexists with immunological typing, partly as a result of the loci chosen as the standard for MLST.

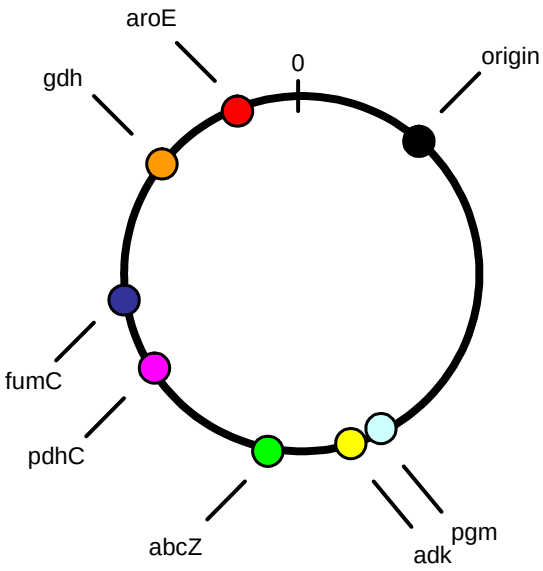


Figure 7 Locations of the seven housekeeping loci in the *N. meningitidis* Z2491 genome (Parkhill *et al.* 2000) that are used for multilocus sequence typing. The origin of replication is marked (origin) and the reference point for nucleotide positions (0).

MLST was pioneered in *N. meningitidis*, but its use is now widespread in many other bacterial species. In *N. meningitidis* the MLST protocol consists of obtaining short nucleotide sequence fragments, about 450 base pairs (bp) in length, in seven loci distributed about the 2.2 megabase (Mb) genome (Parkhill *et al.* 2000), as shown in Figure 7. Seven housekeeping genes were chosen out of twelve proposed genes to meet certain criteria, based on assessing those criteria in a collection of 107 isolates assembled to represent the global diversity observed in carriage and disease samples to date (Maiden *et al.* 1998).

Those genes were required to have intermediate levels of genetic diversity to facilitate typing; the level of diversity had to meet the desired balance of sensitivity and specificity. Genes were excluded that were thought to be unusually influenced by natural selection or recombination. That reinforced the requirement for intermediate levels of diversity, and suggested the use of housekeeping loci (Urwin and Maiden 2003). The function of the seven MLST genes is summarised in Table 2. Congruence between analyses of genetic clustering based on MLST and MLEE was also taken into account (Maiden *et al.* 1998; see later for more details). The fragment length of ~450bp results from the length of sequence that could practicably be determined in the sequence trace using a single gel electrophoresis in 1996 (Urwin and Maiden 2003).

Having obtained the nucleotide sequences, each allele can be assigned an arbitrary label which is a number that roughly reflects the order in which the allele was discovered. This allele number is analogous to the number assigned to electromorphs

Table 2 Function of the seven loci used in MLST in *N. meningitidis*

Locus	Function
abcZ	Putative ABC transporter
adk	Adenylate kinase
aroE	Shikimate dehydrogenase
fumC	Fumarate hydratase
gdh	Glucose-6-phosphate dehydrogenase
pdhC	Pyruvate dehydrogenase subunit C
pgm	Phosphoglucomutase

during MLEE. Each combination of allele numbers observed is known as the allelic profile, and is assigned an arbitrary label known as the sequence type (ST). This label, which is actually a number, is analogous to the ET obtained by MLEE.

The results of MLST are more easily shared, in contrast to the results of MLEE typing. The nucleotide sequences can be stored digitally, usually as text files on a computer, and sent instantly to other laboratories using e-mail. As a result it is straightforward to verify that the nomenclature used to assign allele numbers and STs is consistent from laboratory to laboratory. There is a central repository for *N. meningitidis* MLST data (http://neisseria.org/mlst/) which consists of two databases (Jolley *et al.* 2004). The profiles database contains all deposited nucleotide sequences, allelic profiles and sequence types, and the PubMLST database contains isolate-specific information. The PubMLST database can query the profiles database to obtain the nucleotide sequences for specific isolates, and contains additional information such as the study, country of origin, disease status of the carrier and serogroup. Whereas the profiles database contains only one complete nucleotide sequence of every allele identified to date, many of the entries in the PubMLST database will have the same allelic profile, and may have been sampled in the same, or different, studies.

1.1.3 Control and prevention

Strategies for prevention, or prophylaxis, and control are given different priority based on disease prevalence, economic costs and economic resources, all of which vary from country to country. While antibiotics are used to treat infected individuals and their close contacts (see section 1.1.1.1), control and prevention strategies make

use of vaccines for protecting as-yet uninfected members of the local population in the case of outbreak control, or the population at large in the case of prevention. Principally for economic reasons, vaccination of the population at large, if undertaken at all, is targeted at particular risk groups (see section 1.1.1.3), for example children, military recruits and the immunocompromised.

1.1.3.1 Polysaccharide vaccines

Bivalent (A, C) and tetravalent (A, C, Y, W-135) polysaccharide vaccines exist for meningococcal disease that contain the serogroup-specific capsular polysaccharide molecule. The bivalent vaccine was developed first, and extended because of significant disease caused by the other serogroups (Frasch 1995). In older children and adults, the efficacy of the serogroup A and C polysaccharides have been estimated to be 85-90% in clinical trials and epidemiological use, with a duration of protection of 5-10 years (Rosenstein *et al.* 1998). The polysaccharide vaccines are licensed for use in Europe and North America, but are not widely used because they do not induce strong or lasting immunological memory in the highest risk group, children under 2 years of age (Raghunathan *et al.* 2004). No polysaccharide vaccine for serogroup B meningococcal disease has been developed because of the low immunogenicity of the serogroup B capsular polysaccharide. This is thought to be owing to its close homology to a component of the human extracellular matrix (N-CAM).

The efficacy of serogroup Y and W-135 polysaccharide vaccines has not been investigated, but none of the polysaccharide vaccines substantially reduces meningococcal carriage rates, and as a result, does not induce herd immunity.

Repeated immunization has also been shown to result in immune hyporesponsiveness, although the clinical relevance of this is not well understood (Raghunathan *et al.* 2004). As a result, meningococcal polysaccharide vaccines are not part of the routine immunization schedule in any country. They are used in Europe and North America to protect members of high risk groups including patients suffering from asplenia (absent or defective spleen function that predisposes patients to fulminant bacterial infections), complement deficiency, military recruits, laboratory workers exposed to *N. meningitidis* and travellers to hyperendemic or epidemic areas (Pollard *et al.* 2001). Currently polysaccharide vaccines, in combination with antibiotics depending on the scale of the outbreak, are part of strategies for managing outbreaks in the West (Stuart 2001).

1.1.3.2 Polysaccharide-protein conjugate vaccines

The immunological shortcomings of polysaccharide vaccines are thought to result from the inability of the human T cell receptor to recognise the polysaccharide structure. Polysaccharide-protein conjugate vaccines work by binding the capsular polysaccharide to a protein carrier, which helps in T cell recruitment. This strategy has been successfully utilised in the *Haemophilus influenzae* type B vaccine (Hib; Heath 1998). To date, polysaccharide-protein conjugate vaccines have only been introduced in the United Kingdom, largely in response to concern over the rise in serogroup C meningococcal disease. The meningococcal serogroup C conjugate vaccine (MenC) was introduced into the routine immunization schedule in October 1999, with immunizations at 2, 3 and 4 months of age. Simultaneously, a campaign to immunize all children and young adults from 5 months to 18 years was initiated to induce widespread immunity. As a result there was an 81% reduction in the number

of confirmed cases of invasive meningococcal disease and deaths fell from 67 in 1999 to 5 in 2001. A 66% reduction in carriage in teenagers a year after vaccination has been documented, and substantial herd immunity was found in unvaccinated children who demonstrated a 67% reduction in carriage from 1998/1999 to 2001/2002. Potential problems such as capsule switching, in which the virulent strain undergoes recombination at the serogroup determining locus hence switching serogroup, and serogroup replacement, in which serogroup B disease might occupy the niche vacated by serogroup C disease, have not as yet presented themselves (Snape and Pollard 2005).

Other conjugate vaccines are under development, including a combined serogroup A and C vaccine (MenAC), trials of which were conducted in the United Kingdom and United States prior to the introduction of MenC in the U.K. (Snape and Pollard 2005). However, in North America the conjugate vaccine has only recently been licensed and a number of considerations suggest that it may not become part of the routine immunization schedule, including (i) the fact that polysaccharide vaccines are not currently used in routine immunization (ii) the probable absence of serogroup B from the vaccines (iii) the low prevalence of disease (iii) the cost of the vaccine and (iv) the crowding of the current immunization schedule (Pollard *et al.* 2001). It is likely that long-term, conjugate vaccines will replace polysaccharide vaccines in outbreak management (Stuart 2001). In Africa, where a lack of funding and vaccine research by pharmaceutical companies has led to the situation in which many countries that suffer from sporadic large scale epidemics do not have formal immunization strategies for preventing meningococcal outbreaks, the meningitis vaccine project (MVP), which is a collaboration between the World Health Organisation and the

Program for Applied Technology in Health, has been working on the development of a serogroup A conjugate vaccine (MenA). Clinical trials lasting for three years have begun, with licensure of the vaccine expected in 2008. The first use of the preventative MenA vaccine is anticipated to begin in 2009, with widespread vaccination initially targeted at high risk groups such as young children (Soriano-Gabarró *et al.* 2004).

1.1.3.3 Outer membrane protein vesicle vaccines

The poor immunogenicity of the serogroup B capsular polysaccharide in particular is concerning because serogroup B meningococci are responsible for much of the endemic meningococcal disease worldwide, including Europe and North America. After the capsule, class 1 OMPs are the next most immunodominant meningococcal antigen, followed by OMP classes 2 and 3. Patients suffering from meningococcal disease present bactericidal antibodies directed against these sub-capsular cell surface antigens. Recent research into serogroup B (MenB) vaccines has therefore concentrated on the development of outer membrane protein vesicle (OMV) vaccines.

There are several routes under investigation for OMV vaccine development. OMVs are naturally secreted from the meningococcus in the form of blebs, although they cannot be used in their native form. To prepare OMVs for a vaccine first requires the depletion of lipopolysaccharide (LPS), which is known to induce fever. Insoluble OMVs have been found to have poor immunogenicity, but combining the OMV with capsular polysaccharide makes the complex soluble and more efficacious. Similarly, adsorption of the vaccine on to aluminium hydroxide can increase the bactericidal response of the immune system. It is thought that vaccine efficacy might be further

improved by removal of class 4 OMPs that induce antibody blocking. A number of proteins expressed during pathogenesis are not expressed during natural growth, including iron regulated OMPs and heat-shock proteins, and could be important to a vaccine's immunogenicity. OMPs can be isolated with relative ease, but their native conformation is not conserved upon removal from the phospholipid membrane. Packaging isolated OMPs in such a way as to retain their natural conformation offers an alternative route to OMV vaccine development (Frasch 1995).

An important consideration in OMV vaccine development is the variety of serotypes defined by the subcapsular class 1 OMP. A given vaccine is raised against a particular serotype, so the long-term usefulness of that vaccine will depend both on the fluctuations in serotype prevalence and cross-protection between serotypes. There have been various trials of MenB OMV vaccines in Cuba, Norway, Chile, Brazil and Iceland (Sierra *et al.* 1991; Bjune *et al.* 1991; Zollinger *et al.* 1991; de Moraes *et al*. 1992; Perkins *et al.* 1998). Of these, children are generally less well protected than adults. Overall efficacy was between 50-80%, but in some studies young children had no protection. The duration of protection was short-lived, falling after 8 months. New Zealand introduced an OMV vaccine in 2004 in response to a 14-year epidemic of B:4:P1.7b,4. The vaccine is not predicted to offer broad cross-protection, but was introduced on the basis of the specificity of the epidemic and the high levels of meningococcal disease (Sexton *et al.* 2004). It has been suggested that OMVs based on a combination of two antigenic loci might offer better efficacy and long-term effectiveness (Urwin *et al.* 2004).

1.2 Population biology of *Neisseria meningitidis*

Genetic typing, in particular MLEE and MLST, has allowed patterns of genetic diversity in meningococcal populations to be quantified within and between geographic regions, sampling time points, and virulent and non-virulent isolates. Many studies of meningococcal disease and carriage have been undertaken which have shed light on the extent of diversity, structure of the population, frequency of recombination, influence of selection and overlap between disease-causing and carried strains. As larger-scale studies have been undertaken with MLST providing greater genetic discrimination, models of meningococcal evolution have been proposed and revised. I will discuss the progression of these studies, the techniques used in their analysis and the development of the evolutionary models used to explain them.

1.2.1 The clonal complex

Patterns of genetic diversity revealed by MLEE clearly demonstrate that the population structure of disease-causing *N. meningitidis* is organised into closely-related, genetically homogeneous clusters, which can be visualised through UPGMA dendrograms (Sneath and Sokal 1973; Box 1). The genetic clusters tend to be strongly associated with particular serogroups; serogroup A clusters are known as *subgroups*, whereas in serogroup B and C meningococci the terms *complex* and *cluster* are used. These clusters, or complex of clones (Caugant *et al.* 1988), are routinely recovered from geographically disparate locations over periods of more than 10 years and exhibit strong linkage disequilibrium between loci, suggesting that populations of

Box 1 – Building a UPGMA dendrogram

Initially, there are as many clusters as there are genotypes, and the genetic distance d_{ij} between clusters *i* and *j* is defined as the proportion of loci at which isolates *i* and *j* have different alleles. The number of isolates in cluster *i* is defined initially to be $n_i = 1$.

1. Join the set of clusters *C* that have the smallest distance from one another.
2. Call the new cluster *i* and define the genetic distance between *i* and each of the other clusters *j*, $j \notin C$ as $d_{ij} = \frac{\sum_{c \in C} n_c d_{cj}}{\sum_{c \in C} n_c}$ and let $n_i = \sum_{c \in C} n_c$.
3. If there is more than one cluster left, return to step 1.

meningococci are basically clonal in structure (Caugant *et al.* 1986; Caugant *et al.* 1987).

The use of MLEE has allowed the epidemiological spread of meningococci to be charted, the results of which have demonstrated that clonal complexes differ in their propensity to cause disease, rate of transmission and extent of global dissemination. Only a small number of clonal complexes are responsible for most of the disease worldwide (Caugant *et al.* 1988), the so-called hyper-virulent and hyper-invasive lineages. The clonal complex is thought to constitute the basic unit for epidemic spread (Achtman 1995).

1.2.1.1 Serogroup A lineages

Analysis of serogroup A *N. meningitidis* strains isolated from all major global epidemics in the period 1960-1990 divided the population into a small number of

genetically homogeneous clusters, or subgroups (Wang *et al.* 1992). Eighty-four unique ETs were identified amongst the 290 isolates, and their genetic relationship can be visualised using a UPGMA dendrogram (Figure 8), with the subgroups colour-coded. Genetic distance is defined as the proportion of loci at which a pair of ETs have different alleles (Selander *et al.* 1986). Figure 8 shows that the subgroups are genetically homogeneous. Within each subgroup there is a highly skewed frequency distribution of ETs, with one or two common ETs, and many rare ETs differing from one another at a small number of loci. The genetic distance between subgroups is generally much greater than the average distance within a subgroup. Serosubtypes are highly conserved within subgroups, with most ETs exhibiting a common PorA VR1/VR2 combination.

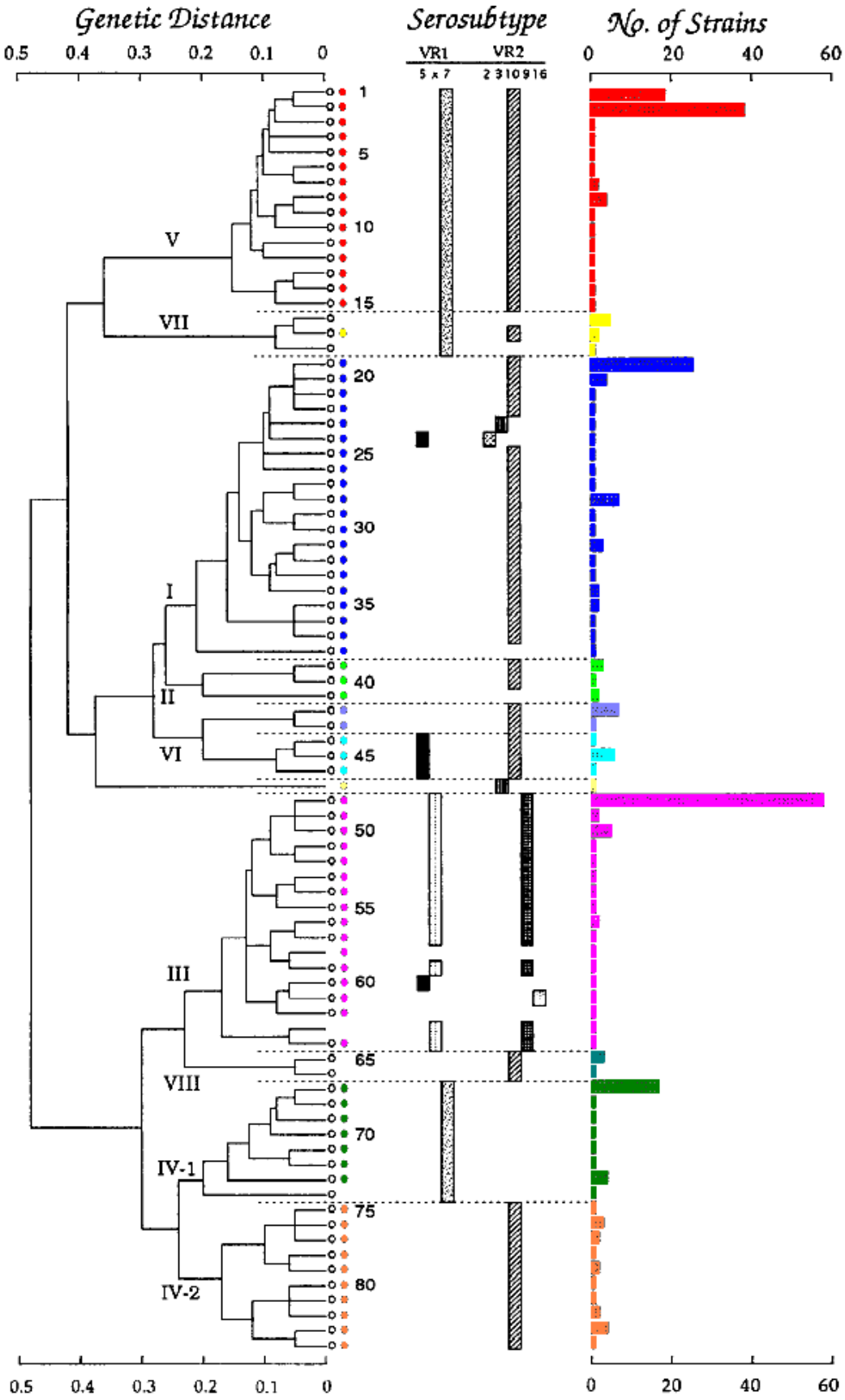


Figure 8 Genetic relationships, serosubtype patterns and relative abundance of 84 ETs in a global sample of disease-causing serogroup A meningococci. Adapted from: Wang et al. (1992).

Within serogroup A, subgroups I/II, III/VIII and IV-1/IV-2 are identified as hyper-virulent lineages (Maiden *et al.* 1998), and these groups have been historically linked to specific epidemics worldwide (Achtman 1995):

- S**ubgroup I** was first isolated in the United Kingdom in 1941, although the subgroup may have originated elsewhere. Since the 1960s, subgroup I meningococcal disease has been responsible for epidemics affecting Niger, North Africa, the Mediterranean, native Americans living in Canada, homeless people in the United States, Nigeria, Rwanda and native peoples of New Zealand and Australia, over a time frame of thirty years. ETs belonging to subgroup I have also caused endemic disease globally.
- **Subgroup III** isolates are first known from China in the 1960s, from where the cluster has spread causing outbreaks in Russia and Norway, then Finland and Brazil in the 1970s, Nepal and China in the 1980s and throughout continental Africa in the late 1980s and early 1990s. It was subgroup III that caused the Haj pilgrimage outbreak in 1987. An estimated 10% of the 1,000 U.S. pilgrims to Mecca returned home carrying subgroup III meningococci. Thereafter sporadic cases were reported in the U.K., France, Israel and the Gambia. Subgroups III meningococci have also been responsible for endemic disease.
- **Subgroup IV-1** is, in contrast, almost entirely restricted to endemic disease in West Africa, persistently isolated over a period of 40 years. Except for two waves of subgroup I epidemics in the 1970s, subgroup IV-1 has been responsible for all epidemic disease isolated from West Africa in the same time period.

- **Subgroup V** bacteria are similarly geographically restricted. In their case no subgroup V strains have yet been isolated anywhere but in China, where they caused an outbreak in the 1970s.

1.2.1.2 Serogroup B and C lineages

Disease-causing isolates belonging to serogroups B and C are less uniform than serogroup A meningococci, and genetic clusters identified in one of these serogroups often contain some isolates expressing the other serogroup, as a result of recombination (Caugant *et al.* 1986). Sub-capsular antigenic expression is also less homogeneous (Caugant *et al.* 1987). Several groups important for disease exist within serogroups B and C that comprise highly genetically similar, low-frequency ETs clustered around a common ET (Achtman 1995):

- **ET-5 complex** bacteria typically belong to serogroup B and are responsible for much endemic disease around the world. ET-5 complex meningococci have caused epidemics in Cuba, Chile, Brazil and New Zealand since 1970, prior to which their isolation was rare. As a result of their global endemicity, reconstructing the spread of particular epidemics has proved difficult.
- **A4 cluster** isolates originated from South Africa and the U.S. in the late 1970s and early 1980s (Caugant *et al.* 1987), and were sampled contemporaneously in Canada and Europe. A4 cluster isolates typically express serogroup B but serogroup C isolates have been associated with increased disease incidence in Brazil since the 1990s.
- **ET-37 complex** meningococci principally express serogroup C. Responsible for disease outbreaks amongst U.S. military recruits in the 1960s, ET-37

complex bacteria have been isolated from endemic infection globally, including North America, Europe, Africa and Asia.

1.2.2 How clonal are bacteria?

Problems exist for the idea that *N. meningitidis* has a basically clonal population structure. Firstly, recombination, which occurs by transformation of naked DNA in the meningococcus, is needed to explain the antigen switching observed not just in serogroup B and C complexes, but also serogroup A complexes and many carriage isolates (Caugant *et al.* 1987; Caugant *et al.* 1988). Secondly, in light of the fact that recombination is known to occur at some level, the use of dendrograms is questionable (Holmes *et al.* 1999). Thirdly, strong levels of linkage disequilibrium may occur in spite of recombination for several reasons (Maynard Smith *et al.* 1993):

i. If the sample contains multiple populations, within which recombination is common, but between which it is rare, then there will be linkage disequilibrium.
ii. Drift causes non-zero linkage disequilibrium even in the presence of random mating.
iii. Epidemic population structure can lead to linkage disequilibrium.
iv. Epistatic fitness interactions between loci can maintain linkage disequilibrium.

Objection (i) applies to any analysis of datasets in which disease-causing isolates are overrepresented relative to carriage. If there exist different subpopulations of *N. meningitidis* that have different propensities to cause disease, and if disease-causing isolates are not a random sample of meningococcal isolates at large, then the problem will be exacerbated. Objection (ii) applies to any population of finite size.

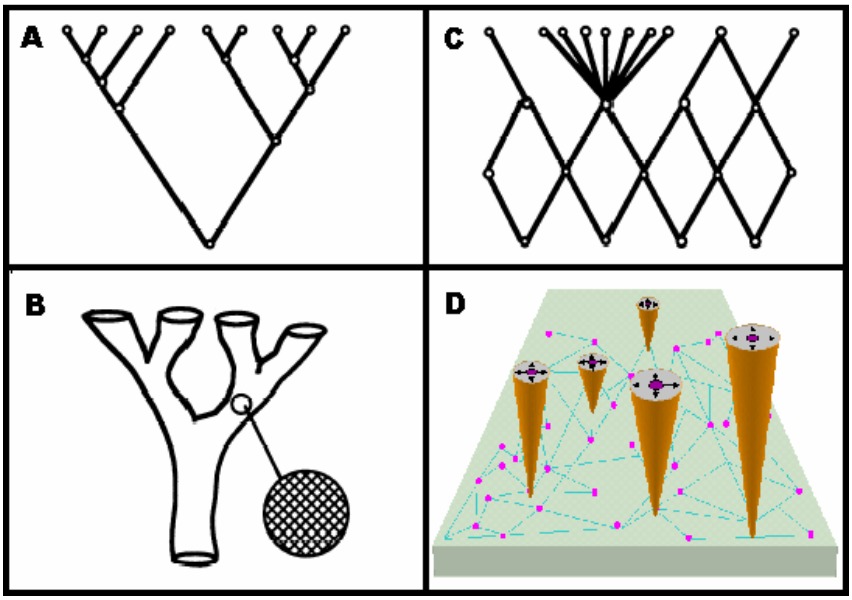


Figure 9 Representation of population structures. **A** Strict clonality is represented by a bifurcating evolutionary tree. **B** Frequent recombination within reproductively isolated subpopulations is represented by a network within a bifurcating population tree. **C** Epidemic clone model in which recent clonal expansion (star-shaped tree) occurs against a backdrop of frequent recombination (network). **D** Alternative representation of the epidemic clone model in which cones represent recent clonal expansion superimposed on to a network of recombination. Source: Maynard Smith *et al.* (1993) and Maynard Smith *et al.* (2000).

Determining what level of linkage disequilibrium (LD) is significantly different to zero is a statistical problem. Objections (iii) and (iv) are the subject of the epidemic clone model of Maynard Smith *et al.* (1993) and the strain theory model of Gupta *et al.* (1996) respectively.

1.2.2.1 Epidemic clone model

Maynard Smith *et al.* (1993) proposed an epidemic model of population structure in which a population undergoing frequent recombination may exhibit high linkage disequilibrium because of a recent epidemic during which a particular lineage

Box 2 – Index of Association

Suppose p_{ij} is the frequency of allele *i* at locus *j*, that $h_j = 1 - \sum p_{ij}^2$ is the probability that two isolates differ at locus *j*, and that *K* is the genetic distance between a pair of isolates, defined as the number of loci at which they differ. Then

$$I_A = V_O / V_E - 1$$

is the index of association, where V_O is the observed variance in *K* and $V_E = \sum h_j (1 - h_j)$ is the expectation of V_O under the null hypothesis of linkage equilibrium. The standard error is calculated using

$$\mathrm{var}(V_E) = \frac{1}{n}\left(\sum h_j - 7\sum h_j^2 + 12\sum h_j^3 - 6\sum h_j^4 + 2\left[\sum h_j - \sum h_j^2\right]^2\right).$$

undergoes rapid clonal growth (Figure 9C,D). This they contrast against a model of strict clonality (Figure 9A), and a model of reproductively isolated populations each of which exhibits frequent recombination (Figure 9B). Maynard Smith *et al.* (1993) use a statistical test for recombination called the index of association (I_A; Brown *et al.* 1980; see Box 2), whose expectation is zero under the null hypothesis of frequent recombination.

Analysis of a collection of over 600 serogroup A, B and C disease-causing isolates and carriage isolates (Caugant *et al.* 1987) yields $I_A = 1.96 \pm 0.05$, which is statistically significant from zero. Thus the null hypothesis of linkage equilibrium, and hence frequent recombination in a panmictic population, in these isolates is rejected, consistent with the observation of strong linkage between MLEE loci (Caugant *et al.* 1987). However, when each of the 37 ET clusters identified by

Caugant *et al.* (1987), is treated as a single individual, $I_A = -0.14 \pm 0.17$. Maynard Smith *et al.* (1993) argue that this is evidence for an epidemic population structure; that each ET cluster is the result of recent growth of a particular clone against a backdrop of frequent recombination. Subsequent studies of MLST data have shown the same pattern of significant I_A for all isolates, but non-significant I_A when each cluster is treated as a single individual (Holmes *et al.* 1999; Jolley *et al.* 2000).

There are a number of problems with the model and analysis. Firstly, the epidemic clone model provides a description of the population structure, but not a description of the evolutionary processes that cause the population structure. Secondly, if recombination is an important process in meningococcal evolution, then it is not clear that epidemic clusters can be identified using a UPGMA dendrogram. Indeed, for a recombining population "there is no justification for constructing trees: one might as well construct a tree for the members of a panmictic sexual population" (Maynard Smith *et al.* 1993). Yet it is unclear how to identify members of an epidemic cluster without a full specification of the evolutionary model. Despite these complaints, the epidemic clone model is useful in illustrating that a clonal view of meningococcal evolution is unsatisfactory.

1.2.2.2 Relative contribution of recombination and mutation

Numerous attempts have been made to quantify the extent of recombination in meningococcal populations. These studies have benefited from the greater resolution afforded by nucleotide sequencing and MLST, which reveals synonymous nucleotide polymorphism that is invisible to MLEE. There is a large body of evidence supporting the importance of recombination in meningococcal evolution from a number of

sources (Feil and Spratt 2001). Mosaicism has been observed in the nucleotide sequences of housekeeping genes (Zhou and Spratt 1992; Feil *et al.* 1995; Feil *et al.* 1996; Zhou *et al.* 1997), which can only be explained by frequent recombination. Comparison of meningococcal sequences with homologues in other neisseriae suggests that importation from closely related commensals may be an important process in addition to intraspecific recombination (Zhou *et al.* 1997; Linz *et al.* 2000). Furthermore, splits decomposition (Bandelt and Dress 1992; Dopazo *et al.* 1993) indicates that the evolutionary history of meningococcal housekeeping genes is better represented by a network than a strictly bifurcating tree which would be produced under clonality (Holmes *et al.* 1999).

In a frequently recombining organism there is no sense in which there is a single phylogenetic tree for a collection of isolates. Recombination will cause there to be different phylogenies at different positions in the genome. The frequency of recombination determines the extent to which these trees are correlated. Therefore the degree of incongruence between phylogenetic trees at distinct loci is a way to quantify the extent of recombination in a population. A subset of 30 out of a global sample of 107 predominantly disease-causing meningococci (Maiden *et al.* 1998) were analysed to quantify the effect of recombination on phylogenetic congruence (Holmes *et al.* 1999; Feil *et al.* 2001). Under the null hypothesis of complete linkage in the absence of recombination, all loci share the same phylogenetic tree topology. For each MLST locus a maximum likelihood (ML) tree was estimated. To test for congruence between the ML tree topology at each locus and all the others, the difference in log likelihood was calculated, having re-optimised the branch lengths for the other trees. A null distribution for the difference in log likelihood was produced using 200 bifurcating

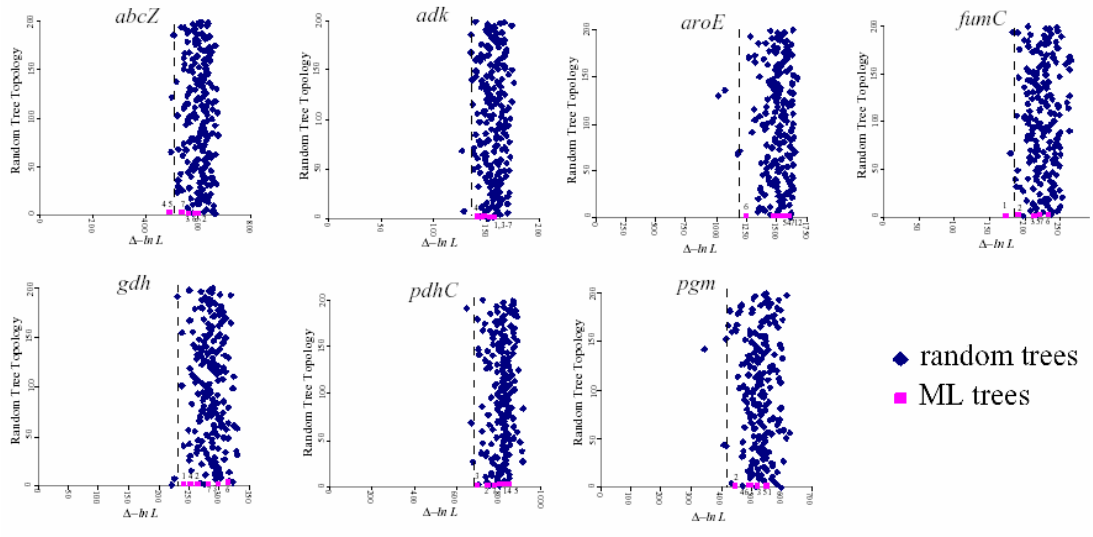


Figure 10 Phylogenetic incongruence amongst MLST loci for 30 isolates representative of global disease. Horizontal axis is the difference in log likelihood between the ML tree for each locus and the ML tree for another locus (pink squares) or a random tree (blue diamonds). Trees are spread vertically in no particular order. Trees to the left of the dashed line are more congruent with the ML tree for that locus than 99% of the random trees. Source: Feil *et al.* 2001, supplementary material (http://www.pnas.org).

topologies simulated uniformly at random. Figure 10 shows that the difference in log likelihood between the ML topology at each locus and the six others (pink squares) is not significantly less than for the random topologies (blue squares). The extent of recombination in *N. meningitidis* is therefore sufficient to create phylogenetically incongruent trees within a 450bp sequence (Feil *et al.* 2001).

Holmes *et al.* (1999) suggested that mutation may not be the primary route by which new allelic variants arise in the meningococcus. MLST data allows the role of mutation and recombination to be disentangled because the nucleotide differences between alleles can be examined. Point mutation changes a single site at a time, whereas recombination can cause mosaicism wherein a whole tract is imported from

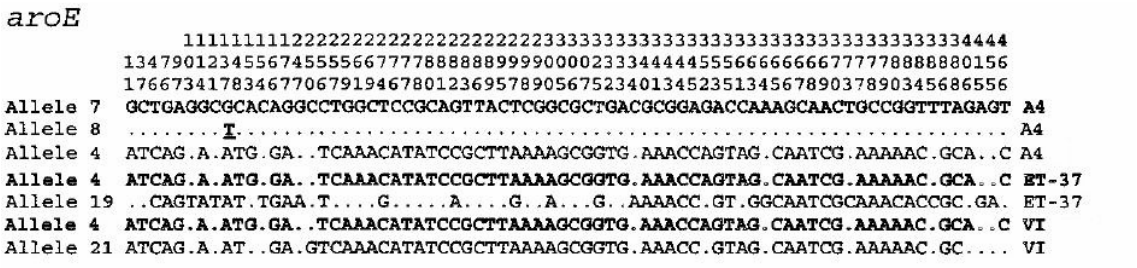


Figure 11 Polymorphic sites in *aroE* within three clonal complexes. The clonal complexes are labelled at the far right. For each clonal complex there is one bold row representing the central, or most common, genotype. The non-bold rows represent all other variants assigned to that clonal complex. Dots indicate that the allele does not differ from the top row sequence. Sites are numbered above in vertical format. Sites that are underlined are putative point mutations. Adapted from: Feil *et al.* 1999.

another allele in the population. In the 107 isolates representing global disease (Maiden *et al.* 1998), nine out of 10 clonal complexes defined by a UPGMA dendrogram contain isolates that differ by one or two loci (Feil *et al.* 1999). In total there are 23 isolates that differ from the ST typical of that clonal complex. Of those, 5 differ from the typical genotype at a single nucleotide position, indicative of a point mutation. The rest differ by five or more polymorphisms, which Feil *et al.* (1999) therefore took to be the result of recombination.

For example, Figure 11 illustrates polymorphism within clonal complexes at the *aroE* locus. In three clonal complexes (A4 cluster, ET-37 complex and subgroup VI) a low frequency variant differs from the common, or central, genotype at *aroE*. Of those, only one (allele 8 in the A4 cluster) differs from the typical genotype by a single nucleotide polymorphism (G→T, underlined in Figure 11). Feil *et al.* (1999) use these observations to estimate that the relative frequency of recombination events to mutation events is 18:5, or 3.6 times as often. Subsequent studies have used the same procedure to estimate that recombination is 4.75 times (Feil *et al.* 2001) or 275 times

(Jolley *et al.* 2000) more frequent than mutation. Whilst there are obvious problems with the estimation procedure, not least of which the lack of quantification of uncertainty, these figures show that recombination is a potent evolutionary force in meningococci, and that most allelic novelty probably arises by recombination of existing alleles rather than *de novo* mutation.

1.2.2.3 BURST

Recombination is an important force in meningococcal evolution, and despite the presence of clusters of closely related genotypes, phylogenetic congruence is all but obliterated even within a 450bp gene fragment (Feil *et al.* 2001). A bifurcating tree is an inadequate description of the ancestry of a collection of meningococcal genotypes (Holmes *et al.* 1999), so the identification of clonal complexes on the basis of UPGMA dendrograms is questionable. Therefore the question arises as how to identify and visualize clusters of meningococcal genotypes.

From the perspective of the epidemic clone model (Maynard Smith *et al.* 1993), a clonal complex is a group of individuals descended from a founding genotype that had a fitness advantage allowing it to proliferate rapidly in the population. As the clonal complex expanded over time it will have experienced mutation and recombination events leading to divergence from the founding genotype. MLST shows that clusters of meningococcal genotypes exist, in which frequent genotypes are closely related to numerous low-frequency genotypes, separated not necessarily by mutation, but commonly by recombination events. Linkage disequilibrium may appear to be high not because of strict clonality, but because short, recent explosive

Box 3 – The eBURST Algorithm

Clonal complex eBURST groups all STs into connected, mutually exclusive sets within which every ST differs from at least one other by no more than a single locus.

Primary founder Within each clonal complex, the ST that differs from the greatest number of STs by no more than a single locus (single locus variants, SLVs), is defined to be the primary founder. In the event of a tie, the number of double locus variants (DLVs) is taken into account, and so on. The frequency of the STs does not come into consideration.

Bootstrap support for the primary founder is obtained as follows. Within the clonal complex, a collection of STs the same size as the number of unique STs is resampled, with replacement, from those unique STs. The primary founder of that collection is then determined. The procedure is repeated 1,000 times, and the bootstrap support for a particular ST is the percentage of resampled collections in which it was determined to be the primary founder. Resampled collections in which that particular ST was not present are excluded.

Subgroups and subgroup founders Moving outwards from the primary founder, STs are defined to be subgroup founders if they are SLVs of two or more STs not currently connected. Having connected all members of the clonal complex, starting farthest from the primary founder, the subfounder-descendant relationship can be reversed if that would increase the number of SLVs connected to the subgroup founder.

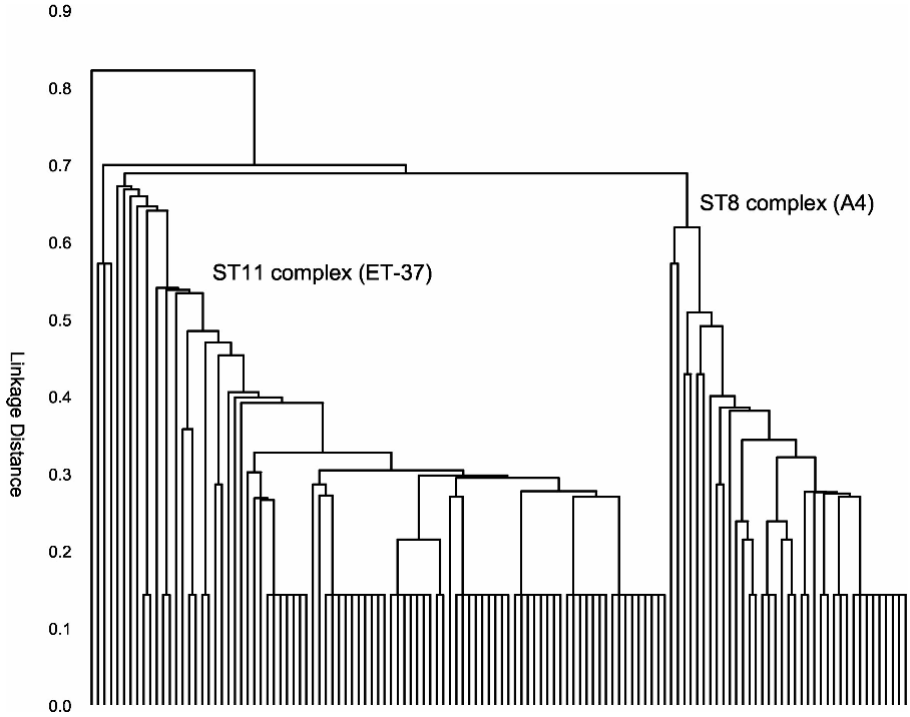


Figure 12 UPGMA dendrogram of the ST8 (A4 cluster) and ST11 (ET-37) clonal complexes. All STs with fewer than 4 alleles different to ST 8 or ST11 were included from the *Neisseria* MLST database. Source: Feil *et al.* (2004).

bursts of selected clones causes LD to temporarily accumulate faster than it can be broken down by recombination.

eBURST (Based Upon Related Sequence Types) is a deterministic algorithm used for clustering STs based on a more realistic account of meningococcal evolution (Feil *et al.* 2004). Although eBURST is non-parametric in the same sense that the UPGMA dendrogram is non-parametric, it does use the informal model of an epidemic population structure to inform the rules used in the clustering algorithm (Box 3). Figure 12 shows a UPGMA dendrogram for the ST8 and ST11 clonal complexes (formerly the A4 cluster and ET-37 complex) for all STs in the *Neisseria* MLST database that differ from ST8 or ST11 at less than 4 loci. Figure 13 shows the

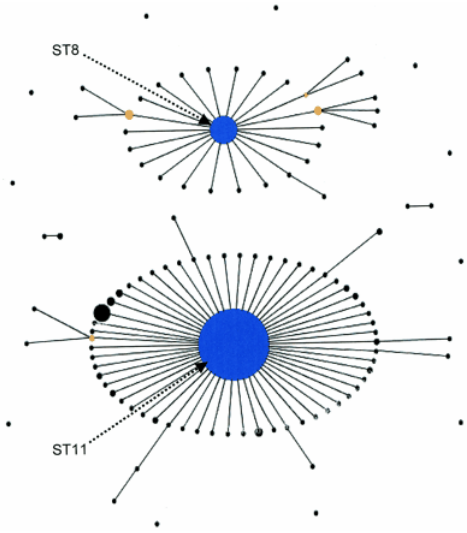


Figure 13 eBURST diagram of the group containing the ST8 and ST11 clonal complexes. A group was defined as STs differing by less than 3 alleles from one another. Clonal complexes (connected nodes in the diagram) were defined as STs differing by less than 2 alleles from one another. Source: Feil *et al.* (2004).

corresponding eBURST diagram for these clonal complexes, and closely related STs (Feil *et al.* 2004). Whereas the dendrogram, of necessity, depicts a hierarchical population structure within each clonal complex, the eBURST diagram depicts a frequent founding genotype (represented by the diameter of the node) surrounded by many SLVs (single locus variants), a number of which may be founders of subgroups themselves. So whilst the dendrogram is constrained to portray a hierarchical population structure, the eBURST diagram is able to, but actually portrays a radiation of rare, closely related genotypes surrounding a core genotype. eBURST assigns bootstrap support of 100% for the primary founders of the ST8 and ST11 complex. Whilst the dendrogram suggests that the ST8 and ST11 complex are closely related, eBURST identifies them as distinct entities and does not, therefore, infer the relationship between the two.

The identification of clonal complexes in *N. meningitidis* is now informed by a combination of historical convention (largely influenced by epidemiological considerations and UPGMA dendrograms), the clusters identified using the eBURST algorithm, and a committee of microbiologists (chosen from amongst delegates of the International Pathogenic Neisseria Conference). As a result much of the nomenclature has been changed, and continues to be revised. The A4 cluster and ET-37 complex have become the ST8 and ST11 complex respectively. Serogroup A subgroups have been merged and renamed to create the ST1 complex (subgroups I/II), ST 5 complex (subgroups III/VIII) and ST 4 complex (subgroups IV-1/IV-2).

However, there are problems with the eBURST algorithm, and hence the clonal complexes it produces. These problems are essentially the result of the non-parametric nature of eBURST. In the absence of a statistical model, it is impossible to assign uncertainty to the groupings. It is likely that a statistical description of the epidemic clone model would report considerable uncertainty in the group designations. It is not clear what the null model is for the bootstrap support that eBURST calculates for the assignment of primary founders (Box 3), and at any rate it is calculated conditional upon the groupings, the reliability of which is unknown. The primary founder is not necessarily present in the sample, and in the absence of an explicit evolutionary model it is impossible to comment on the ancestral relationships between clonal complexes, or the age of those complexes. Finally, there is no framework for falsifying the model if the fit is poor.

1.2.3 Strain theory

An epidemic population structure is not the only explanation for high levels of linkage disequilibrium despite frequent recombination. One alternative recognised but not explored by Maynard Smith *et al.* (1993) is that selection can cause non-random associations of alleles across loci. Epistasis between loci can cause LD not just at the loci under selection, but across the genome. Strain theory suggests that the interaction between the host immune system and antigenic loci can cause epistasis if there is some immunological cross-protection between alleles at individual loci. This epistasis might explain the paradox that pathogens such as *N. meningitidis* appear to persist as strains despite the constant exchange of genetic material (Gupta *et al.* 1996).

1.2.3.1 Immune selection can structure the pathogen population

Suppose there are two distinct loci, A and B that encode antigens. Both loci are dimorphic (Figure 14). There are four genotypes (i.e. combinations of alleles at the two loci) ■●, ■○, □● and □○. Genotypes that are different at both loci are called discordant. So ■● and □○ are discordant, and ■○ and □● are discordant. Gupta *et al.* (1996) propose an SI-type model (Anderson and May 1991) that is defined by the differential equations

$$\frac{dz_i}{dt} = \lambda_i(1 - z_i) - \mu z_i,$$
$$\frac{dy_i}{dt} = \lambda_i(1 - z_i)[1 - \gamma(1 - \varphi_i)] - \sigma y_i,$$

where z_i and y_i are the proportion of hosts susceptible to and infectious with genotype i respectively. $1/\mu$ and $1/\sigma$ are life expectancy and duration of infectiousness respectively in the host. λ_i is the per-capita force of infection for genotype i, which is the rate at which susceptible hosts become infected. λ_i equals the transmission

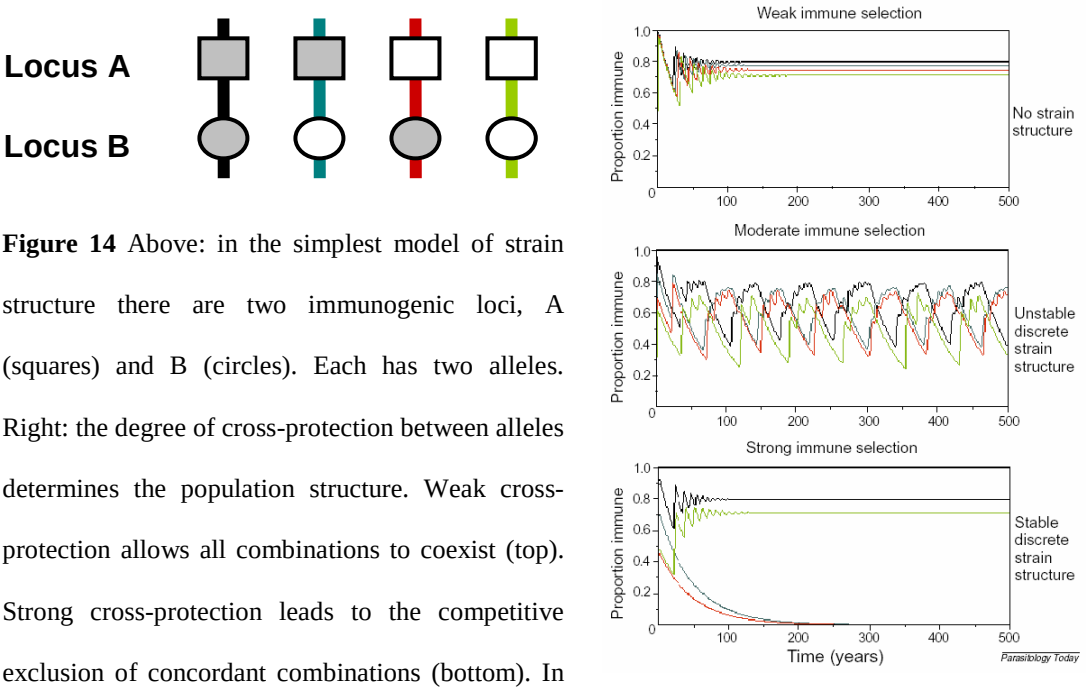


Figure 14 Above: in the simplest model of strain structure there are two immunogenic loci, A (squares) and B (circles). Each has two alleles. Right: the degree of cross-protection between alleles determines the population structure. Weak cross-protection allows all combinations to coexist (top). Strong cross-protection leads to the competitive exclusion of concordant combinations (bottom). In between lies unstable switching between discordant pairs (middle). Source: Gupta and Anderson (1999).

coefficient β_i multiplied by the frequency of genotype i in the host population following recombination; the loci are assumed to be unlinked. φ_i is the proportion of the host population with immunity to genotypes concordant to i, and γ is the degree of cross-protection afforded against a concordant genotype.

Figure 14 shows that the key parameter determining the behaviour of the model is the degree of cross-protection, γ. When there is weak cross-protection, so that encountering a particular allele in one genotype confers no immunity to that allele in other genotypes, all genotypes can coexist. When cross-protection is strong, concordant pairs are in direct competition, resulting in exclusion of one or other discordant pair. In the model which pair is out-competed depends on the initial genotype frequencies. At intermediate levels of cross-protection, the population switches between discordant pairs intermittently.

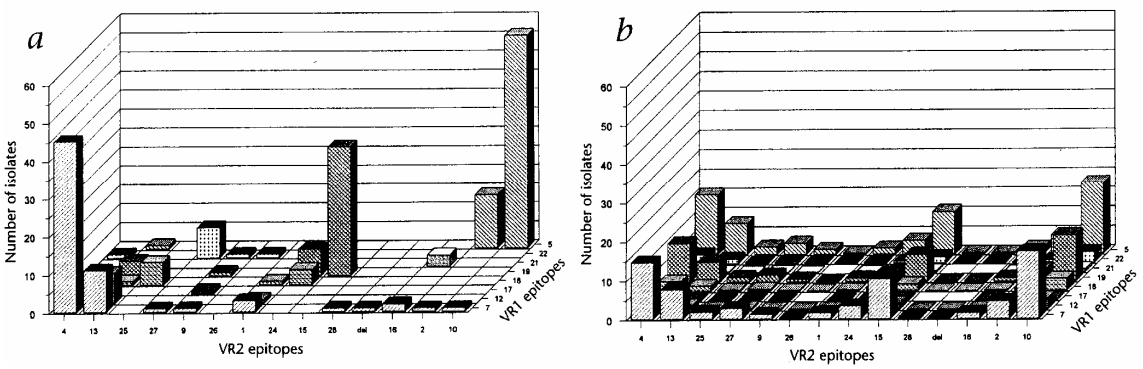


Figure 15 Association between epitopes of the VR1 and VR2 region of the PorA outer membrane protein. (**a**) Observed frequency distribution. (**b**) Expected frequency distribution under linkage equilibrium. Source: Gupta *et al.* (1996).

1.2.3.2 Evidence for meningococcal strain structure

Gupta *et al.* (1996) suggest that the population structure of meningococci can be explained by immune selection causing exclusion of immunologically overlapping genotypes. Serogroup B and C meningococci sampled from England and Wales in 1989-1991 were serosubtyped for the VR1/VR2 combination at the *porA* locus (Feavers *et al.* 1996). Figure 15a shows the observed frequencies of VR1/VR2 epitope combinations, and Figure 15b the expected frequencies under linkage equilibrium. What is striking about Figure 15a is that broadly speaking each VR1 epitope is associated with only a single VR2 epitope at any appreciable frequency (and vice versa). Not only do the data reject the null hypothesis of random association ($p < 0.01$ based on a χ^2 test with 15 d.f.), but the non-random association of particular epitopes to the exclusion of other combinations is the pattern predicted by the strain theory model (Gupta *et al.* 1996).

However, there are a number of problems with this analysis. In a finite population drift can cause associations between loci, that is LD, even at unlinked loci by chance alone. The VR1/VR2 regions of *porA* are tightly linked, so LD would be expected to be even higher. Thus zero LD is not the appropriate null model. Because the appropriate null model has not been tested, it is not possible to be sure that the associations of VR1/VR2 epitopes are non-random after all. Secondly, a statistic sensitive to the mutual exclusivity of genotypes imposed by strain structure, and not merely to LD *per se*, would be required to show that immune selection is responsible for the observed LD, and not some other process.

1.2.4 Neutral models

Owing to high carriage rates and low incidence of disease, it has been postulated that *N. meningitidis* is an accidental pathogen (Levin and Bull 1994; Maiden 2002). That is to say, that disease-causing strains are so rare that they cannot possibly be important for transmission or long-term persistence of meningococcal populations. Indeed, most epidemics are relatively modest in size, and subsequently die out, suggesting that pathogenicity might be an evolutionary dead end for the meningococcus (Stollenwerk *et al.* 2004). If virulence is indeed detrimental, or at best equivocal to the evolutionary success of meningococci, then selection for epidemic-causing variants may not be an important explanation for the structure of meningococcal populations. As noted by Maynard Smith *et al.* (1993), drift alone can cause non-zero levels of linkage disequilibrium in a finite population. Studies show that a purely neutral model with drift does not adequately explain the observed patterns of genetic diversity even in carriage studies (Fraser *et al.* 2005). However, Fraser *et al.* (2005) claim that when the effects of local transmission or sampling bias

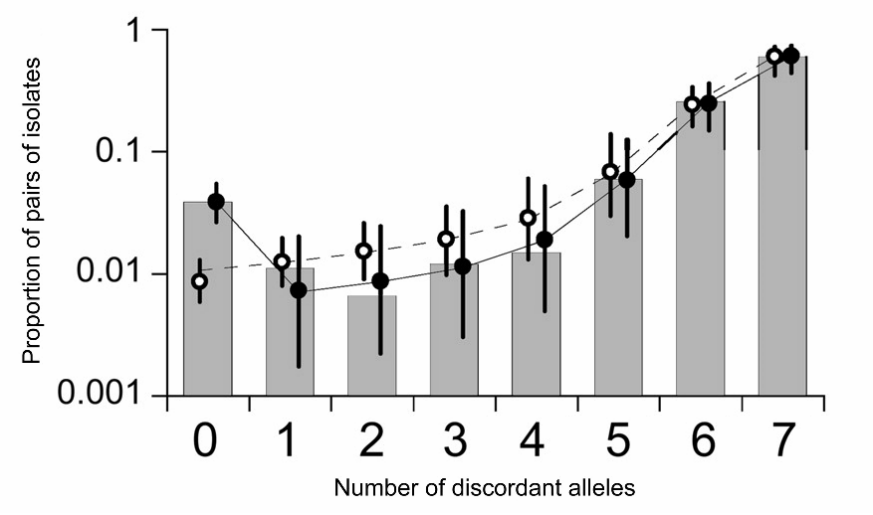


Figure 16 Allelic mismatch distribution for Czech carriage study (grey bars). The horizontal axis shows the number of loci at which a pair of isolates can differ (up to 7 for MLST), and the vertical axis the proportion of pairs that differ at that number of loci. Open circles show the fit under the standard neutral model, and the filled circles show the fit under the neutral microepidemic model. Source: Fraser *et al.* (2005).

are taken into account, meningococcal evolution may amount to no more than a neutrally evolving commensal with the occasional accidental progression to pathogenesis. This they call the neutral microepidemic model.

1.2.4.1 Standard neutral model

In the approach of Fraser *et al.* (2005), the patterns of genetic diversity observed in the population are summarised by a small number of statistics. Some of these statistics are used to estimate the parameters for the model, and the model is then assessed for goodness-of-fit. Figure 16 shows the observed allelic mismatch distribution (grey bars) in a population of carried meningococci sampled from the Czech Republic in 1993 (Jolley *et al.* 2000). The allelic mismatch distribution shows the proportion of pairs of individuals that differ at 0, 1, …, 7 loci.

In the simplest evolutionary model, known as the standard neutral model, members of the population reproduce with equal vigour. The population size is constrained so that it remains at a constant size *N*. Each generation the total rate of mutation amongst all individuals is $\theta/2$ per base pair, and the total rate of recombination amongst all individuals is $\rho/2$ per base pair. Under the standard neutral model with infinitely many alleles, the expected frequency of each of the grey bars in Figure 16 is known (Kimura 1968). Fraser *et al.* (2005) assume that the allelic mismatch distribution is multinomially distributed with these expected frequencies. In an attempt to account for the fact that this is wrong, owing to the non-independence of the different classes in the allelic mismatch distribution, they calculate standard errors by taking the observed degrees of freedom to be n rather than $n(n-1)/2$, where $n = 217$ is the number of isolates.

The estimated population rates of mutation and recombination are $\theta = 8.2$ and $\rho = 5.7$ respectively, which suggests that recombination events occur 1.44 times less frequently than mutation events, in contrast to previous work (Feil *et al.* 1999; Feil *et al.* 2001), including analysis of the same data (Jolley *et al.* 2000). No confidence intervals were published, even using the approximate correction for the degrees of freedom. Simulations using the estimated parameters did not produce the observed allelic mismatch distribution (open circles, Figure 16). The observed homozygosity (proportion of identical isolates) lay above the standard error, indicating that the standard neutral model is inadequate to explain the observed patterns of genetic diversity in a population of carried meningococci (Fraser *et al.* 2005).

1.2.4.2 Neutral microepidemic model

The neutral microepidemic model is a mathematically simple extension to the treatment of the standard neutral model by Fraser *et al.* (2005), based on the idea that in natural populations of infectious agents there exist localised transmission chains. If a sample contains multiple isolates from the same short transmission chain, or microepidemic, then there will be an excess of homozygosity (Fraser *et al.* 2005). In a eukaryote this would be analogous to assembling a population sample taking multiple members of the same family. So the model is essentially neutral evolution with biased sampling.

An extra parameter, h_e, is added to the neutral model which allows homozygosity (the proportion of individuals that are identical) to vary freely from the mutation and recombination rates. As a result, the observed and expected homozygosity match exactly (Figure 16, filled circle, 0 discordant alleles). This simple extension appears to fit the data well, because the rest of the allelic mismatch distribution lies well within the standard errors from simulation (filled circles, Figure 16). Under this model the parameter estimates were $\theta = 10.2$ and $\rho = 13.6$, suggesting that recombination occurs 1.33 times more frequently than recombination, which agrees better with previous work. Fraser *et al.* (2005) modelled the biased sampling scheme as taking an average of σ individuals from each of n_c microepidemic transmission chains. Using a simple relationship between the observed homozygosity and these parameters, estimates of $n_c = 9$ and $\bar{\sigma} = 13.1$ are obtained. This suggests that nine microepidemic clusters have been over-represented by an average of 13.1 isolates.

Interestingly, Buckee *et al.* (2004) have used simulations to show that when the meningococcal population is subdivided because of clustering in the host contact network, such as in the microepidemic model, strain theory predicts that structuring of meningococci into antigenically discordant types will only occur locally. Because of the random way in which a particular set of antigenically discordant types come to predominate locally, no particular set will predominate across the population as a whole, so elevated LD between loci at the level of the whole population is no longer predicted. As a result, strain theory and the neutral microepidemic model appear to be mutually exclusive explanations for elevated patterns of LD in meningococci.

There are several advantages to formulating an explicit statistical model, such as the neutral microepidemic model. Firstly, the parameters can be estimated and the uncertainty in these estimates quantified (although the latter was not performed in this case). Secondly, by making the model mathematically explicit its interpretation is less vulnerable to vague verbal reasoning, and more precise hypotheses can be evaluated. Thirdly, the model can be used to make predictions, including predicting other aspects of the data. These predictions can then be used to validate the model. Fraser *et al.* (2005) showed that the nearest-neighbour distribution (the distance to each isolate's most similar non-identical neighbour) simulated under the estimated parameters was a good fit to the observed distribution. Goodness of fit testing allows the model to be falsified if it is a poor description of the data, although a more thorough investigation than performed in this example might be carried out. MLST provides full nucleotide sequence data from loci distributed around the genome. The approach used by Fraser *et al.* (2005) discards much of that information by reducing the sequence information into the number of pairwise allelic differences between isolates. Throwing away

information in this way results in lower power and greater uncertainty in parameter estimates, and less sensitive goodness-of-fit testing. The objective of population genetics techniques is to model nucleotide evolution in a statistical framework, obtain estimates for evolutionary parameters of interest, and refine the models using model criticism techniques.

1.3 Population genetics in epidemiology

Genetic diversity in pathogen species contains information about evolutionary and epidemiological processes, including the origins and history of disease, the nature of the selective forces acting on pathogen genes and the role of recombination in generating genetic novelty[1]. The role of population genetic analysis is to extract as much information from the nucleotide sequences as possible by using realistic evolutionary models. This section reviews recent applications of such methods to pathogenic organisms other than *N. meningitidis*, and compares the use of population genetic, or population-model based, approaches to evolutionary inference with phylogenetic, or population-model free, methodologies.

1.3.1 Pathogen biology

Like any other organism, a pathogen has an evolutionary history that is reflected in the distribution of genetic diversity within the species. What makes a pathogen special is that this evolutionary history is dominated by the successful and ongoing

[1] Section 1.3 was originally written as a review article: D. J. Wilson, D. Falush and G. McVean (2005) Germs, genomes and genealogies. *Trends in Ecology and Evolution* **20**: 39-45. All three authors contributed to writing the text.

colonisation of a host. Therefore, analyses of pathogen genomes can not only tell us things about the history of disease (when did the epidemic begin?), but also inform efforts to understand (which genetic changes made the ancestral organism pathogenic?) and control the disease (which is the best target for a vaccine, will vaccines be effective in different populations?).

The statistical and analytical tools available for comparing molecular sequences (DNA, RNA or protein) from representative pathogen isolates are becoming increasingly sophisticated. The first part of this section summarises recent research where molecular sequences alone have been used to understand pathogen biology. It will focus on: the reconstruction of a pathogen's origin and history; the nature of immune-mediated selection acting on pathogen genomes; and the role of recombination in generating genetic novelty. The second part will discuss the different methodologies that can be applied to molecular sequence data; in particular the use of phylogenetic methods versus population genetic ones. Phylogenetic methods were originally developed for the analysis of sequences from different species and make no assumptions about how population-level processes such as genetic drift, natural selection, changes in population size or geographical structure influence the shape of underlying gene trees. Population genetic approaches gain extra power to understand such factors by explicitly modelling their effects on tree-shape, and treating quantities of interest as explicit parameters for estimation. Integrating epidemiological models into a population genetics framework allows the estimation of epidemiologically relevant parameters.

1.3.2 The origin and history of pathogens

Tracing the origins and history of pathogen species provides information about what causes new epidemics and how they spread. Phylogenies constructed from samples of contemporary pathogen diversity reconstruct the history of those ancestors that have left descendants, the depth and shape of which can tell us about the size and structure of historical populations. For example, explosive growth generates characteristic 'star-like' phylogenies as seen in the HIV viruses and subtypes (Lemey *et al.* 2003; Robbins *et al.* 2003; Lemey *et al.* 2004). Historical changes in the pathogen population size may also be detected, e.g. the major increase in population size of the hepatitis C virus during the first half of the 20th century (Pybus *et al.* 2003). Dating events in phylogenies constructed from contemporary genetic diversity requires an independent estimate of the nucleotide (or amino acid) substitution rate. For many species such estimates are very approximate; e.g. estimates of the mutation rate in *Plasmodium* obtained by comparing *P. falciparum* (the most virulent human malaria parasite) and *P. reichenowi* genes (the most closely related Chimpanzee malaria parasite) differ by up to three-fold depending on the age postulated for the human-chimp split and which codon positions (2-fold or 4-fold degenerate positions) are used in the comparison (Rich *et al.* 1998). However, when isolates sampled from different time-points are available they provide internal calibration points (Drummond *et al.* 2002; Drummond *et al.* 2003b).

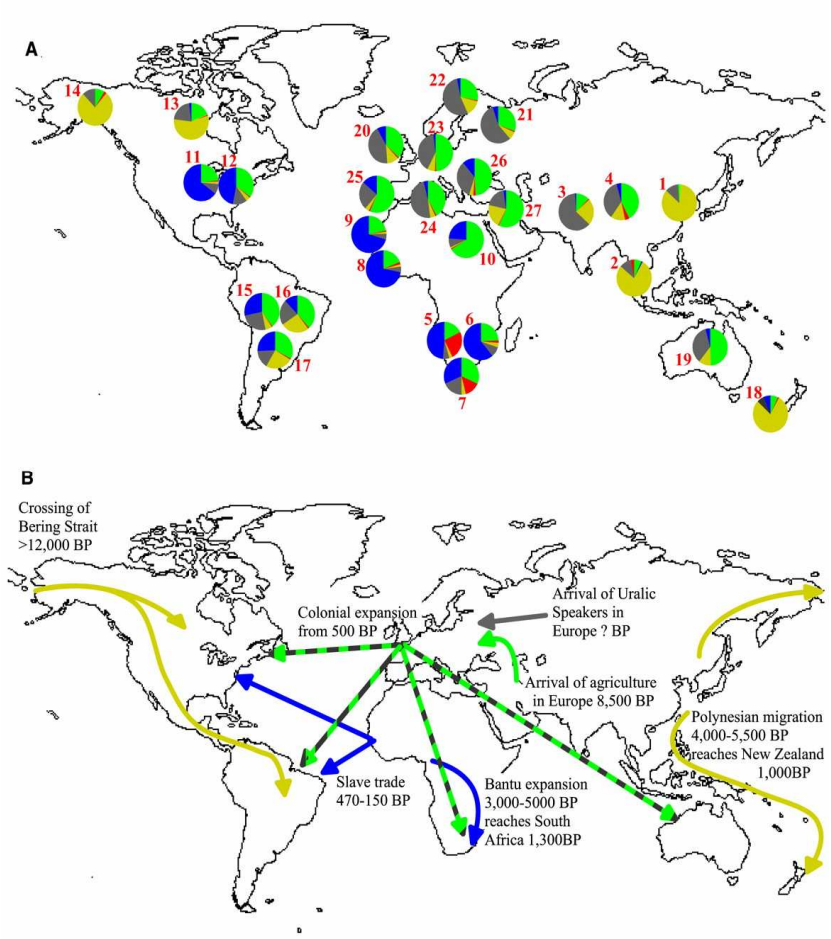


Figure 17 Putative and modern migration routes of *Helicobacter pylori*, as inferred by a population genetics clustering method (Falush *et al.* 2003a). **A** Ancestral sources of modern populations as a fraction of the genome. Five ancestral source populations were identified: two from Europe (green and grey), two from Africa (blue and red), and one from Asia (yellow). **B** proposed migration routes of those ancestral populations. Source: Falush *et al.* (2003b).

Other features of a pathogen's history may also be recovered. In highly recombining species, clustering algorithms (Falush *et al.* 2003a) allow the reconstruction of ancestral population structure and subsequent admixture, without subjective definition of population groups. Figure 17 shows the application of this technique to the enteric bacterium *Helicobacter pylori*. Reconstructing the ancestral populations revealed that the ancient migratory routes of *H. pylori* closely resemble that of their human hosts (Falush *et al.* 2003b). Where populations can be defined *a priori*, inferences can be made about the relative sizes, migration rates and dates of population separation. For example, analyses of natural populations of the Chestnut blight fungus, Cryphonectria hypovirus 1, have shown that transmission rates in the wild are much higher than those observed in lab experiments (Carbone *et al.* 2004).

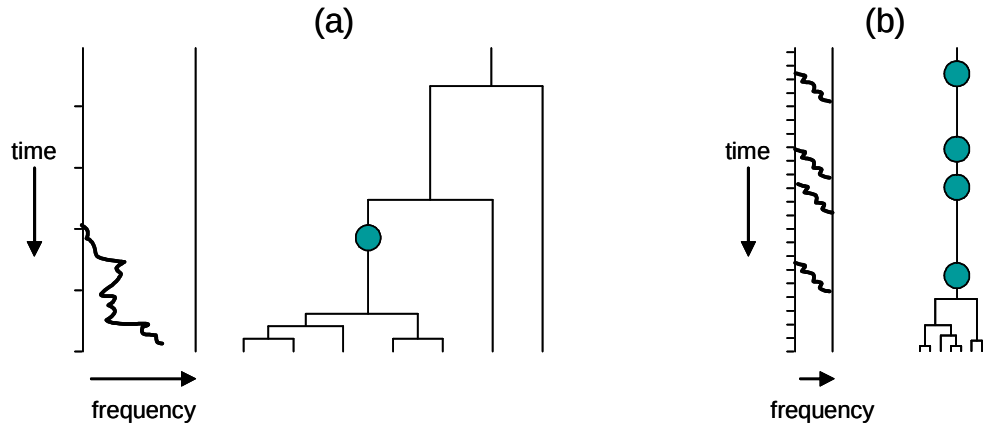


Figure 18 The ability to detect adaptive changes depends on the timescale of evolution. (a) an adaptive mutation occurred since the mrca of the sample, so the genealogy of the sample is distorted. The signature of selection will be visible in the frequency spectra of linked sites. (b) No adaptive mutation has occurred since the mrca, so the genealogy is unaffected. The signature of selection will be visible only by comparison to a closely related species, which would reveal an elevated rate of non-synonymous change.

1.3.3 Immune-mediated selection on pathogen genomes

For pathogen species, the selective pressures arising from the host immune system are a major influence on its evolution. Selection occurs both at the individual level, through the interaction of pathogen antigens and systems of innate and acquired immunity, and also at the population level, through the dynamics of herd immunity and cross immunity. How such factors influence patterns of genetic variation within pathogen populations depend on the relative timescales of host and pathogen adaptation. In species such as HIV-1 where rates of adaptation in the pathogen are high (Rambaut *et al.* 2004), immune-escape mutants will arise and be selected for within hosts. The effect of such selection is to transiently distort patterns of pathogen genetic variation within the host through the hitch-hiking effect (see Figure 18), a

pattern detected in longitudinal samples from HIV-1 infected patients (Shriner *et al.* 2004). However, immune-escape mutations do not generally provide an advantage to viruses infecting other hosts, who are unlikely to have encountered a virus with the same antigen type. Instead, diversifying selection within infected individuals results in pathogen species characterized by diverse and rapidly changing antigenic variation, the hallmark of which is an excess of protein-changing variation (relative to putatively neutral, non-protein changing variation) at antigenic genes during the course of the infection. Such a pattern is seen in HIV-1, particularly in the *env* gene, by use of codon-based phylogenetic methods (de Oliveira *et al.* 2004; Choisy *et al.* 2004).

Another route to detecting diversifying selection comes from comparison of within-species variation to between-species divergence. Because immune-escape mutants are unlikely to ever become fixed within a species, high levels of protein-changing variation at antigenic genes do not necessarily translate into high rates of change between species. For example, in a study of the gene encoding the erythrocyte-binding antigen EBA-175 in *P. falciparum* and the corresponding gene in *P. reichenowi*, there appears to be an excess of within-species variation relative to between-species divergence. The effect is not seen in the related gene *eba-140*, which suggests that *eba-175* is under within-host diversifying selection, probably as a result of interaction with the human immune system (Baum *et al.* 2003).

When cross-immunity is strong and rates of pathogen adaptation are slower, the pathogen population can theoretically become structured into different antigenic types Gupta *et al.* 1996; Haraguchi and Sasaki 1997; Lythgoe 2002; see section 1.2.3). Such types are maintained, or ‘balanced’, over time by frequency-dependent selection.

Structuring may be detected by comparing patterns of genetic variation to those expected under simple mathematical models of genetic variation, such as the neutral coalescent. In particular, balancing selection can result in genes with elevated levels of genetic diversity, changes in the distribution of allele frequencies and can inhibit drift by maintaining genetic variation within multiple populations despite geographic isolation. Such patterns are observed at *ama1* (Polley *et al.* 2003), a gene of *P. falciparum* which encodes an important antigen that represents a potential vaccine target.

Genome wide structuring of genetic variation (in the sense that the population is clustered into groups of closely-related individuals) is found in many pathogen populations. However, in addition to balancing selection, non-selective factors (e.g. genetic drift, bottlenecks, geographic and demographic stratification) and short-term selective processes (repeated partial selective sweeps associated with the origin of novel strains) can also generate similar patterns of linkage disequilibrium. Assigning the contributions of each of these factors to the observed disequilibrium represents a major challenge. Another potential explanation for stable maintenance of diverse types is antibody-dependent enhancement, where primary infection enhances rather than restricts the severity of subsequent infection by another strain (Ferguson *et al.* 1999), a process thought to be important for dengue virus.

1.3.4 The relevance of recombination

The tools available for inferring evolutionary history depend considerably on the biology of the pathogen. If recombination is rare, or hosts are only ever infected by a single pathogen strain, reconstruction of a single phylogeny is the natural starting

point for any analysis. In contrast, if recombination between different strains is common, different parts of the genome will have different phylogenetic histories, thus limiting the use of phylogenetic methods. In recombining species, instead of reconstructing a phylogenetic tree when a single tree may not exist, data sets can be described by summaries of the data such as the frequency distribution of polymorphisms, levels of linkage disequilibrium and measures of differentiation between populations (these summaries are also applicable to non-recombining species). Such summaries are the starting point for making inferences about the evolutionary history of the pathogen species, so knowing whether a species is recombining or not is critical in the choice of appropriate analyses.

Recombination also has major implications in studies that attempt to map phenotypically important genes by association, or through the hitch-hiking effect of adaptive mutations (Anderson 2004), because the rate of recombination determines the density of markers required to reliably detect causative mutations. Furthermore, estimates of important quantities, such as mutation rates, selection parameters (Anisimova *et al.* 2003; Shriner *et al.* 2003) or the age of a species' most recent common ancestor (mrca), are strongly biased if data from a recombining species are treated as having come from a clonal species (Schierup and Hein 2000).

The simplest way of detecting recombination from gene sequences is the identification of mosaic sequences, as in section 1.2.2.2. For example, in an alignment of sequences from avian influenza A, a highly pathogenic strain was shown to have a 30-nucleotide insert in the haemagglutinin gene relative to the low pathogenic strains, which is 100% identical to part of the neuraminidase gene (Suarez *et al.* 2004). More

sophisticated approaches to detecting mosaic structures have recently been developed, for example scanning methods that detect recombinant forms such as those observed in HIV-1 among characterized subtypes (Strimmer *et al.* 2003), and methods for weakly linked markers that detect admixture between subpopulations as in *Helicobacter pylori* (Falush *et al.* 2003a; Falush *et al.* 2003b).

Mosaic identification effectively assumes that all recombination events are very recent, and that genomes can be separated into 'pure' and 'mosaic'. In unstructured (panmictic) recombining species such a distinction is not valid, in which case an alternative is to try to identify the positions along the molecular sequence at which the phylogenetic tree changes. Many methodologies for detecting shifts in phylogeny have been developed, with recent work focusing on methods that aim to accommodate uncertainty about the tree reconstructions (Suchard *et al.* 2002; Husmeier and McGuire 2003). These methods work well at detecting a low number of recombination breakpoints along a sequence; for example in an alignment of the entire 3.2 kb genome of four strains of hepatitis B, two changes in topology were detected (Husmeier and McGuire 2003). Yet for many pathogens the rate of recombination is sufficient that changes in phylogeny are expected every few base pairs (Posada *et al.* 2002; Awadalla 2003).

For most species the rate of recombination relative to mutation is sufficiently high that there is little information about the underlying tree at any given position in the genome, and therefore little chance of exactly detecting recombination breakpoints. Under such circumstances the impact of recombination can be summarised either by a nonparametric estimate of the minimum number of recombination events in the

history of the gene samples, assuming no recurrent or back mutation (Myers and Griffiths 2003), or by a model-based estimate of the rate of recombination relative to genetic drift (Stumpf and McVean 2003). Coalescent methods can estimate recombination rates under models with recurrent and back mutations (Kuhner *et al.* 2000; McVean *et al.* 2002), and have demonstrated very high levels of recombination in various pathogens, including HIV-1 (McVean *et al.* 2002) and *P. falciparum* (Baum *et al.* 2003). Because genetic exchange can only occur between pathogen genomes in the same host, coalescent approaches measure the effective recombination rate, which can provide an indication of the rate of multiple infection (Bowden *et al.* 2004). Genomes with high intrinsic recombination rates, such as *P. falciparum* (Su *et al.* 1999) and HIV-1 (Zhuang *et al.* 2002; Levy *et al.* 2004) can therefore exhibit either high or low levels of historical recombination depending on the wider pathogen epidemiology (Anderson *et al.* 2000; McVean *et al.* 2002). Recombination has important biological, as well as methodological, consequences. Recombination (both homologous and non-homologous, or illegitimate) is an important source of genetic novelty, particularly at antigenic loci such as the haemagglutinin- and neuraminidase-encoding genes of influenza (Steinhauer and Skehel 2002; Li *et al.* 2004), where the origin of novel strains by recombination is known as antigenic shift.

1.3.5 Phylogenetic and population genetic approaches to inference

Diverse biological questions in disparate pathogen species naturally require a variety of approaches to analysing molecular sequence data. However, there is a broad distinction between those approaches which derive from the phylogenetic background and those that are rooted in population genetics modelling. The key distinction is that phylogenetic models make no assumptions about how population-level processes

(such as genetic drift, natural selection, inbreeding, restricted gene flow) influence the shape of genealogies (or gene trees) underlying samples of genetic material from within populations, while population-genetic approaches model such factors explicitly.

Phylogenetic approaches were first developed for the analysis of molecular sequences sampled from different species and have become widespread in the analysis of pathogen species diversity (Nielsen and Yang 1998; Nielsen and Huelsenbeck 2002; Lemey *et al.* 2003; Robbins *et al.* 2003; Yang *et al.* 2003; Grenfell *et al.* 2004; Leslie *et al.* 2004; Rambaut *et al.* 2004; Sheridan *et al.* 2004). In addition to estimating phylogenetic trees, such approaches can be used to date epidemics (Korber *et al.* 2000; Lemey *et al.* 2003; Robbins *et al.* 2003), detect recombination events (Husmeier and McGuire 2003) and identify sites of diversifying selection (Nielsen and Yang 1998; Suzuki and Gojobori 1999). However, because phylogenetic approaches were originally designed to analyse sequences from different species, they naturally assume that the shape of the tree itself is not informative about the quantities of interest. Post-hoc interpretation of tree-shape has, however, been important in the analysis of pathogen diversity; e.g. the observation of ladder-like trees for influenza has shaped theories of antigenic drift and shift (Fitch *et al.* 1997; Ferguson *et al.* 2003; Grenfell *et al.* 2004; Smith *et al.* 2004).

Population-genetic methods, in contrast, are based on mathematical models of populations; initially the ‘bean-bag’ genetics of Fisher, Wright and Haldane and more recently the coalescent theory of Kingman (1982a, 1982b) and Hudson (1983). Coalescent models describe in a probabilistic manner how population-level processes

influence the shape of genealogies underlying samples of gene sequences from within a population, and the resulting patterns of genetic variation. The standard neutral model (which underlies coalescent theory) assumes selective neutrality, constant population size and random mating, but can be extended to consider complexities such as population growth, inbreeding, geographical subdivision and different forms of natural selection (see Nordborg 2003 for a review).

The difference between phylogenetic and population-genetic approaches leads to conceptual differences in how data are analysed. Where phylogenetic approaches make statements about the tree and the substitutions mapped on to it, population-genetic approaches use the same genealogy to make statements about parameters of the coalescent model. For example, phylogenetic methods summarise variability among sequences by the branch lengths of the estimated tree, whereas population genetic methods estimate the population mutation rate $\theta/2$, which is the product of the per generation mutation rate and the effective population size of a species, N_e. Likewise, phylogenetic methods detect adaptive evolution by the relative rate of protein-changing and silent substitutions on the tree, whereas population-genetic methods estimate the selection coefficient of individual mutations from their effect on the shape of the genealogy (Przeworski 2003).

1.3.6 Advantages and disadvantages of population genetics

The benefit of fitting an explicit population-genetic model is that it gives extra power to detect phenomena of interest, and test specific hypotheses. For example, phylogenetic methods cannot test for population growth, because only in a model-based context do the star-like genealogies that population growth generates differ

from those expected without growth (without a model, all genealogy shapes are equally probable). Similarly, phylogenetic methods cannot detect single adaptive substitutions (Figure 18a) because distortion to allele frequencies caused by the hitch-hiking effect is only quantifiable by comparison to the standard neutral model (without a model all allele frequency distributions are equally probable). More generally, comparison of data to the expectations of the standard neutral model is a route to learning about which biological processes have been important in shaping genetic diversity. Many statistical methods for testing the (null) standard neutral model are available. These are either goodness of fit tests that aim to reject the null model (e.g. Tajima's [1989] D, Fu and Li's [1993] D*, Fay and Wu's [2000] H, the McDonald-Kreitman test [McDonald and Kreitman 1991] and the HKA test [Hudson *et al.* 1987]: discussed in Kreitman [2000] and Nielsen [2001]), or likelihood-based approaches that compare models with and without parameters of interest.

The problem of fitting a population-model to the data is that the biological simplifications required in order to make the model tractable may also render it meaningless. The coalescent process derives from a simplification of reproduction in natural populations. For pathogens, where successful reproduction requires both replication within hosts and transmission between hosts, population genetics must either incorporate epidemiological parameters explicitly in models of ancestry, or demonstrate that ignoring epidemiology still provides useful and meaningful inferences.

Both tasks are very much in their infancy. There is hope that the dynamics of simple epidemiological models, such as the susceptible-infectious-susceptible (SIS) model,

may give rise to genealogical models that are identical to those in well-characterized non-pathogen population-genetic models, such as metapopulations. That is the subject of the next section. However, where multiple strains with different epidemiological characteristics are considered, e.g. the epidemic-clone model for bacterial populations (Maynard Smith *et al.* 1993), it seems likely that novel population genetic models are required.

1.4 Coalescent models of *Neisseria meningitidis*

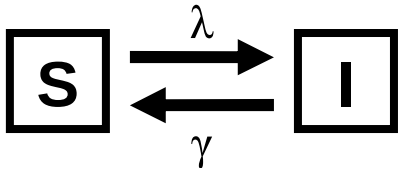
Undoubtedly the coalescent is a useful framework for evolutionary modelling, particularly for recombining organisms. However, its development has primarily been concerned with modelling populations of eukaryotic diploids, and it is not immediately obvious that the coalescent in its native form can be applied directly to obligate microparasites such as bacteria and viruses. Examples from section 1.2 show that it is imperative to specify the appropriate null model; failing to do so can render an analysis essentially meaningless. In this section I will argue that the coalescent is the appropriate null model. I will begin by briefly discussing the models commonly used in the epidemiology of microparasites then I will formally introduce the coalescent in a metapopulation. Finally I will discuss how the two can be combined, providing an integrated approach for modelling the evolution of microparasites.

1.4.1 Epidemiological models

Anderson and May (1991) review the staple differential equation models used for microparasites. These models are endlessly adaptable, so I will concentrate on the two most fundamental models that are in common usage. The SIS (susceptible-infectious-

susceptible) model is appropriate for microparasites that either (i) induce no immunity, or, (ii) cannot be cleared and remain infectious. The SIRS (susceptible-infectious-refractory-susceptible) model is appropriate for microparasites that do induce immunity, either temporary or life-long.

1.4.1.1 SIS



The host population is grouped into a proportion *I* that is infected and a proportion *S* that is susceptible. Susceptible individuals become infected at a rate λ, which is proportional to the prevalence of infectious individuals, offset by a transmission coefficient β. The magnitude of β reflects the transmissibility of the organism. Assuming that the per capita force of infection, λ, is proportional to the density of infectious individuals is known as strong homogenous mixing. The alternative assumption that λ is independent of *I* is known as weak homogenous mixing. Here I will assume strong homogenous mixing, so that $\lambda = \beta I$. Infected individuals clear the infection and return to the susceptible class at rate γ. $1/\gamma$ is the average duration of infection.

The changes in the proportion of infectious individuals over time, *t*, can be expressed as a differential equation.

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta IS - \gamma I \ .$$

Normally it is the equilibrium state of the model that is of interest, unless the emergence of a new infectious agent is being modelled (e.g. Pybus 2001). At equilibrium, the rate of change of *I* with respect to *t* is zero, so

$$I^* = 1 - \frac{\gamma}{\beta},$$

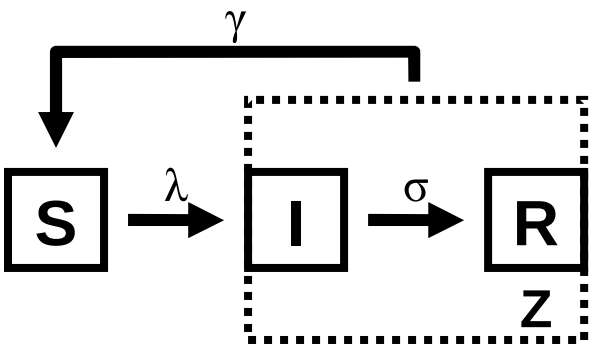
where an asterisk indicates the equilibrium frequency. The basic reproductive number R_0 is defined as the average number of secondary infections caused by a single primary infection in a totally susceptible population. This number is relevant because unless $R_0 \geq 1$ the infection will go extinct. A simple relationship is

$$1 - S^* = 1 - \frac{1}{R_0}, \qquad (1)$$

(Anderson and May 1991) implying that $R_0 = \beta/\gamma$. Therefore, for the infection to persist, $\beta > \gamma$. From these equations it is apparent that the dynamics of the model depend on the product of the transmission coefficient and the duration of infection.

The SIS model, as stated here, is equivalent to the SI (susceptible-infectious) model (in which the infection cannot be cleared), in which case $1/\gamma$ is the life expectancy of the host. Because the host population size is assumed to remain constant, the susceptible class is replenished with births at a rate equal to the mortality rate γ. A model including clearance of infection and births/deaths is straightforward, but is closely approximated by the SIS model when the life expectancy of the host is much greater than the average duration of infection.

1.4.1.2 SIRS



SIRS can be used to model disease that induces natural immunity, such as meningococcal disease. In addition to the susceptible and infectious class of the SIS model, a proportion *R* of the host population is refractory, and immune to reinfection. Infectiousness is lost at rate σ, and immunity is lost at rate γ. Class *Z* is the proportion of the host population infectious or immune. $1/\sigma$ is the average duration of infectivity. $1/\gamma$ is the average duration of immunity, or analogously, in an SIR (susceptible-infectious-refractory) model (where immunity is life-long) host life expectancy. A model containing host mortality and loss of infectiousness is very close to the SIRS model when host life expectancy greatly exceeds average duration of immunity.

The model can be represented by the differential equations

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta IS - \sigma I,$$
$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \beta IS - \gamma Z,$$

which can be solved to give $S^* = \sigma/\beta$, $I^* = \frac{\gamma(\beta-\sigma)}{\beta\sigma}$, $Z^* = 1-\sigma/\beta$ and, using Equation 1, $R_0 = \beta/\sigma$. For the infection to persist in the host population, $\beta > \sigma$.

The dynamics of this model depend principally on the product of the transmission coefficient and the duration of infectiousness, rather than the duration of immunity.

1.4.2 Metapopulations and the coalescent

1.4.2.1 The coalescent

The coalescent is a description of the ancestral history, or genealogy, of a random sample from a population that is evolving according to the standard neutral model (see section 1.2.4.1). In the standard neutral model the population has a constant size, and individuals reproduce with equal vigour. In its original formulation (Kingman 1982a, 1982b) the coalescent models the genealogy of n genes sampled from a non-recombining population of size N individuals, where it is assumed that N is large (formally, $N \to \infty$).

In the standard neutral model, also known as the Wright-Fisher model (Fisher 1930; Wright 1931) the reproductive success of members of the current generation, measured in number of offspring in the subsequent generation, follows a symmetric multinomial distribution. The Wright-Fisher model is a model of evolution forwards-in-time. The coalescent is a model of evolution backwards-in-time (see Nordborg 2003 for a review). Specifically, it is a model of the evolutionary history of genes backwards-in-time. Suppose the ploidy of the population is P. Whereas a diploid organism ($P = 2$) normally has two parents, a particular gene in that organism's genome has a single parent gene. Backwards-in-time, genes in the current generation choose their parent genes uniformly at random from the PN genes in the previous generation. From this, the waiting time until a pair of genes share an ancestor in

common can be found. The probability that a pair of genes have yet to find a common ancestor after *PNt* generations is

$$\left(1-\frac{1}{PN}\right)^{PNt},$$

which, as the population size gets very large ($N \to \infty$) equals approximately e^{-t}. So the waiting time, in units of *PN* generations, for the common ancestor of a pair of genes is exponentially distributed with rate 1. This is known as the rate of coalescence. For a sample of *n* genes there are $n(n-1)/2$ potential coalesce events, so the waiting time (in units of *PN* generations) to the first coalescence is exponentially distributed with rate $n(n-1)/2$. The chance of multiple simultaneous coalesce events is vanishingly small for large *N*.

1.4.2.2 The coalescent with recombination

Hudson (1983) described a way to simulate the genealogy of a sample of *n* genes in the presence of recombination. A genealogical tree with recombination as well as coalescence is no longer bifurcating, but can be a network, or graph. Griffiths and Marjoram (1997) provided a mathematical description of a coalescent genealogy with recombination, which they called the ancestral recombination graph (ARG). When there is recombination, the ancestral lineages not only merge together when they find a common ancestor, but also split apart, as a result of recombination. When there are *n* gene sequences, recombination events occur at rate $n\rho/2$ (per *PN* generations). So the waiting time for the next coalescence or recombination event (backwards in time) is exponentially distributed with rate $(\lambda_C + \lambda_R)$, where

$$\lambda_C = \binom{n}{2},$$

$$\lambda_R = \frac{n\rho}{2},$$

and the relative probability of coalescence is

$$\Pr(\text{coalescence}) = \frac{\lambda_C}{\lambda_C + \lambda_R}.$$

1.4.2.3 Coalescence in a metapopulation

A metapopulation model (Wright 1940; Levins 1968, 1969) is a simple extension of the standard neutral model in which the population is subdivided into subpopulations, or demes. Migration occurs between the demes, and sporadically demes go extinct. In a model in which there are a constant number of occupied demes *D*, unoccupied demes are recolonised at the same rate that occupied demes go extinct. Wakeley and Aliacar (2001) show that under certain conditions, the genealogy of a sample of genes taken from a metapopulation is a straightforward extension of the coalescent.

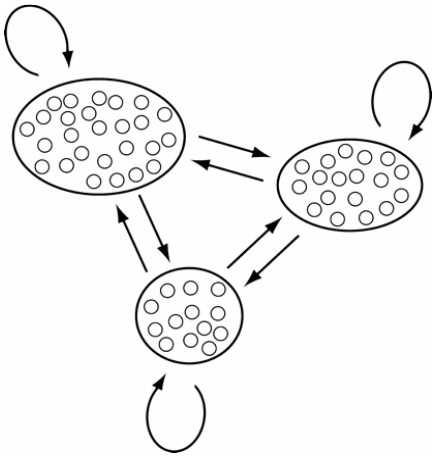


Figure 19 An example of a metapopulation model with many demes. There are $K = 3$ types of deme, which may differ in their population size, extinction/recolonisation rates and migration rates. Individuals can move between demes by migration or recolonisation events, indicated by the arrows. Source: Wakeley and Aliacar (2001).

Fundamental to the model of Wakeley and Aliacar (2001) is that there are a large number of (occupied) demes D, so that the sample size is much smaller than the number of demes (formally, $D \to \infty$). In addition, there can be K different types of deme that can differ in their population size, rates of extinction/recolonisation and rates of migration. Demes of type i have population size N_i, extinction/recolonisation rate E_i per PN_i generations, and migration rate M_i per PN_i generations. Note that this is the backwards migration rate, which means that M_i is the rate at which individuals migrate into deme i from other demes. When a deme of type i is recolonised, it has k_i founders, and the deme population is instantaneously repopulated to N_i individuals. A proportion β_i of all demes are of type i. Figure 19 illustrates the metapopulation model.

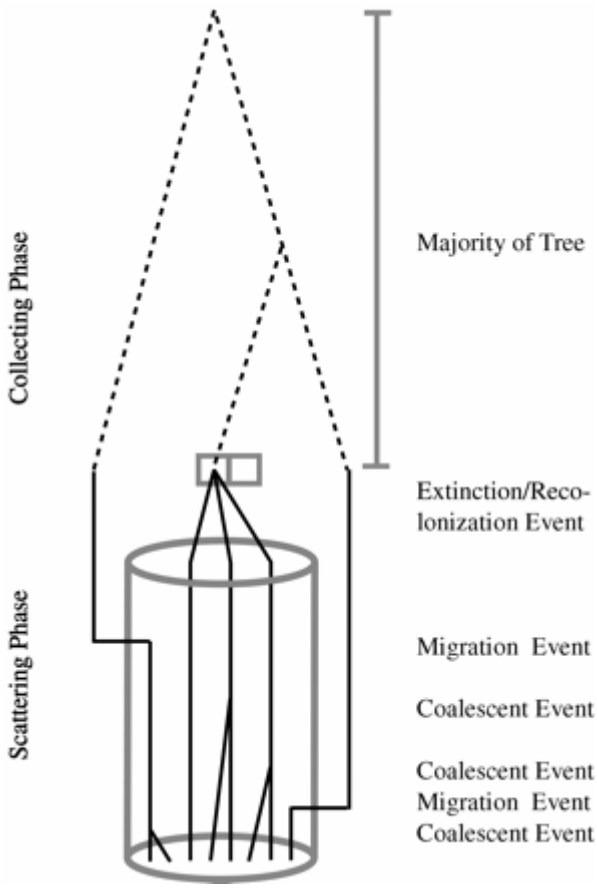


Figure 20 The genealogy of a metapopulation is divided into the scattering phase and the collecting phase. In this example, 8 genes were sampled from a single deme. In the scattering phase a sequence of coalescence, migration and recolonisation events rapidly change the configuration of the ancestral lineages amongst the demes. At the end of the scattering phase there are only 3 lineages left, each in a separate deme. During the collecting phase these coalesce according to a standard coalescent with an altered timescale. Source: Wakeley and Aliacar (2001).

As a consequence of the large number of demes, the genealogy of a sample from the model described above is straightforward. Suppose the sample, of size *n*, was taken from *d* demes so that $\mathbf{n} = (n_1, \ldots, n_d)$ describes the sample configuration, with demes labelled 1…*d* and $n = \sum_{i=1}^{d} n_i$. Wakeley and Aliacar (2001) show that the genealogy of this sample consists of two parts, that they call the scattering phase and the collecting phase (Figure 20). In the scattering phase the ancestral lineages rapidly

coalesce, migrate or undergo recolonisation until there is a single lineage in each deme. Backwards-in-time, recolonisation is equivalent to coalescence if $k_i = 1$, or to a combination of coalescence and migration if $k_i > 1$. The scattering phase for deme i takes around PN_i generations or less. The collecting phase describes the rest of the genealogical history, which resembles a standard coalescent genealogy but with a different timescale. That is to say that the collecting phase is a standard coalescent process with effective population size

$$N_e = \frac{ND}{2(M+E)F}, \tag{2a}$$

where

$$F = \frac{1+E/k}{1+2M+E}, \tag{2b}$$

in the case of a single deme type ($K = 1$, subscripts for k, N, M and E suppressed) (Wakeley and Aliacar 2001; Wakeley 2004). F has a natural interpretation in the coalescent metapopulation model. It is the inbreeding coefficient, which is to say that it is the probability that the ancestral lineages of a pair of sequences sampled from the same deme coalesce during the scattering phase (Wakeley and Aliacar 2001).

This separation of timescales relies on the assumption that D is much larger than N. When migration or recolonisation occurs, and the ancestral lineage of the migrant or coloniser moves to another deme (the source deme), the probability that the source deme is also occupied by another ancestral lineage is on the order of magnitude of $1/D$. Thus certain types of events occur with vastly different rates.

- **Fast timescale.** Coalescence within demes and migration or recolonisation in which the source deme is unoccupied occur with rates on the order of PN generations.

- **Slow timescale.** Migration or recolonisation in which the source deme is occupied occur with rates on the order of PN/D generations, which is very much slower for large D.

There are several important consequences of the separation of timescales. The scattering phase is so short relative to the collecting phase that if the mutation rate is finite in the collecting phase then no mutation events occur during the scattering phase. Recombination is easily incorporated into the model (Wakeley and Aliacar 2001; Lessard and Wakeley 2004), but if the recombination rate is finite in the collecting phase then no recombination events occur during the scattering phase. When a recombination event occurs during the collecting phase, there are transiently two ancestral lineages in one of the demes, analogous to during the scattering phase. The lineages rapidly either coalesce back together again, or move to another deme owing to migration/recolonisation. In the case of coalescence (or recolonisation when $k = 1$), which occurs with appreciable probability, the recombination event has no effect on the genealogical history of the genes. As a result, the observed recombination rate ρ_{obs} is lower than would be expected for a standard coalescent process with the specified effective population size, resulting in higher than expected LD:

$$\rho_{obs} = \rho(1 - F), \tag{3}$$

when there is a single deme type (note that there is a typographical error in Equation 28 of Wakeley and Aliacar 2001; John Wakeley personal communication).

1.4.3 Epidemiology and the coalescent

The model of coalescence in a metapopulation is useful because it could easily describe a population of hosts, each of which is infected with a population of microparasites.

- Each host is represented by a deme
- The population size of a deme is the parasite load
- Primary infection corresponds to recolonisation of a deme
- Secondary infection corresponds to migration between demes
- Clearance of infection corresponds to a deme extinction

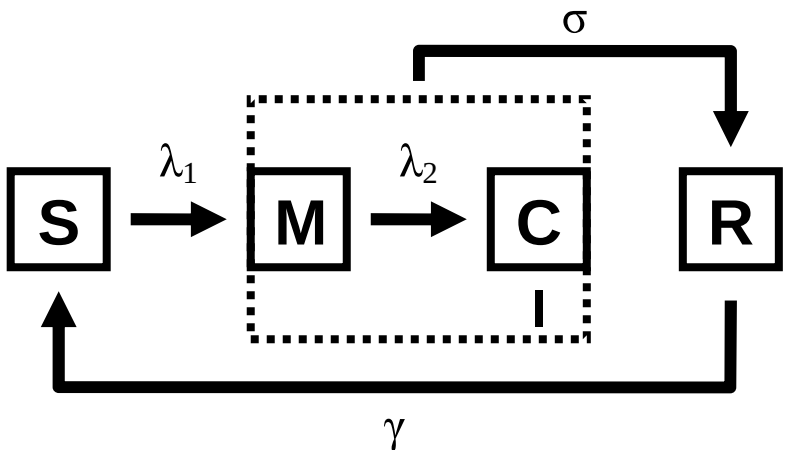
Of the epidemiological models discussed above, SIRS is appropriate for modelling *N. meningitidis* because it includes natural immunity. Although a simplification of the truth, I will show how to incorporate it into the simplest case ($K = 1$) of Wakeley and Aliacar's (2001) metapopulation model. Doing so provides some valuable insights into modelling microparasites using the coalescent. The versatility of the coalescent metapopulation model means that incorporation of more complex epidemiological models, for example modelling age structure, would be straightforward (see for example Laporte and Charlesworth 2002).

1.4.3.1 SIRS with superinfection

Before the SIRS model can be integrated with the metapopulation model it is helpful to expand it slightly so as to distinguish between primary and secondary infection. In the metapopulation analogy, primary infection of a previously susceptible host corresponds to recolonisation of an unoccupied deme. Secondary infection, on the

other hand, corresponds to migration between occupied demes. Separating primary and secondary infection in this way is important, both epidemiologically because secondary infection may differ in its success rate, and evolutionarily because only recombination within multiply infected hosts leads to the emergence of allelic novelty and mosaic genomes.

Suppose that *S* is the proportion of hosts that are susceptible, and *R* is the proportion of hosts that are refractory (immune), as before. The proportion of hosts that are singly infected is *M*, whilst the proportion of hosts that are co- or super-infected is *C*. $I = M + C$. Each generation (which may be thought of as the average time it takes for the complete intra-host population of microparasites to turn over), the probability that a susceptible host becomes infected is given by the per-capita force of infection Λ_1, where $\Lambda_1 = B_1 I$ and B_1 is the transmission coefficient for primary infection $(0 < B_1 < 1)$. When a host is infected for the first time, the intra-host parasite population is assumed to immediately attain its carrying capacity N_P. It is assumed that primary infection results from a single founding genotype. The probability that an infected host (be it singly or multiply infected) is reinfected is given by the per-capita force of secondary infection Λ_2, where $\Lambda_2 = B_2 I$, and B_2 is the transmission coefficient for secondary infection $(0 < B_2 < 1)$. When a host is reinfected, it is assumed that a single parasite genotype enters the intra-host population at initial frequency $1/N_P$. Infected hosts (be it single or multiply infected) become refractory with probability Σ per generation $(0 < \Sigma < 1)$. Refractory hosts lose immunity and become susceptible once more with probability Γ per generation $(0 < \Gamma < 1)$.



It is assumed that N_P is large, and the epidemiological parameters B_1, B_2, Σ and Γ are small so that in the limit as $N_P \to \infty$,

$$\beta_1 = \lim_{N_P \to \infty} PN_P B_1,$$

$$\beta_2 = \lim_{N_P \to \infty} PN_P B_2,$$

$$\gamma = \lim_{N_P \to \infty} PN_P \Gamma,$$

and

$$\sigma = \lim_{N_P \to \infty} PN_P \Sigma$$

are finite, where P is the ploidy of the parasite. The host population size N_H is assumed to be sufficiently large that the rate of change of the proportion of susceptible, singly infected, multiply infected and refractory individuals can be described deterministically by the differential equations

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \gamma R - \beta_1 IS, \\
\frac{\mathrm{d}M}{\mathrm{d}t} &= \beta_1 IS - \beta_2 IM - \sigma M, \\
\frac{\mathrm{d}C}{\mathrm{d}t} &= \beta_2 IM - \sigma C, \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \sigma I - \gamma R,
\end{aligned}$$

where time t is measured in units of PN_P generations. In units of PN_P generations, the per capita forces of primary and secondary infection are $\lambda_1 = \beta_1 I$ and $\lambda_2 = \beta_2 I$ respectively.

1.4.3.2 Metapopulation with SIRS

Whereas in a standard metapopulation model the number of demes is usually assumed to be independent and fixed, in the SIRS metapopulation model the number of demes is dynamic, and dependent upon the epidemiological parameters. To integrate the SIRS model and the metapopulation model, I will assume that infection rates are at equilibrium in the host population. It is possible to use the SIRS model to model the emergence of the microparasite in the metapopulation. The number of infected hosts, which corresponds to the number of demes, can be found by solving the differential equations under equilibrium conditions. At equilibrium, a proportion $S^* = \sigma / \beta_1$ of hosts are susceptible, so $R_0 = \beta_1 / \sigma$ from Equation 1. For the microparasite to persist in the host population, R_0 must be greater than one, so $\beta_1 > \sigma$. The equilibrium frequency of infected hosts is

$$I^* = \frac{\gamma(\beta_1 - \sigma)}{\beta_1(\sigma + \gamma)} = \frac{\gamma}{\sigma + \gamma}\left(1 - \frac{1}{R_0}\right), \qquad (4)$$

which means that for a host population of size N_H, there will be $I^* N_H$ infected hosts. This is analogous to $D = I^* N_H$ occupied demes in the metapopulation. The relative frequency of multiple to single infection is given by

$$\frac{C^*}{I^*} = \frac{\beta_2 \gamma(\beta_1 - \sigma)}{\beta_2 \gamma(\beta_1 - \sigma) + \beta_1 \sigma(\sigma + \gamma)}.$$

This tends to zero for small β_2, and tends to 1 for large β_2.

In the SIRS metapopulation model, the duration of infectiousness is $1/\sigma$, regardless of whether hosts are singly or multiply infected. At equilibrium, $\sigma = \beta_1 S^*$ because

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta_1 I^* S^* - \sigma I^* = 0 .$$

So the rate at which demes (hosts) are recolonised (suffer primary infection) and go extinct (clear infection) occurs at rate $E = \beta_1 S^*$ per PN_P generations. Similarly, the rate at which demes (hosts) experience immigration (secondary infection) occurs at rate $M = \beta_2 I^*$. Therefore using Equation 2, the effective population size of the collecting phase for the SIRS metapopulation model is

$$N_e = \frac{N_P I^* N_H}{2\left(\beta_2 I^* + \beta_1 S^*\right)F}, \tag{5a}$$

where

$$F = \frac{1+\beta_1 S^*}{1+2\beta_2 I^* + \beta_1 S^*} . \tag{5b}$$

It is interesting to remark that, whereas the genealogy of the SIRS metapopulation is straightforward (a coalescent process with an altered timescale, with a correction for the sample configuration), the effective population size for the genealogy is a complex function of the epidemiological parameters, with little hope to disentangle them. However, the inbreeding coefficient itself, *F*, which might be thought of more as a population genetic parameter than an epidemiological parameter, could be estimated from the data. Supposing mutations occur at rate $\theta/2$ per site per PN_e generations according to the infinite sites model (Watterson 1975), then for a pair of sequences of length *L* sampled from different demes, the expected number of pairwise differences is

$$E(\pi_T) = \theta L .$$

For a pair of sequences sampled at random from the population, the expectation is the same because for a large number of demes, a truly random sample has zero probability of sampling the same deme twice. For a pair of sequences sampled from the same deme, the expected number of pairwise differences is

$$\begin{aligned} E(\pi_i) &= (1-F)\times E(\pi_T) \\ &= \theta L(1-F), \end{aligned}$$

because for θ to be finite on the timescale of the collecting phase (N_e) it must be zero on the timescale of the scattering phase. As a result, the only source of variation within a deme must be multiple infection. This is an important implication of the model. A moment estimator of the inbreeding coefficient would be

$$\hat{F} = \frac{\overline{\pi}_T - \overline{\pi}_i}{\overline{\pi}_T},$$

where $\overline{\pi}_i$ and $\overline{\pi}_T$ are the observed average number of pairwise differences within and between demes respectively.

There are many simplifications in a SIRS model; however it is useful to see how such an epidemiological model can be integrated into a population genetics framework, and how the key parameters of the two models relate to one another. Patterns of genetic diversity in microparasite populations can potentially reveal a great deal about the evolutionary history of the population, so it is important to appreciate the relationship between, for example, prevalence and effective population size. The SIRS metapopulation model introduced here results in a straightforward genealogical model, but the relationship between prevalence and effective population size is not linear, which suggests that some thought is needed before inferring changes in parasite prevalence over time directly from genetic data. There are other important

insights from the model, such as the relationship between the observable rate of recombination in a sample of sequences and the rate of recombination within a host. This insight might help reconcile molecular genetic and population genetic estimates of the recombination rate in microparasites. The relationship between observable and actual rates of recombination is investigated further in section 2.2.2. That within-host variation can only be explained by multiple infection in the SIRS metapopulation model is another important insight. Obviously such a result depends on the assumptions of the model, and if the data appear to contradict this prediction, that says something interesting about the validity of the model. Finally, it is significant that the simple SIRS model, the appropriate null model in an epidemiological setting, gives rise to a simple coalescent model, suggesting that the coalescent is the appropriate null model for the population genetics of microparasites. What is more is that the versatility of Wakeley and Aliacar's (2001) model of coalescence in a metapopulation means that more complex epidemiological models can be integrated into a population genetics framework.