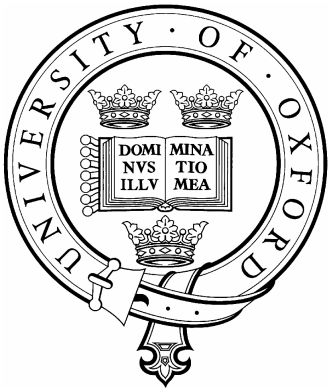
University of Oxford



Multilocus sequence analysis of the pathogen *Neisseria meningitidis*

Daniel J. Wilson

St. John's College

Mathematical Genetics and Bioinformatics

Department of Statistics

A thesis submitted for the degree of Doctor of Philosophy

September 2005

Multilocus sequence analysis of the pathogen *Neisseria meningitidis*

Daniel J. Wilson, St. John's College
D.Phil. thesis, Trinity Term 2005

ABSTRACT

Neisseria meningitidis is the bacterium responsible for meningococcal meningitis and septicaemia in humans. Meningococcal disease is primarily a disease of young children, characterized by rapid deterioration from first symptoms to death, with an 11% fatality rate and a global distribution. Patterns of genetic diversity in meningococcal populations provide an account of their evolutionary history and structure, which can be inferred by population genetics modelling. Understanding these phenomena can inform control and prevention strategies, and provides interesting case studies in evolution. The aim of this thesis is to develop population genetics techniques for inferring the evolutionary history of meningococci.

I begin by reviewing the field, and justifying the use of coalescent methods in modelling microparasite populations. Inference on carriage populations of meningococci under the standard neutral model and the neutral microepidemic model is performed using a modification to approximate Bayesian computation. AMOVA and Mantel tests are used to quantify the differentiation between carriage and disease populations, and the extent to which geography and host age structure carriage populations. The results are used to propose revised coalescent models for meningococcal evolution.

The role of natural selection in shaping meningococcal diversity is investigated 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, I apply it to the *porB* antigen locus and seven housekeeping loci in *N. meningitidis*. There is strong evidence for selection imposed by the host immune system in the antigen locus, but not the housekeeping loci which are functionally constrained. Finally I discuss the future direction of population genetic approaches to understanding infectious disease.

Acknowledgements

Thanks go to members of the Mathematical Genetics and Bioinformatics Group of the Department of Statistics and the Bacterial Population Structure and Public Health Group of the Department of Zoology. In particular I would like to thank Adam Auton, Ella Chase, Daniel Falush, Bob Griffiths, Rosalind Harding, Chris Holmes, Keith Jolley, Stephen Leslie, Jonathan Marchini, Noel McCarthy, Chris Spencer and Rachel Urwin. I would also like to thank Graham Coop and Don Conrad of the Human Genetics Department at the University of Chicago. Ziheng Yang kindly provided C computer code and Jeremy Derrick the molecular structure of the PorB molecule. Special thanks to Gillian Kay and my parents Brian and Janet, who in addition to support also kindly helped to proof-read the thesis. Finally I would like to thank my co-supervisors, Gil McVean and Martin Maiden.

This thesis was supported by a research studentship from the Biotechnology and Biological Sciences Research Council. Much of the computational work presented here was conducted on a multi-node AMD compute cluster that was bought with a grant awarded by the Wolfson Foundation to Peter Donnelly. Part of the work in this thesis was presented at the Society for Molecular Biology and Evolution, Newport Beach, California, June 2003, the London Mathematical Society Symposium on Mathematical Genetics, Durham, July 2004, the 14th International Pathogenic Neisseria Conference, Milwaukee, Wisconsin, September 2004, and the 10th Congress of the European Society for Evolutionary Biology, Krakow, Poland, August 2005. I would like to thank the BBSRC, the London Mathematical Society and St. John's College, Oxford for financial support in attending these conferences.

Thanks also to my examiners, Brian Charlesworth and Jotun Hein.

Table of Contents

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

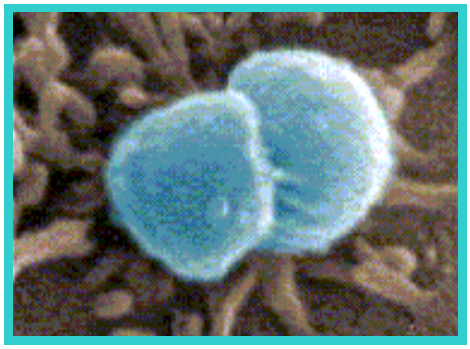


Figure 1 Neisseria meningitidis is a diplococcus ordinarily resident in the nasopharynx. Source: scanning electronic micrograph, Sanger Centre.

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

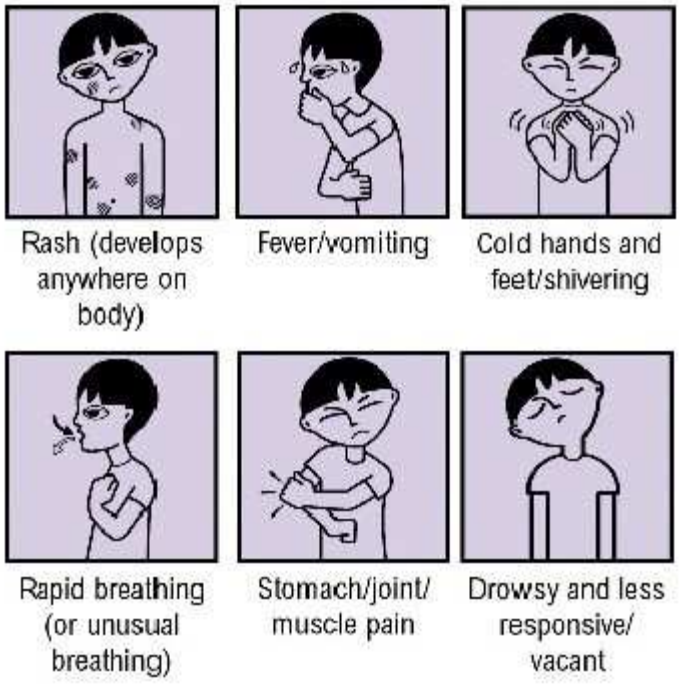


Figure 3 Symptoms of meningococcal septicaemia. Source: Meningitis Research Foundation (2005).

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.

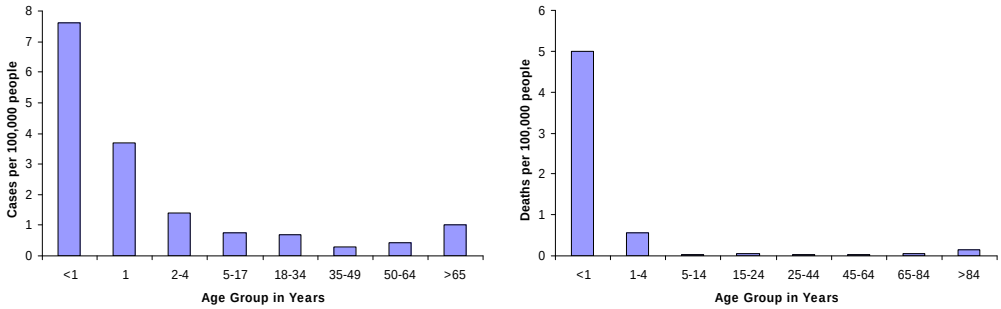


Figure 4 Left: rate of incidence of meningococcal disease with age in the United States, 2000. Right: fatality rate of meningococcal disease with age in the United Kingdom, 2002. Sources: Centers for Disease Control and Prevention (2000), National Office of Statistics (2002).

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

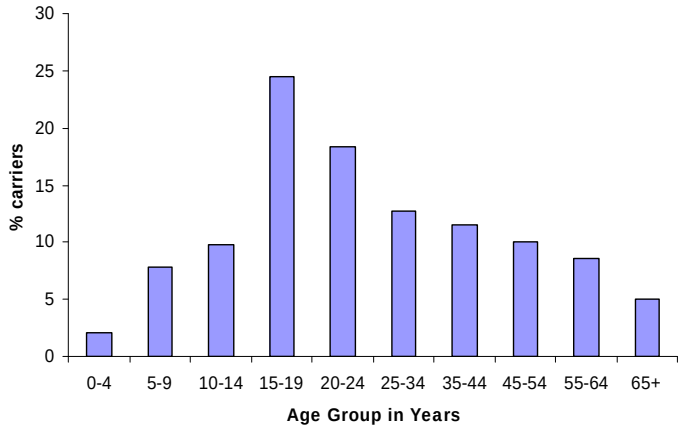


Figure 5 Percentage carriage with age in Gloucestershire, United Kingdom, 1986. Source: Cartwright *et al.* (1987).

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).

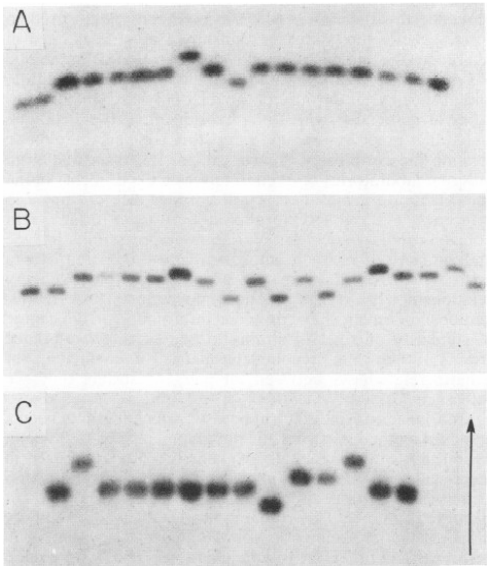


Figure 6 Gel illustrating electrophoresis of three enzymes of *Escherichia coli*. The arrow indicates the direction of migration of the enzymes across the gel. Source: Selander *et al.* (1986).

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.

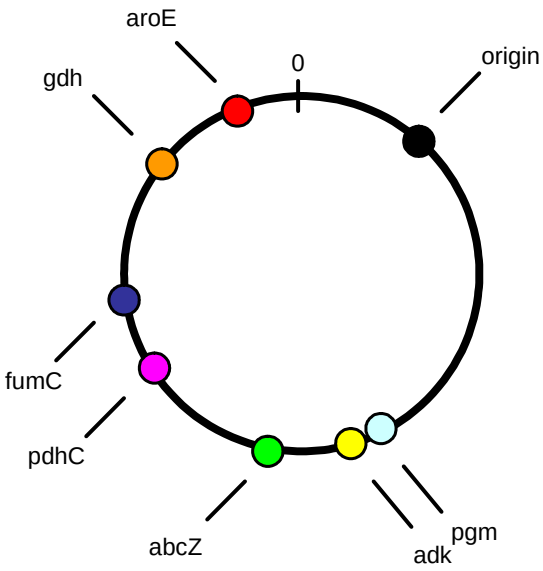


Figure 7 Locations of the seven housekeeping loci in the *N. meningitidis* Z2491 genome (Parkhill *et al.* 2000) that are used for multilocus sequence typing. The origin of replication is marked (origin) and the reference point for nucleotide positions (0).

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.

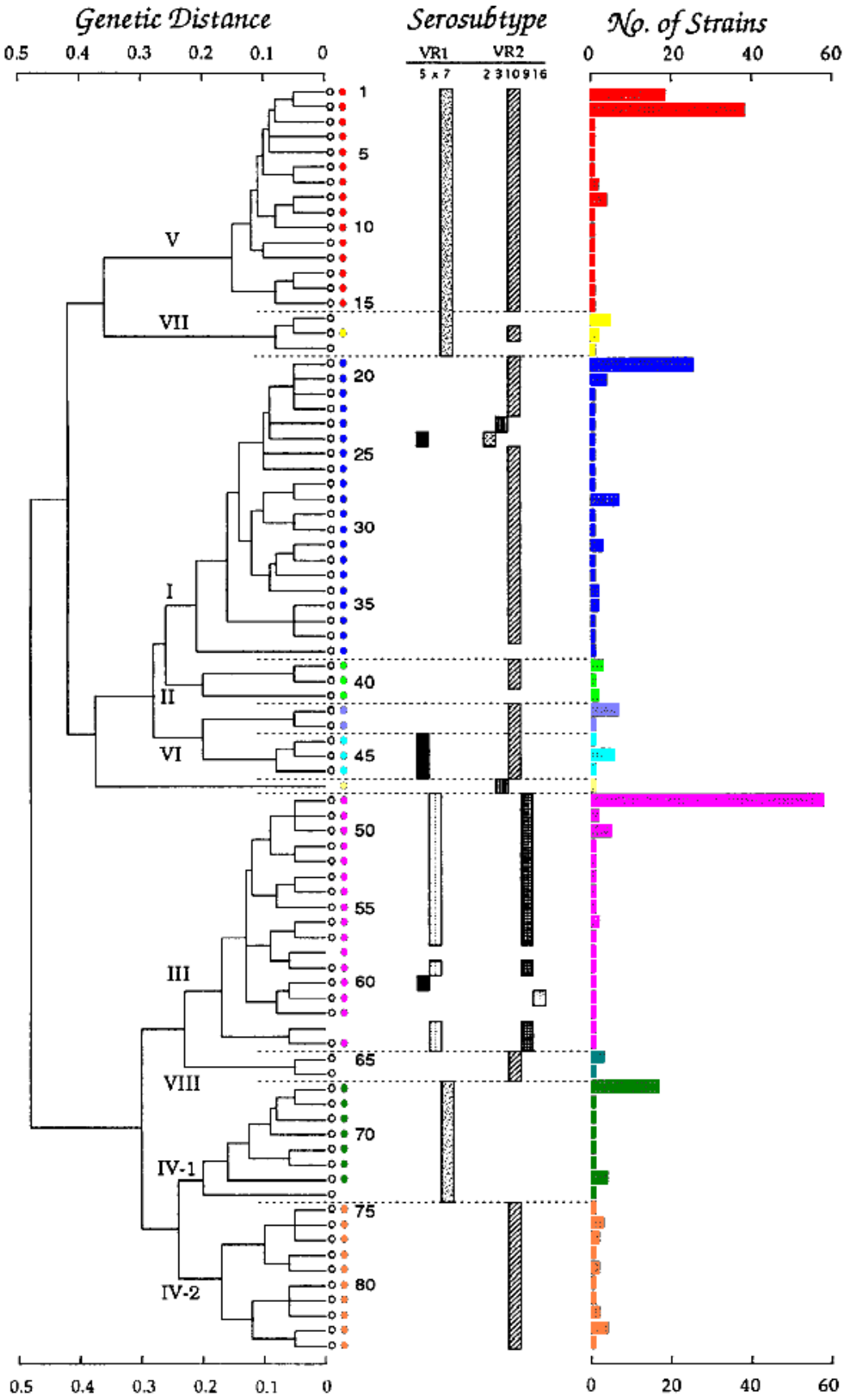


Figure 8 Genetic relationships, serosubtype patterns and relative abundance of 84 ETs in a global sample of disease-causing serogroup A meningococci. Adapted from: Wang et al. (1992).

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.

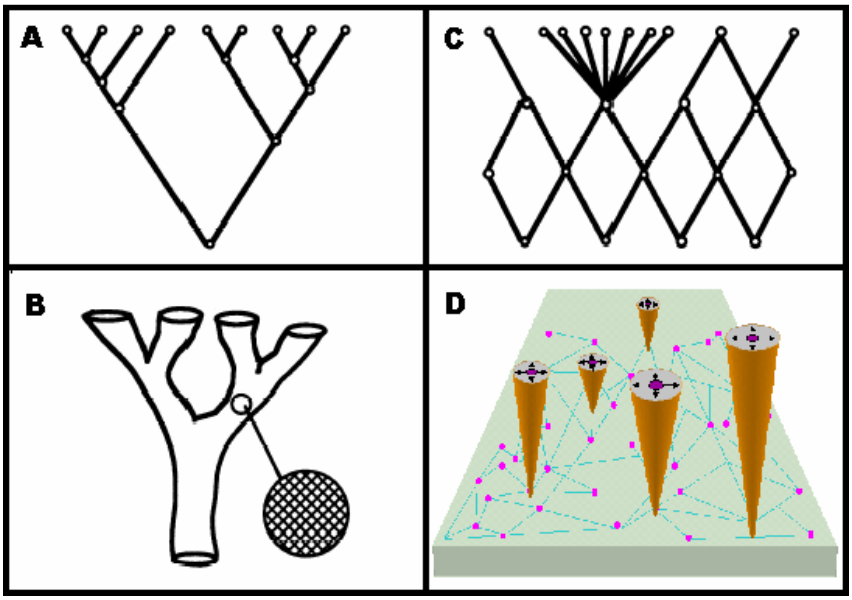


Figure 9 Representation of population structures. **A** Strict clonality is represented by a bifurcating evolutionary tree. **B** Frequent recombination within reproductively isolated subpopulations is represented by a network within a bifurcating population tree. **C** Epidemic clone model in which 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

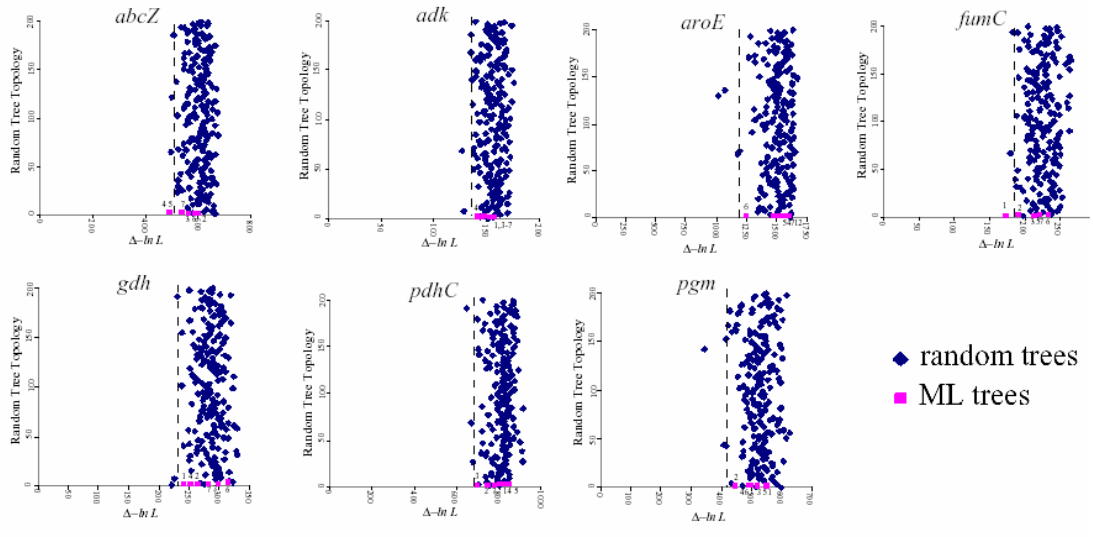


Figure 10 Phylogenetic incongruence amongst MLST loci for 30 isolates representative of global disease. Horizontal axis is the difference in log likelihood between the ML tree for each locus and the ML tree for another locus (pink squares) or a random tree (blue diamonds). Trees are spread vertically 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

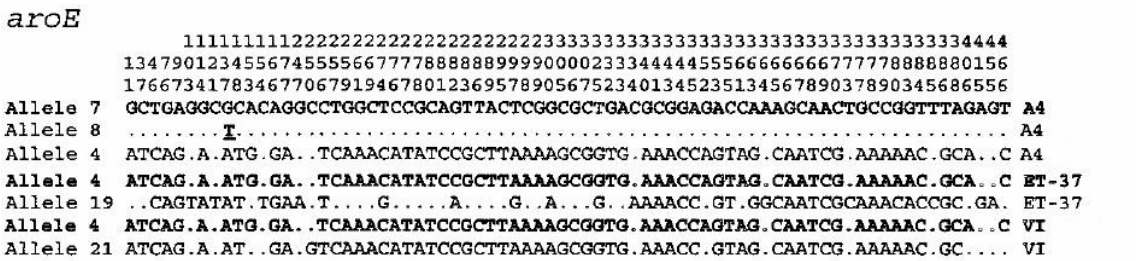


Figure 11 Polymorphic sites in *aroE* within three clonal complexes. The clonal complexes are labelled at the far right. For each clonal complex there is one bold row representing the central, or most common, genotype. The non-bold rows represent all other variants assigned to that clonal complex. Dots 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.

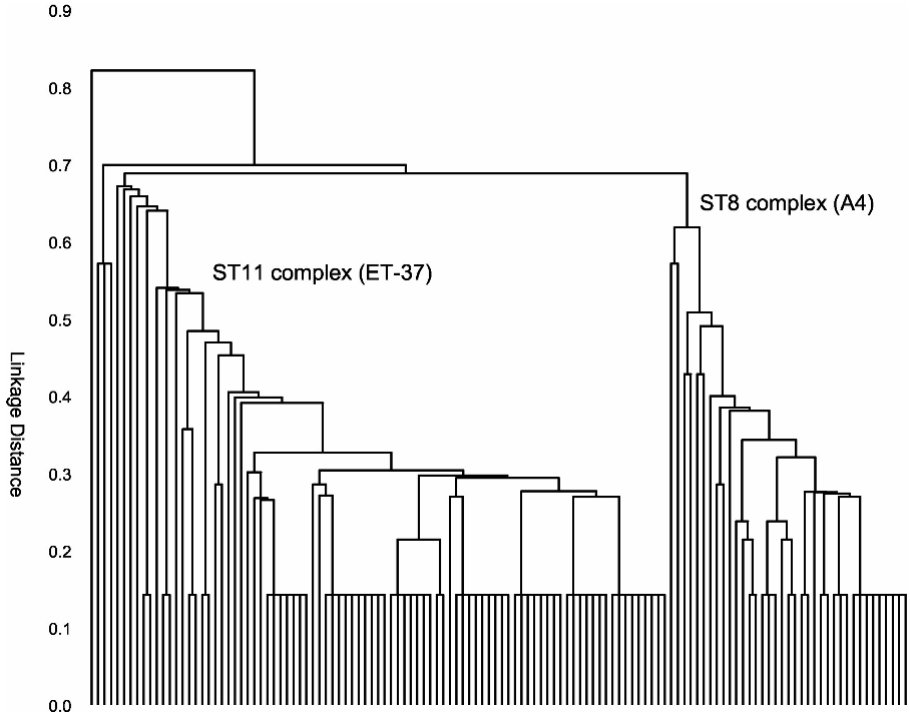


Figure 12 UPGMA dendrogram of the ST8 (A4 cluster) and ST11 (ET-37) clonal complexes. All STs with fewer than 4 alleles different to ST 8 or ST11 were included from the *Neisseria* MLST database. Source: Feil *et al.* (2004).

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

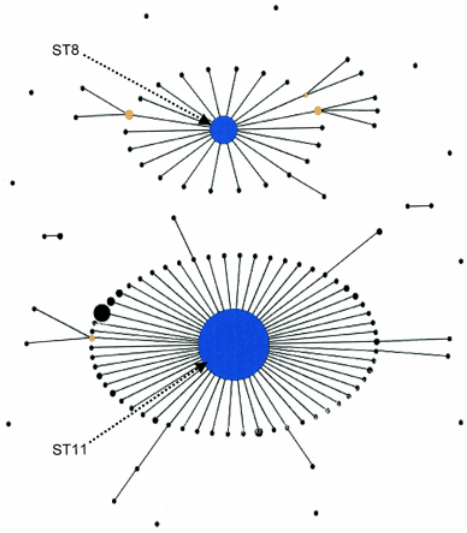


Figure 13 eBURST diagram of the group containing the ST8 and ST11 clonal complexes. A group was defined as STs differing by less than 3 alleles from one another. Clonal complexes (connected nodes in the diagram) were defined as STs differing by less than 2 alleles from one another. Source: Feil *et al.*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

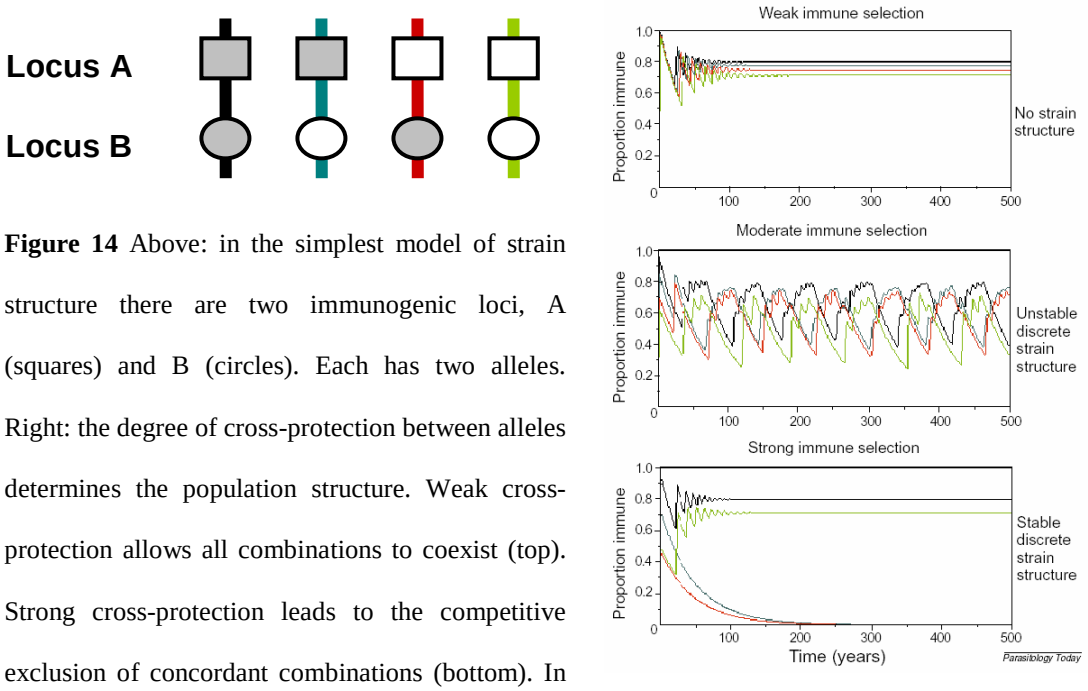


Figure 14 Above: in the simplest model of strain structure there are two immunogenic loci, A (squares) and B (circles). Each has two alleles. Right: the degree of cross-protection between alleles determines the population structure. Weak cross-protection allows all combinations to coexist (top). Strong 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.

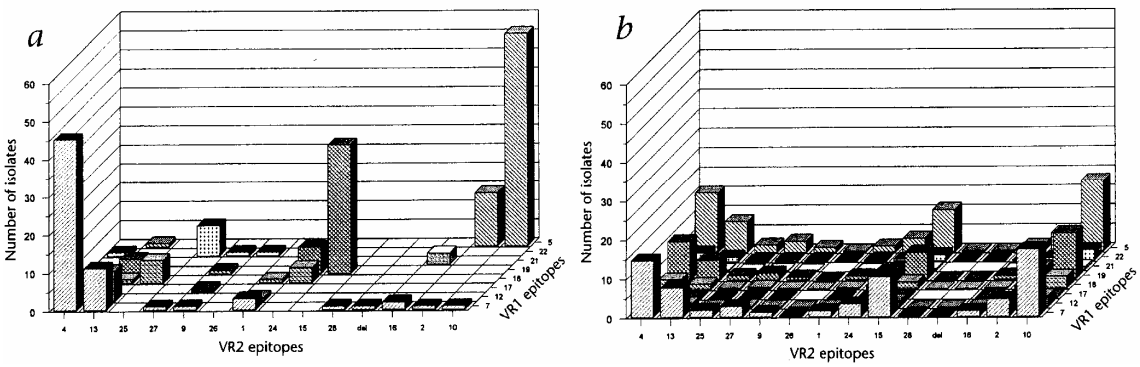


Figure 15 Association between epitopes of the VR1 and VR2 region of the PorA outer membrane protein. (**a**) Observed frequency distribution. (**b**) Expected frequency distribution under linkage equilibrium. Source: Gupta *et al.* (1996).

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

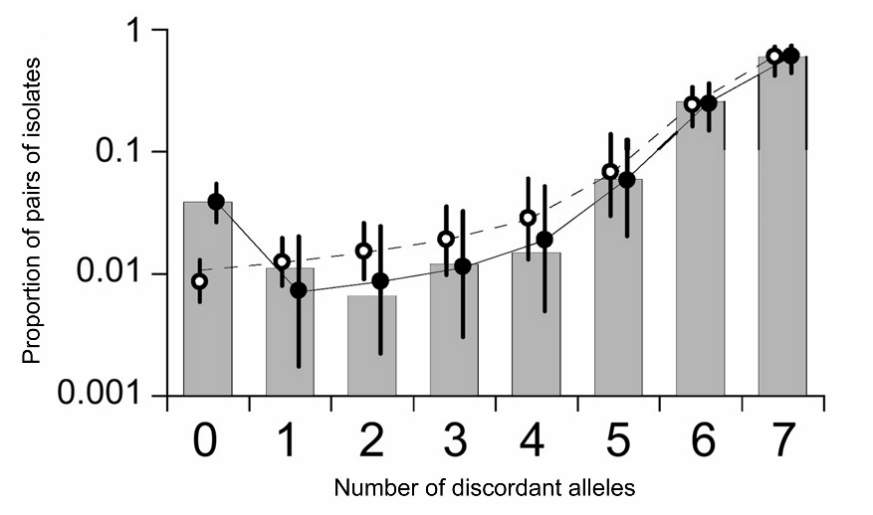


Figure 16 Allelic mismatch distribution for Czech carriage study (grey bars). The horizontal axis shows the number of loci at which a pair of isolates can differ (up to 7 for MLST), and the vertical axis the proportion of pairs that differ at that number of loci. Open circles show the fit under the standard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

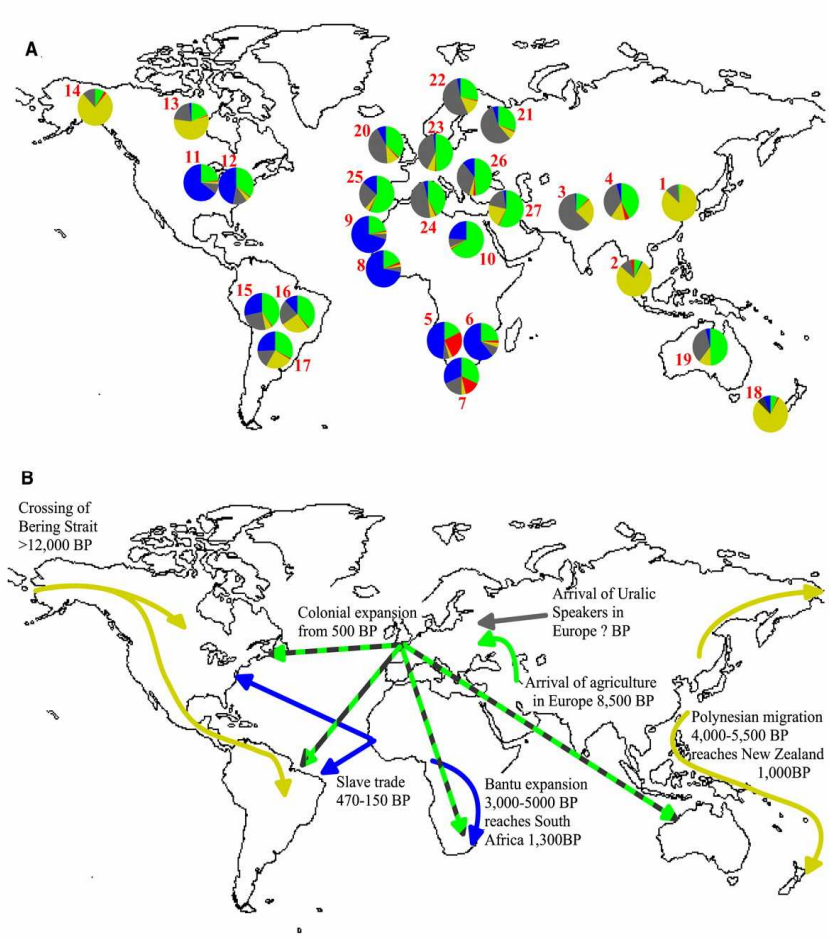
[1] Section 1.3 was originally written as a review article: D. J. Wilson, D. Falush and G. McVean (2005) Germs, genomes and genealogies. *Trends in Ecology and Evolution* **20**: 39-45. All three authors contributed to writing the text.

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).



From D. Falush, T. Wirth, B. Linz, J.K. Pritchard, M. Stephens et al., 2003 Traces of Human Migrations in Helicobacter pylori Populations. Science 299 (5612): 1582-1585. Reprinted with permission from AAAS.

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).

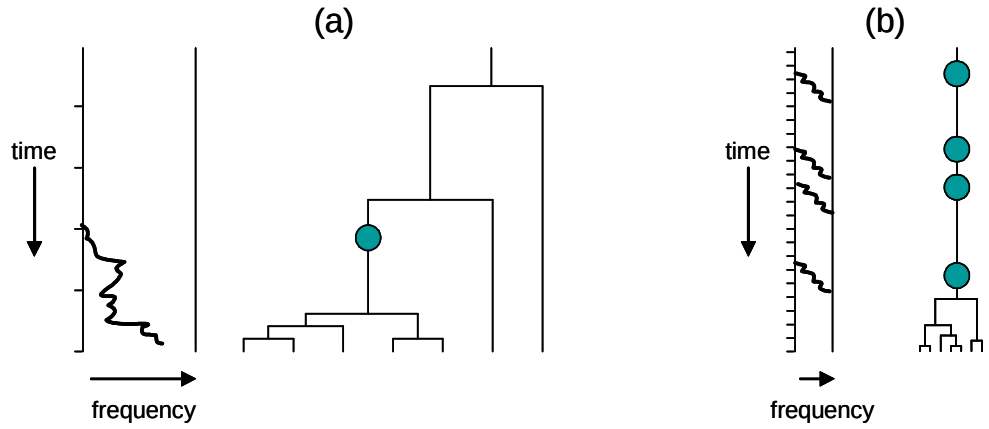


Figure 18 The ability to detect adaptive changes depends on the timescale of evolution. (a) an adaptive mutation occurred since the mrca of the sample, so the genealogy of the sample is distorted. The signature of selection will be visible in the frequency spectra of linked sites. (b) No adaptive mutation 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*

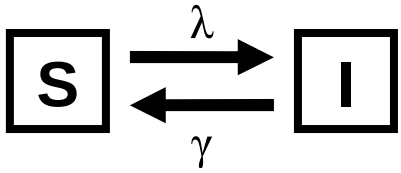
Undoubtedly the coalescent is a useful framework for evolutionary modelling, particularly for recombining organisms. However, its development has primarily been concerned with modelling populations of eukaryotic diploids, and it is not immediately obvious that the coalescent in its native form can be 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

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

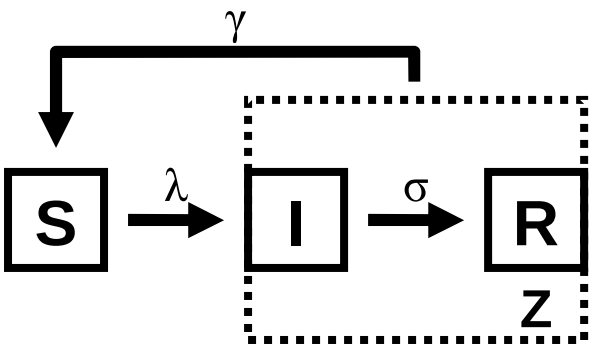
where an asterisk indicates the equilibrium frequency. The basic reproductive number R_0 is defined as the average number of secondary infections caused by a single primary infection in a totally susceptible population. This number is relevant because unless $R_0 \geq 1$ the infection will go extinct.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

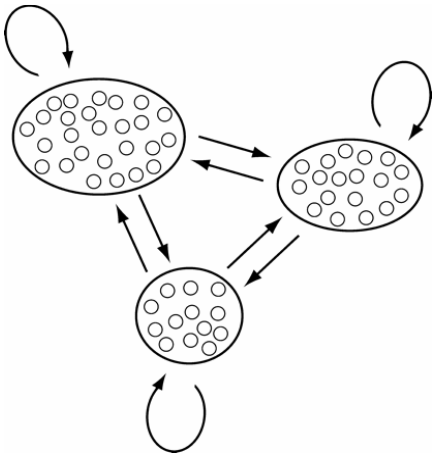


Figure 19 An example of a metapopulation model with many demes. There are $K = 3$ types of deme, which may differ in their population size, extinction/recolonisation rates and migration rates. Individuals can move between demes by migration or recolonisation events, indicated by the arrows. Source: Wakeley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

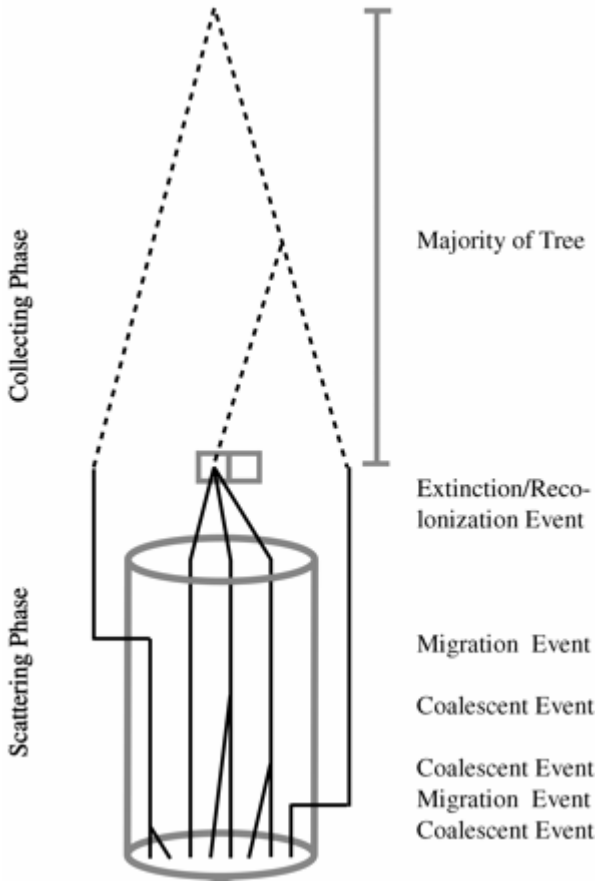


Figure 20 The genealogy of a metapopulation is divided into the scattering phase and the collecting phase. In this example, 8 genes were sampled from a single deme. In the scattering phase a sequence of coalescence, migration and recolonisation events rapidly change the configuration of the ancestral 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

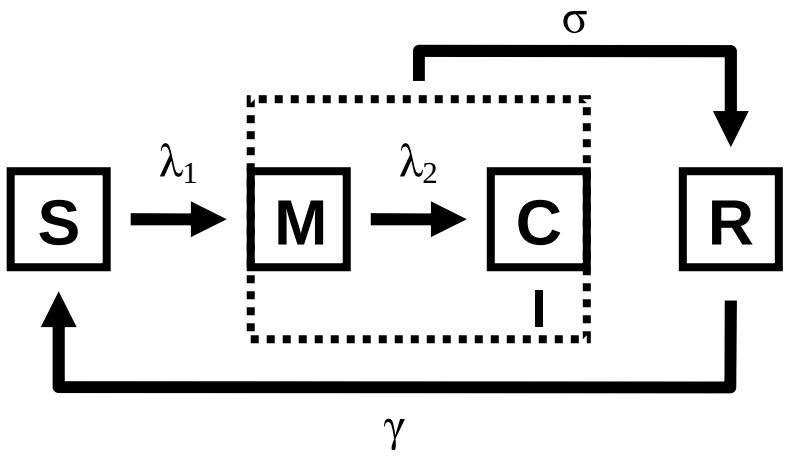
Of the epidemiological models discussed above, SIRS is appropriate for modelling *N. meningitidis* because it includes natural immunity. Although a simplification of the truth, I will show how to incorporate it into the simplest case ($K = 1$) of Wakeley and Aliacar's (2001) metapopulation model. Doing 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.

Chapter 2

Population genetics of *Neisseria meningitidis*

As population studies of *Neisseria meningitidis* have become more numerous and technological developments such as nucleotide sequencing have provided greater resolution for characterising the genetic diversity in those studies, the view of meningococcal biology has itself evolved from a model of clonal descent with a subsidiary role for recombination to a model of a highly recombining population in which high levels of linkage disequilibrium (LD) persist despite frequent horizontal gene transfer. This shift in opinion has been facilitated by applying a variety of mathematical modelling techniques to genetic data. Analysis based on the purely verbal epidemic clone model of Maynard Smith *et al.* (1993) is *post hoc* in the sense that it relies on the identification of clonal complexes using UPGMA trees. Feil *et al.* (1999) count the number of historic mutation and recombination events in an *ad hoc* manner based on observable patterns of genetic mosaicism. Gupta *et al.* (1996) and Holmes *et al.* (1999) utilise more coherent statistical models; however, these are disparate and do not share a common thread. For example, the χ^2 test of Gupta *et al.* (1996) rejects a null model of linkage equilibrium. Linkage equilibrium might be rejected even in a neutrally evolving, panmictic population because of random drift. Holmes *et al.* (1999) rejected two null models, one of complete linkage disequilibrium and one of linkage equilibrium within a gene. While these phylogenetic tests together establish that recombination occurs at intermediate levels in meningococci, the non-parametric nature of the tests means that the actual rate has not been satisfactorily quantified.

Using the coalescent to model the ancestral history of recombining genes offers a coherent approach to evolutionary inference. In the previous chapter I argued that the coalescent is an appropriate starting point for modelling the ancestral history of microparasites, where the effective population size is a complex function of the epidemiological rates of transmission and duration of infection. In this chapter I will test to see whether the coalescent model is an adequate description of meningococcal evolution using housekeeping genes that were sequenced from commensal meningococci in a population of healthy carriers. Two methods of inference are used for fitting the coalescent to *N. meningitidis*, and their respective merits and conclusions are compared. Model adequacy is evaluated by estimating parameters and performing goodness-of-fit tests. By investigating the way in which the model is a poor fit to the data, the standard coalescent can be refined, and in Chapter 3 I investigate the importance of population structure on patterns of genetic diversity in meningococci.

2.1 Description of a carriage population

As discussed in Chapter 1, meningococcal carriage rates are on the order of 10% of the population at large, whereas the rate of disease is closer to 5 persons per 100,000, several orders of magnitude lower. As a result it has been recognised that the overwhelming transmission of *N. meningitidis* occurs between asymptomatic carriers. Samples collected from hospitals and health laboratories comprise, usually solely, of disease-causing meningococci, which represent a minor fraction of all meningococci. Thus carriage studies are extremely important for understanding the normal

transmission cycles of meningococci, and from there the circumstances that lead to invasive disease. In this chapter, 217 isolates collected from healthy young adults in the Czech Republic in 1993 (Jolley *et al.* 2000) are analysed. The isolates were taken from throat swab specimens of 1,400 individuals aged 15 to 24 from nine main sampling locations consisting of schools and workplaces in Prague, České Budejovice, Hradec Králové, Kutna Hora, Plzeň, Olomouc and Opava. All the individuals were healthy with no known contact to patients with invasive disease. The carriage rate was 11.1%. Fragments of seven housekeeping genes were sequenced for MLST (*abcZ*, *adk*, *aroE*, *fumC*, *gdh*, *pdhC* and *pgm*; see Chapter 1), and these are analysed here.

The coalescent is a useful guide to quantifying patterns of genetic variation because under the standard neutral model certain statistics are natural summaries of the data. Under simple mutation models, such as the infinite sites (Watterson 1975) and infinite alleles (Kimura 1968) model, particular summaries of patterns of genetic diversity are related in a direct way to evolutionary parameters such as the mutation rate and recombination rate. In this section, those summaries are used to gain a precursory understanding of the evolution of meningococcal populations before likelihood-based statistical inference is performed explicitly.

2.1.1 Diversity

In the coalescent (Kingman 1982a,b) the time to the most recent common ancestor (mrca) for a pair of sequences is exponentially distributed with rate 1 in units of PN_e generations, where P is the ploidy ($P = 1$ for haploids) and N_e is the effective population size. The simplest model of mutation is the infinite sites model (Watterson

1975) in which a locus of length L undergoes mutation at rate $L\theta/2$ per PN_e generations. The parameter θ is related to the mutation rate per generation, μ, by

$$\theta = 2PN_e\mu .$$

The number of mutation events in t PN_e generations is Poisson distributed with mean $L\theta t/2$, so the average number of mutations that occur in the genealogy of a pair of sequences is $L\theta$. Under the model there are an infinite number of potential sites that undergo mutation, so all mutation events are observed. As a result, the expected number of pairwise differences between a pair of sequences i and j is

$$E(\pi_{ij}) = L\theta . \quad (1)$$

Therefore, the average number of pairwise differences in a sample of size n,

$$\bar{\pi} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\pi_{ij} ,$$

is a natural summary of diversity because Equation 1 implies that $E(\bar{\pi}) = L\theta$. A commonly-used moment estimate for the mutation parameter is $\hat{\theta}_\pi = \bar{\pi}/L$.

The number of segregating sites is another natural summary because in the infinite sites model each mutation results in a new segregating site. The expected sum of branch lengths, T, for the genealogy of n sequences is

$$E(T) = 2\sum_{k=1}^{n-1}\frac{1}{k}, \quad (2)$$

in units of PN_e generations. This is known as the Watterson constant, and was originally calculated for a sample taken from a population evolving according to the standard neutral model by Watterson (1975). The expected number of segregating sites S for a sequence of length L is

$$E(S) = L\theta\sum_{k=1}^{n-1}\frac{1}{k}, \quad (3)$$

because the number of segregating sites equals the number of mutation events, which is Poisson distributed with expectation linear in *T*. Equation 3 suggests a second moment estimate for the mutation parameter $\hat{\theta}_S = S/\left(L\sum_{k=1}^{n-1}1/k\right)$. This is known as Watterson's estimate of the mutation rate.

Table 1 Meningococcal diversity

Locus	*L*	$\bar{\pi}$	*S*	$\hat{\theta}_\pi \times 10^3$	$\hat{\theta}_S \times 10^3$
abcZ	433	19.6	75	45.2	29.1
adk	465	4.07	25	8.76	9.02
aroE	490	32.9	135	67.2	46.2
fumC	465	9.11	48	19.6	17.3
gdh	501	7.13	26	14.2	8.71
pdhC	480	22.9	83	47.7	29.0
pgm	450	20.1	81	44.6	30.2
Total	3284	115.8	473	35.3	24.2

Table 1 shows that there is considerable heterogeneity in diversity between housekeeping loci, ranging from $\bar{\pi}$ = 4.07, *S* = 25 for *adk* up to $\bar{\pi}$ = 32.9, *S* = 135 for *aroE*. The two measures of diversity $\bar{\pi}$ and *S* give a similar account of diversity in the housekeeping genes, and provide comparable estimates of the mutation parameter θ, ranging from 0.00876 for *adk* to 0.0672 for *aroE*. Across loci, the average proportion of sites that differ between sequences is 3.5%, and 14% of sites are

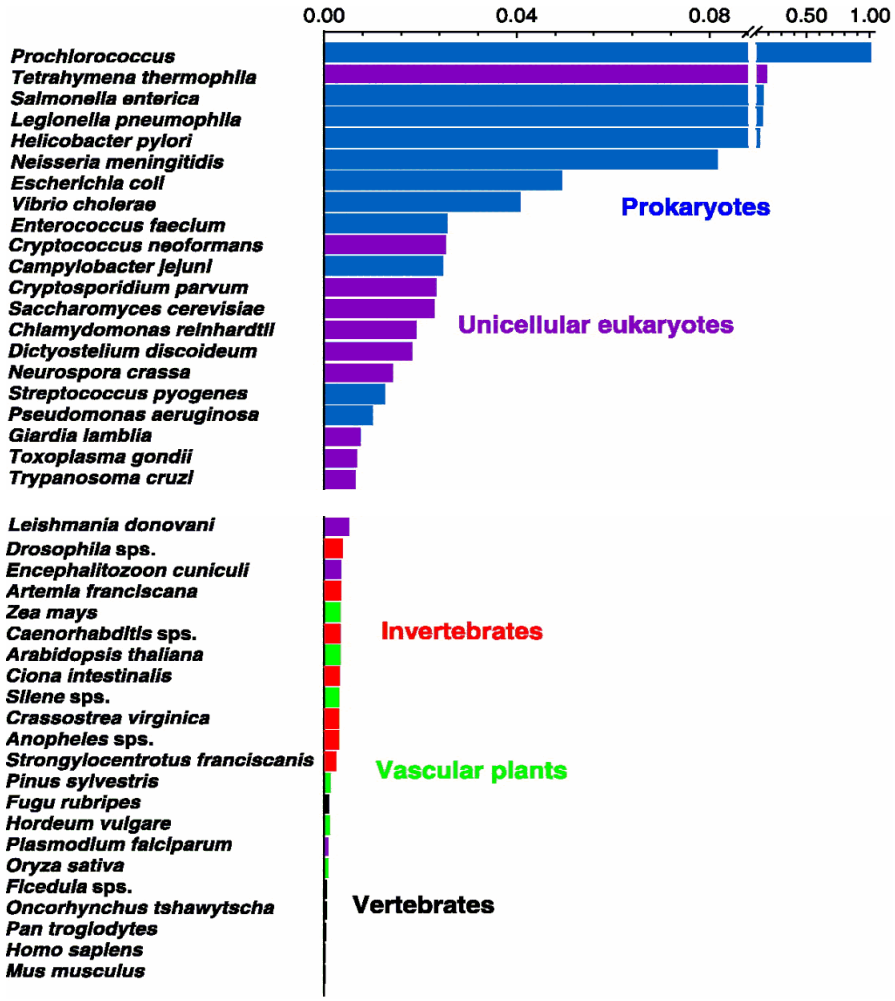


Figure 1 Estimates of the population mutation rate θ/P for different taxa. Source: Lynch and Conery (2003).

From M. Lynch and J.S. Conery, 2003 The Origins of Genome Complexity. Science 302 (5649): 1401-1404. Reprinted with permission from AAAS.

segregating. The Watterson estimate of θ is 0.0242 per site for the concatenated sequence.

The diversity of these housekeeping genes is of the same order of magnitude as that observed in other prokaryotes (Figure 1), which is considerably larger than for unicellular eukaryotes, and more so for multicellular eukaryotes. In general there is an inverse relationship between θ and organism size (Lynch and Conery 2003). The estimates of θ in Figure 1 are for synonymous changes only, in an attempt to estimate the neutral mutation rate. Therefore, the estimate of $\theta = 0.08$ for *N. meningitidis*

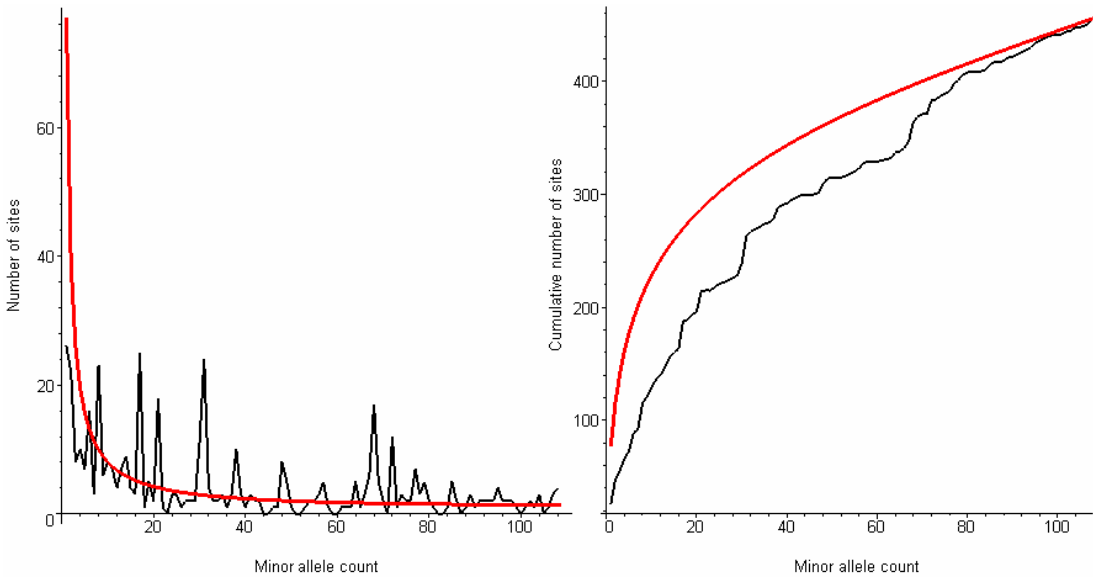


Figure 2 Observed distribution of minor allele count across biallelic segregating sites. In both figures the red line indicates the neutral expectation. Left: plot of the number of sites with a given minor allele count. Right: plot of the cumulative number of sites with a given minor allele count.

(Figure 1) is higher than that observed here, presumably because of functional constraint in the housekeeping genes. Lynch and Conery's estimate is based on 11 sequences of housekeeping genes from a collection of 107 isolates representing global disease (Maiden *et al.* 1998).

2.1.2 Frequency distributions

In the coalescent with infinite sites mutation, the expected number of mutations η_i that reach abundance i ($i = 1,2,\ldots,n-1$) is $E(\eta_i) = \theta / i$ (Fu 1996). In a real data set it is not usually possible to determine whether a particular allele is derived or ancestral, so it is necessary to take the folded distribution,

$$E(\eta_i + \eta_{n-i}) = \theta / i + \theta / (n-i),$$

where i $(i = 1,2,\ldots,n/2)$ is the count of the less frequent allele (the minor allele) for a biallelic site; in the infinite sites model, segregating sites can only be biallelic. Figure 2 shows the observed frequency distribution of minor alleles, aggregated over biallelic sites at all seven loci (left hand graph, black line). In total there were 456 biallelic sites. The red line indicates the neutral expectation, using the Watterson estimate of θ. Because of the small number of sites involved, it is difficult to assess whether there is any deviation from the neutral expectation. In the right hand graph, the observed cumulative distribution for the number of sites with a given minor allele is plotted (black line), with the neutral expectation (red line). It is clear that there is a dearth of sites with a small minor allele count. That is to say that there is an excess of sites with intermediate frequency alleles. Such a pattern might be caused by ascertainment bias when choosing the seven MLST loci to type, if loci with high diversity were preferred. The effect of ascertainment depends on the size of the sample used for ascertainment. That 109 meningococcal isolates were used to choose the MLST loci, and loci with intermediate rather than high diversity were preferred suggests that the observed excess of intermediate frequency alleles is not readily explained by ascertainment bias (Urwin and Maiden 2003). An alternative explanation for an excess of intermediate frequency alleles is ancestral population structure, which is investigated in more detail in section 2.2.4 and Chapter 3.

Rather than report the average diversity between pairs of sequences, the whole distribution of pairwise differences can be plotted to demonstrate the degree of genetic clustering in a population. In a clonal population, a deep branch at the root of the evolutionary tree that partitions the population into k and $(n-k)$ individuals respectively will cause a bi-modal distribution because the ${}^{k}C_{2} + {}^{n-k}C_{2}$ pairwise

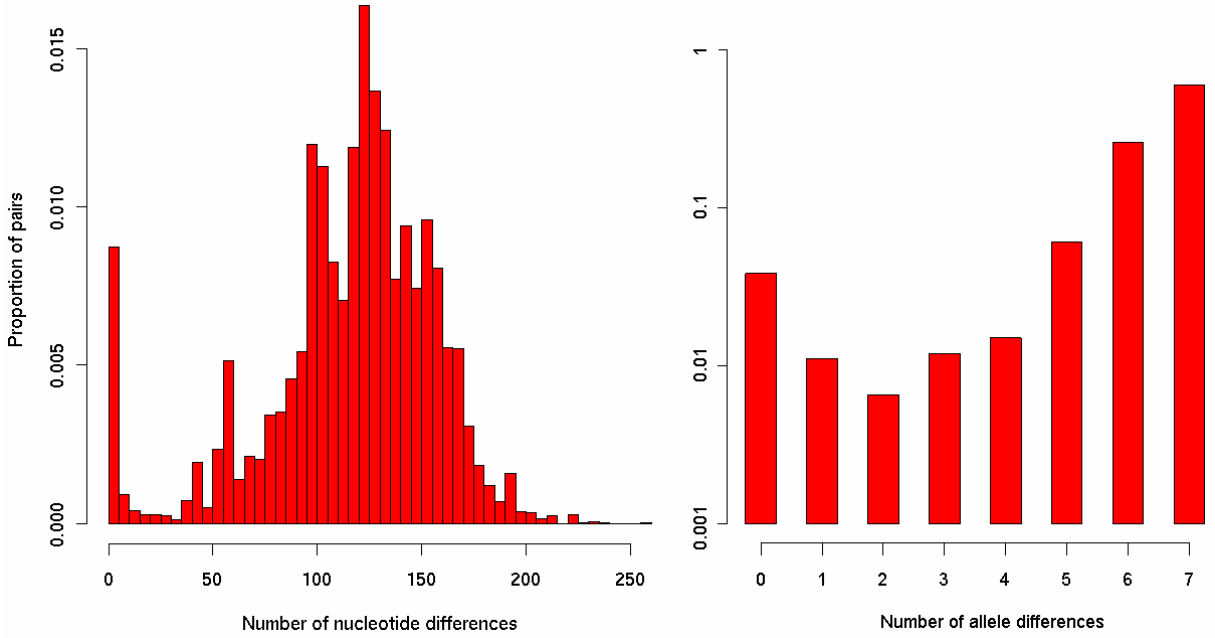


Figure 3 Mismatch distributions for isolates sequenced at the seven MLST loci. Left: histogram of the number of nucleotide differences between pairs of isolates, out of 3284 bp in total. Right: bar chart of the number of allele differences between pairs of isolates, out of 7 loci.

comparisons within each partition will exhibit fewer differences than the $k(n-k)$ pairwise comparisons across the root-branch partition. In a recombining population the bimodality may be less pronounced because shifts in the topology of the evolutionary tree along the sequence cause the population to be partitioned differently at different parts of the sequence. In the extreme case of linkage equilibrium, the distribution would be binomial, which is unimodal. On the other hand, strong population structure might maintain a deep partition in spite of recombination.

Figure 3 shows the mismatch distributions for the Czech carriage study, plotted as a histogram of the pairwise number of nucleotide differences (left hand graph) and a bar chart of the pairwise number of allele differences (right hand graph). The nucleotide mismatch distribution is bi-modal, with a peak at zero and a peak near 120, indicating that there is some genetic clustering of individuals, whether it be caused by limited

Table 2 Recombination-sensitive statistics

Locus	$V(\pi)$	R_m	$\mathrm{cor}(r^2, d)$	$\mathrm{cor}(D', d)$	$\mathrm{cor}(G4, d)$
abcZ	132.7	10	-0.235	-0.329	-0.332
adk	7.2	3	-0.216	0.104	0.151
aroE	657.5	19	-0.434	-0.095	-0.061
fumC	27.8	12	-0.111	-0.164	-0.116
gdh	20.6	5	-0.251	-0.316	-0.264
pdhC	193.4	14	-0.255	-0.139	-0.091
pgm	144.9	9	-0.373	0.030	0.008

recombination or population structure. Likewise, the allelic mismatch distribution is bimodal, with peaks at the extreme values of zero and 7. In agreement with previous work (Holmes *et al.* 1999), these graphs demonstrate that whatever the rate of recombination may be in this population of meningococci, it is not sufficiently high to obliterate genetic structuring.

2.1.3 Recombination

Coalescent theory tells us that the variance in the number of pairwise differences is sensitive to the rate of recombination in a standard neutral model. Specifically,

$$V(\pi) = \left[\frac{n+1}{3(n-1)}\right]\theta + f(\rho, n)\theta^2, \qquad (4)$$

where $f(\rho, n)$ is a function of the recombination rate and sample size (Wakeley 1997). Hudson (1987) and Wakeley (1997) have exploited this relationship to obtain

moment estimators of the recombination rate, similar to those in section 2.1.1, but less simple. The observed variance in the number of pairwise differences, shown for each locus in Table 2, ranges from 7.2 for *adk* up to 657.5 for *aroE*. In fact sorting the loci by the magnitude of $V(\pi)$ produces exactly the same ordering as sorting the loci by the magnitude of $\overline{\pi}$.

Amongst other things, recombination causes genetic incompatibilities in an alignment of nucleotide sequences. For two biallelic loci A and B there are four possible haplotypes: *AB*, *Ab*, *aB* and *ab*. Under the infinite sites model with no recombination, it is impossible to observe all four haplotypes in a sample of sequences. Such a scenario is called a genetic incompatibility, because the data are incompatible with the genetic model. Incompatibility can be caused by violation of either the mutation model (recurrent mutation can cause all four haplotypes to arise) or the assumption of no recombination (a shift in topology can allow all four haplotypes to arise). When the mutation rate is low, the infinite sites model is a reasonable approximation, and genetic incompatibility is indicative of recombination. Several authors (Hudson and Kaplan 1985; Myers and Griffiths 2003) have used the number of genetic incompatibilities to estimate a lower bound on the number of recombination events in the ancestral history of the sequences under an infinite sites model. Their estimators are known as R_m and R_h respectively. Whilst the lower bound on the number of recombination events in a finite sites model (where recurrent mutation is allowed) must always be zero, R_m or R_h can be used nonetheless as a statistic that is sensitive to the recombination rate.

Whilst it is true that R_h is a more efficient lower bound than R_m in the sense that $R_h \geq R_m$ under the infinite sites model (Myers and Griffiths 2003), the former is considerably more computationally intensive because it involves an optimisation step, and for that reason using R_h is not strictly deterministic. Myers and Griffiths (2003) give an efficient way to calculate R_m. Define

$$B_{ij} = \begin{cases} 1 & \text{if sites } i \text{ and } j \text{ are incompatible} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

then $R_m = R_m^{(L)}$ can be solved iteratively using

$$R_m^{(j)} = \max\left\{R_m^{(i)} + B_{ij}; i = 1,2,\ldots, j-1\right\}$$

and the boundary condition $R_m^{(1)} = 0$. I calculated R_m for each locus; the values are displayed in Table 2. The lowest value of R_m was 3 for *adk*, and the highest was 19 for *aroE*. This reflects the extreme status of these two loci for the other measures of diversity and recombination. However, sorting the loci by the magnitude of R_m does not produce exactly the same order as sorting them for $V(\pi)$.

Recombination causes a breakdown in linkage disequilibrium (LD) along the sequence. There are various ways to measure LD between a pair of sites. A natural measure is to take the difference between the observed haplotype frequency and that expected under linkage equilibrium. Take, for example, two biallelic loci A and B, as before. The LD for haplotype *AB* can be expressed as

$$D_{AB} = f_{AB} - f_A f_B, \quad (6)$$

where f is the observed frequency of the haplotype or allele. In this simple example, $(D_{AB} = D_{ab}) = -(D_{Ab} = D_{aB})$. The expectation of D is zero under linkage equilibrium. For a pair of biallelic loci, Equation 6 can be interpreted as a covariance in allele

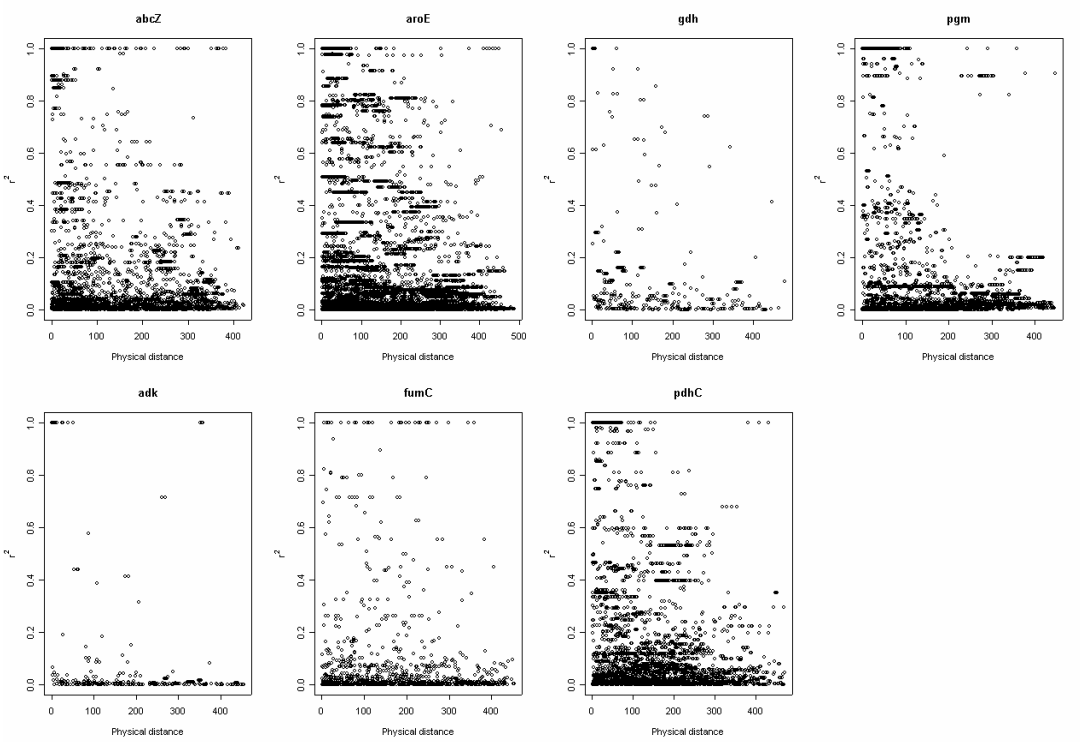


Figure 4 Breakdown in linkage disequilibrium, as measured by r^2, with physical distance in each of seven housekeeping loci from the Czech carriage study.

frequencies. A natural way to compare LD from different pairs of biallelic loci is to standardise this covariance, i.e. calculate the correlation coefficient

$$r = \frac{f_{AB} - f_A f_B}{\sqrt{f_A(1-f_A) f_B(1-f_B)}}, \tag{7}$$

or, to remove the arbitrary sign, r^2 (Hill and Robertson 1968). Even under complete linkage, r^2 can only equal one if the allele frequencies are the same. To overcome the problem, Lewontin (1964) introduced D', which scales the covariance by its theoretical maximum given f_A and f_B.

$$D' = \begin{cases} \dfrac{f_A f_B - f_{AB}}{\min\{f_A f_B, (1-f_A)(1-f_B)\}} & \text{if } f_{AB} < f_A f_B \\ \dfrac{f_{AB} - f_A f_B}{\min\{f_A(1-f_B), (1-f_A) f_B\}} & \text{if } f_{AB} \geq f_A f_B \end{cases}. \tag{8}$$

To quantify the breakdown in LD along a sequence, one can look for a decrease in r^2 or D' with increasing physical distance. Figure 4 illustrates the decay in r^2 with physical distance for each of the seven housekeeping loci. Each data point corresponds to a pair of sites. The decay in LD can be quantified as the correlation between the LD statistic and physical distance, *d*. In the presence of recombination, LD is expected to decrease as physical distance increases, so the correlation coefficient should be negative.

Table 2 displays the correlation between LD and distance for both r^2 and D'. A third LD statistic, G4, is also used, which corresponds to the four-gamete test of Hudson and Kaplan (1985). G4 is defined as $(1 - B_{ij})$ for a pair of sites *i* and *j* (see Equation 5); it equals zero if there is a genetic incompatibility and one otherwise. Incompatibility is expected to increase with distance in the presence of recombination; therefore G4 should also show a negative correlation with distance. For all three correlation coefficients in Table 2, the stronger the correlation, the stronger the relationship is between LD and distance. All three correlation coefficients show broadly the same pattern: a negative correlation indicative of recombination. Sorting the loci by the magnitude of the correlation, $\mathrm{cor}(D', d)$ and $\mathrm{cor}(G4, d)$ produce the same order with *abcZ* exhibiting the strongest relationship between LD and distance, and *pgm* the weakest. This pattern differs, however, from that presented by $\mathrm{cor}(r^2, d)$, for which *aroE* shows the strongest correlation and *fumC* the weakest, and the other recombination-sensitive statistics. These differences may amount to the relative sensitivity of the statistics to the mutation rate, of which each is necessarily also a function. To learn more about the evolutionary parameters for these loci and

assess the adequacy of any particular model, it is necessary to fit a statistical model formally to the data.

2.2 Fitting the standard neutral model

The purpose of fitting a statistical model to genetic data, as opposed to a purely descriptive analysis, is (i) to obtain estimates of the parameters which are presumably of some evolutionary relevance, and (ii) to challenge the model by exploring its deficiencies and in so doing refine our understanding of the process of evolution that underlies the data. For all the elegance of the standard neutral coalescent, the ease with which results can be obtained for various quantities of interest and the efficiency of simulation (see section 2.2.3), performing likelihood-based inference under the coalescent is not straightforward. No analytic expressions exist for the likelihood of a sample of gene sequences, or haplotypes **H**, under the coalescent. Therefore the likelihood must be evaluated numerically.

The likelihood of **H** can be computed with reference to a given genealogy, or set of genealogies, *G*. In principal, the likelihood might be calculated from

$$P(\mathbf{H} \mid \Theta) = \int P(\mathbf{H} \mid \Theta, G) P(G) \mathrm{d}\, G,$$

where $P(\mathbf{H} \mid \Theta)$ is the likelihood of the parameters Θ given the data **H**, $P(G)$ is the probability of the genealogy, specified by the coalescent, and $P(\mathbf{H} \mid \Theta, G)$ is the conditional likelihood of the data given the genealogy, obtained using the pruning

algorithm (Felsenstein 1981) for a finite sites mutation model[1]. In practice, the integral needs computing numerically, and a naïve approach would be to calculate

$$P(\mathbf{H} \mid \Theta) \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathbf{H} \mid \Theta, G^{(i)}),$$

for large M, where $G^{(i)}$ is simulated from $P(G)$. However, for the coalescent this method is not feasible because almost all trees will contribute a negligible amount to the sum. Only once in a million draws might the conditional likelihood contribute significantly (Stephens 2003). Various techniques have been employed in an attempt to solve this problem (discussed further in Chapter 4). Amongst these is the composite likelihood approach (Hudson 2001; McVean *et al.* 2002), which has been used to estimate recombination rates in *N. meningitidis*.

2.2.1 Composite likelihood inference

This approach relies on approximating the likelihood as the product over all pairs of columns in the alignment

$$P(\mathbf{H} \mid \Theta) \approx \prod_{i,j} P(\mathbf{H}_{\cdot i}, \mathbf{H}_{\cdot j} \mid \Theta), \tag{9}$$

where $\mathbf{H}_{\cdot i}$ represents the n sequences at the ith column in the alignment. To simplify matters further, McVean *et al.* (2002) assume that the mutation rate is known, using an estimate that is modified to allow for finite-sites mutation

$$\hat{\theta}_{Mc} = \frac{\ln(L) - \ln(L - S)}{L \sum_{k=1}^{n-1} 1/k},$$

[1] Strictly speaking, the likelihood function $L(\Theta)$ is defined to be proportional to the conditional probability density function $P(\mathbf{H}|\Theta)$. However, I have used *likelihood* synonymously for $L(\Theta)$ and $P(\mathbf{H}|\Theta)$.

and only biallelic sites are used for inference. The recombination rate $\rho = 2PN_e r$ is estimated by assuming that the rate of recombination between a pair of sites separated by d_{ij} nucleotides is

$$r_{ij} = rd_{ij}.$$

In *N. meningitidis*, homologous recombination occurs by donor-recipient style transformation in which a fragment of naked DNA is endocytosed by the cell from the environment and incorporated into the recipient's genome (Lorenz and Wackernagel 1994). The fragment length of the recombinant DNA tract can be modelled as exponential with mean $\bar{t}$ (Wiuf and Hein 2000). In such a model, only recombination events that have one, but not both, breakpoints between a pair of loci affect the linkage of the loci. As a result, the effective rate of recombination between loci *i* and *j* separated by distance d_{ij} is

$$\int_{-\infty}^{0} \frac{r}{2}\left[\exp\left\{-\frac{-u}{\bar{t}}\right\} - \exp\left\{-\frac{d_{ij}-u}{\bar{t}}\right\}\right] \mathrm{d}u + \int_{0}^{d_{ij}} \frac{r}{2}\exp\left\{-\frac{d_{ij}-u}{\bar{t}}\right\} \mathrm{d}u = r\bar{t}\left(1-\exp\left\{-d_{ij}/\bar{t}\right\}\right), \tag{10}$$

where $r/2$ is the rate of initiation of recombination per bp per generation, u is the position at which recombination is initiated, $\exp\{-(-u)/\bar{t}\} - \exp\{-(d_{ij}-u)/\bar{t}\}$ is the probability that the tract terminates between loci *i* and *j* if it initiates outwith, and $\exp\{-(d_{ij}-u)/\bar{t}\}$ is the probability that the tract length is longer than $(d_{ij}-u)$ if it initiates between them. For loci separated by much less than $\bar{t}$, the rate is approximated by

$$r\bar{t}\left(1-\exp\left\{-d_{ij}/\bar{t}\right\}\right) = rd_{ij}, \tag{11}$$

because for $x << 1$,

$$1-\exp\{-x\} \approx x.$$

For loci that are weakly linked, the effective rate of recombination is

$$\lim_{d_{ij}/\bar{t}\to\infty} r\bar{t}\left(1-\exp\left\{-d_{ij}/\bar{t}\right\}\right) = r\bar{t}\ . \qquad (12)$$

Thus estimates of recombination between pairs of distant loci can be contrasted to estimates between pairs of proximate loci, and the tract length estimated.

By using only biallelic loci, the nucleotides can be converted from `A`, `G`, `C` and `T` into `0` and `1`, where `0` represents the rare allele. For a pair of biallelic loci there are a possible

$$1+N+\frac{N(N-1)(N+4)}{6}+\frac{(N-1)(N+2)}{2}$$

unordered, unlabelled, exchangeable sample configurations, where $N = n/2$. $P(\mathbf{H}_{\cdot i}, \mathbf{H}_{\cdot j} \mid \Theta)$ can be calculated for any given r_{ij} using the importance sampler of Fearnhead and Donnelly (2001), and the computation proceeds by calculating this pairwise likelihood for a finite number of values of r_{ij}, which is then stored in a look-up table. A single value of θ is used in generating the look-up table. Whilst the importance sampling step is extremely computationally intensive and increasingly so for increasing sample size n, the computation of the composite likelihood (Equation 9) is rapid. Maximum likelihood estimates of ρ can then be obtained using the interpolated composite likelihood curve. The method is implemented in the package LDhat (available from http://www.stats.ox.ac.uk/~mcvean).

Table 3 Composite likelihood estimates of recombination and mutation rates[2]

Locus	$\hat{\theta}_{Mc} \times 10^3$			$\hat{\rho} \times 10^3$		
	Czech Carriage	Czech Disease	Global Disease	Czech Carriage	Czech Disease	Global Disease
abcZ	36.0	33.5	36.7	19.2	7.5	8.1
adk	6.8	8.7	7.2	12.4	5.4	2.7
aroE	61.2	80.5	79.9	9.7	2.6	6.2
fumC	18.3	18.8	16.5	23.2	34.0	23.2
gdh	10.3	11.7	11.1	17.6	31.1	13.0
pdhC	35.7	34.2	35.2	22.5	8.9	17.8
pgm	38.3	33.2	33.7	16.8	6.7	22.9

2.2.2 Parameter estimates

The method was applied to three meningococcal datasets, including the Czech carriage study[2]. The other two datasets comprised a previously unpublished collection of 53 disease-causing isolates sampled from the Czech Republic during 1993 (Jolley *et al.* 2005), and a collection of 107 disease causing isolates representing global diversity (Maiden *et al.* 1998). Because the computation time of the composite likelihood method increases disproportionately with the number of sequences, several

[2] Parameter estimates using LDhat were obtained by Gil McVean, and have been published: K.A. Jolley, D.J. Wilson, P. Kriz, G. McVean and M.C.J. Maiden (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Molecular Biology and Evolution* **22**: 562-569.

random samples of 100 sequences were taken from the Czech carriage and disease collections for analysis, and the results averaged. The estimates of the mutation and recombination rates are shown in Table 3.

For the Czech carriage study, the estimates of the mutation rate θ are very close to the Watterson estimates ($\hat{\theta}_S$, Table 1). Except for *adk*, $\hat{\theta}_{Mc}$ is higher than $\hat{\theta}_S$, reflecting that fact that in a finite sites mutation model the sequences become saturated with mutations so the estimate based on an infinite sites assumption is downwardly biased. There is no obvious relationship between the estimates of the recombination rate and the summary statistics presented in Table 2, partly because these statistics are also sensitive to the mutation rate. In contrast to the mutation rates, which are lowest for *adk* and highest for *aroE*, *aroE* exhibits the lowest recombination rate (0.0097) and *fumC* the highest (0.0232). Interestingly, the relative mutation and recombination rates appear to be generally conserved between carriage and disease collections, which is reassuring from the perspective of measuring parameters that are evolutionarily meaningful. Overall, the mutation rates were comparable between carriage and disease collections, but the rate of recombination was diminished in disease-causing isolates for four of the seven loci.

By calculating a composite likelihood for pair of sites at different loci, a recombination rate of $2PN_e r\bar{t}$ (see Equation 12) was estimated at 28.2. By calculating a composite likelihood for all loci using only pairs of sites at the same locus, a recombination rate of $2PN_e r$ (see Equation 11) was estimated at 0.0256. Therefore the mean tract length $\bar{t}$ was estimated to be 1,100 bp.

Table 4 Relative importance of recombination and mutation[2]

Locus	ρ/θ			Relative rate of diversification
	Czech Carriage	Czech Disease	Global Disease	Czech Carriage
abcZ	0.53	0.23	0.22	13.3
adk	1.83	0.62	0.38	8.8
aroE	0.16	0.03	0.08	5.9
fumC	1.27	1.81	1.41	13.7
gdh	1.70	2.66	1.17	13.4
pdhC	0.63	0.26	0.51	16.5
pgm	0.44	0.20	0.68	10.8

Obtaining estimates of the mutation and recombination rate using a statistical (albeit approximate) model and performing inference using established techniques allows the relative contribution of recombination to mutation, r/μ to be quantified by taking

$$\frac{\rho}{\theta} = \frac{2PN_e r}{2PN_e \mu}. \tag{13}$$

The estimates are shown in Table 4, which range from 0.16 for *aroE* to 1.83 for *adk* in the Czech carriage study. The ranges are not dissimilar for the other isolate collections. Across loci, the rate of mutation and the rate of recombination appear to be broadly of the same order of magnitude. Note that r/μ is actually twice the relative rate at which recombination events occur ($r/2$) to mutation (μ).

The evolutionary significance of the relative rates of recombination and mutation depend more upon the rates at which each process causes genetic diversification, rather than their underlying rates of incidence. For every recombination event, an average of 1,100 bp is affected, which is a much greater number of sites than a point mutation affects. Of those, the proportion that will change as a result can be calculated using the average diversity at each locus, which is estimated using $\bar{\pi}/L$ from Table 1. Thus, the relative rate of diversification is calculated as

$$\frac{r/2\times\bar{t}\times\bar{\pi}/L}{\mu}=\frac{1}{2}\frac{\rho}{\theta}\times\bar{t}\times\frac{\bar{\pi}}{L}, \qquad (14)$$

the results of which are given in Table 4 for the Czech carriage study. The relative rate of diversification ranges from 5.9 for *aroE* to 16.5 for *pdhC*, indicating that in terms of generating genetic novelty, recombination is some ten times more important than mutation. This is consistent with previous estimates in the (broad) range of 3.6 – 275 (Feil *et al.* 1999; Jolley *et al.* 2000; Feil *et al.* 2001).

Possible confusion arises from the assumption made in Equation 13 that the effective population size for mutation and recombination is the same. In Chapter 1 a SIRS metapopulation model for microparasites was used as a basis for coalescent modelling in *N. meningitidis*. In that model, the effective population size for mutation (say N_θ) and recombination (say N_ρ) differ, so that

$$N_\rho = N_\theta\left(\frac{1+\beta_1 S^*}{1+2\beta_2 I^*+\beta_1 S^*}\right) \qquad (15)$$

(see Chapter 1, Equations 3 and 5b), where β_1 and β_2 are the primary and secondary rates of infection respectively, and I^* is the equilibrium prevalence of infection and S^* the equilibrium frequency of susceptible hosts, both of which are also a function of

the average duration of infection and rate of loss of immunity. Note that Equation 15 implies that $N_\rho \leq N_\theta$. This result suggests that the estimates of ρ in Table 3, and the estimates of ρ/θ in Table 4 should be up-weighted by some unknown amount. However, the estimated relative rate of diversification (Table 4) does not need to be adjusted. In the metapopulation model, N_ρ is lower than N_θ because a certain fraction of ancestral recombination events immediately coalesce again, rather than the two lineages migrating to separate hosts by independent transmission events. These invisible recombination events have no effect on diversity, and therefore do not contribute to the relative rate of diversification. If the estimates of ρ and ρ/θ were up-weighted using Equation 15, the estimated relative rate of diversification would need to be correspondingly down-weighted.

2.2.3 Simulating under the coalescent

Simulating from the model has a variety of applications, including exploratory analyses, inference, goodness-of-fit testing and prediction. Simulating the ancestry of a sample of sequences under the coalescent is efficient, particularly compared to simulating using an individual-based Wright-Fisher model (Fisher 1930; Wright 1931) in which the whole population is modelled. For a sample of n sequences, the genealogy is simulated as follows (Hudson 1990), where time is measured in units of PN_e generations.

1. Initially there are $k = n$ lineages.
2. Calculate the rates of coalescence and recombination respectively as

$$\lambda_C = \frac{k(k-1)}{2}$$
$$\lambda_R = \frac{k\rho}{2}.$$

3. Generate an exponentially distributed random variate with rate $\lambda_C + \lambda_R$ for the waiting time until the next ancestral event.
4. With probability $\lambda_C / (\lambda_C + \lambda_R)$ choose two lineages uniformly at random to coalesce, and decrement *k* by 1. Otherwise, choose a lineage uniformly at random to recombine, and increment *k* by 1. The recombination breakpoint is chosen uniformly at random along the sequence.
5. Repeat from step 2 until *k* = 1.

Because the rate of coalescence is quadratic in *k* and the rate of recombination is only linear in *k*, the algorithm will finish in finite time (Griffiths and Marjoram 1997). A particularly useful speed-up is to calculate an effective recombination rate η, which excludes sites in a lineage that are not ancestral to the sample, unless the non-ancestral sites are surrounded by sites that are ancestral to the sample. Except in the latter case, recombination breakpoints are then not allowed to occur at non-ancestral sites.

Having simulated the genealogical history, mutations can be superimposed using a finite-sites mutation model with *C* states. The forward-in-time transition probability matrix, $\mathbf{P}^{(t)}$, gives the probability $p_{ij}^{(t)}$ of being in state *j* time $t\,PN_e$ generations after being in state *i*, and can be found by exponentiating the mutation rate matrix **G** such that

$$\mathbf{P}^{(t)} = e^{t\mathbf{G}} \qquad (16)$$

(Grimmett and Stirzaker 2001). The bifurcating genealogy at a single site is known as the marginal genealogy.

1. For each site, the state of the oldest node in the marginal genealogy (the mrca) is drawn from the stationary distribution of the mutation rate matrix, assuming it is ergodic.
2. For each node that is a descendant of the current node, the state of the descendant is drawn from a multinomial distribution with parameters $\left(p_{i1}^{(t)}, p_{i2}^{(t)}, \ldots, p_{iC}^{(t)}\right)$, where i is the state of the current node and t is the length of the lineage connecting the nodes.
3. Step 2 is repeated for each of the descendant nodes until the terminal nodes (the contemporary sample) are reached.

2.2.4 Goodness-of-fit testing

There are two good reasons for performing goodness-of-fit testing for an evolutionary model, which is easily implemented using coalescent simulation. The first is to ask the polarised question, “Does the model adequately fit the data?” It is essential that any model be falsifiable, and the way to falsify a model is through goodness-of-fit testing. However, on the understanding that all models are deficient in some respect, the second purpose of goodness-of-fit testing is to ask the more pertinent question, “In what way does the model fail to fit the data?” Addressing this latter question is an integral part of the iterative process of refining our models, and hence our understanding, of evolution. It is arguably the principal role of mathematical modelling in biology.

In a maximum likelihood framework, goodness-of-fit testing can be performed by taking some summary statistic of the data, generating a null distribution for that statistic by simulating under the estimated parameters, and calculating the probability of observing as such an extreme value of the statistic under the model. This probability is usually called a *p*-value. Statistics are chosen that summarise some aspect of the data that either (i) it is important the model describes well and/or (ii) it is suspected the model does not describe well. Of all the statistics used to test departures from the standard neutral model, Tajima's *D* (Tajima 1989) is perhaps the most well-used. Tajima's *D* exploits the fact that the pairwise diversity estimator $\hat{\theta}_\pi$ and Watterson's estimator $\hat{\theta}_S$ of the mutation rate use different information. Under neutrality, the two have equal expectation, but under various departures from the standard neutral model, the two will differ. Tajima's *D* is normalised so that it has expectation zero and a variance of approximately one under the standard neutral model.

$$D = \frac{\overline{\pi} - S/a_n}{\sqrt{e_1 S + e_2 S(S-1)}}, \tag{17}$$

where

$$e_1 = \frac{n+1}{3a_n(n-1)} - \frac{1}{a_n^2},$$

$$e_2 = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{na_n} + \frac{b_n}{a_n^2}\right),$$

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i} \qquad \text{and} \qquad b_n = \sum_{i=1}^{n-1}\frac{1}{i^2}.$$

Two extreme departures from the standard neutral model can be envisaged. In the first, the tree is close to star-like so that coalescent events occur closer to the root than expected, possibly because of demographic growth or a recent selective sweep. This

Table 5 Tajima's *D* in meningococcal populations[3]

Locus	Czech Carriage	Czech Disease	Global Disease
abcZ	1.15	–0.268	1.063
adk	0.817	0.512	0.392
aroE	0.926	–0.966	0.498
fumC	0.328	–0.221	0.157
gdh	1.355	1.126	**1.742**
pdhC	**1.433**	**1.944**	**1.842**
pgm	0.811	0.541	0.286
Concatenated	**1.101**	0.106	0.833

causes an excess of low-frequency variants, so *S* is elevated relative to $\bar{\pi}$, and *D* is negative. In the second, coalescent events occur closer to the tips than expected, possibly because of population subdivision causing a deep root branch. This scenario causes a dearth of low-frequency variants, so *S* is diminished relative to $\bar{\pi}$, and *D* is positive.

In addition to Tajima's *D*, goodness-of-fit testing was conducted using the number of unique haplotypes, *H*. The number of unique haplotypes can be thought of as a balance between recombination, which will act to increase *H* by creating novel combinations of alleles, and population structure, which will act to decrease *H* by preventing recombination between genetically isolated subpopulations. The observed number of unique haplotypes was 88 in the Czech carriage study, and 50 on average

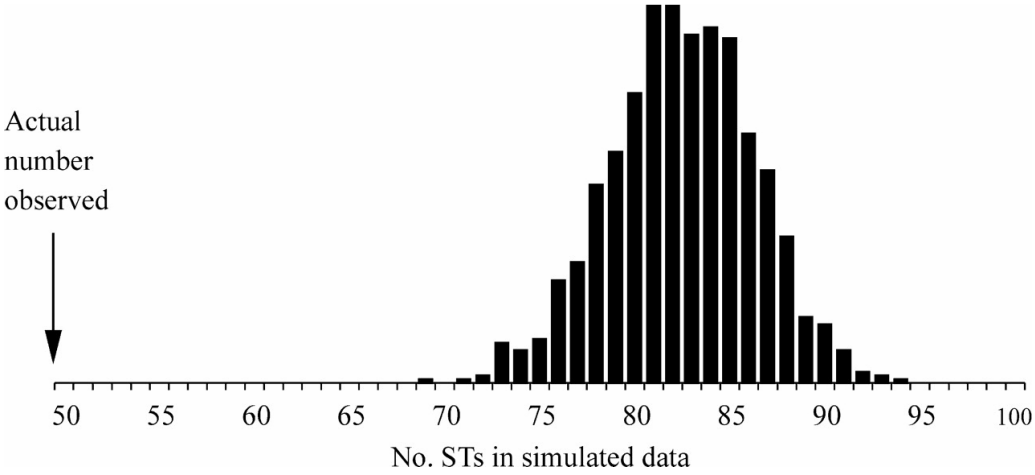


Figure 5 Null distribution of the number of unique haplotypes (STs) under the parameters estimated by LDhat for a sub-sample of 100 sequences[3]. The observed number was 50, which is outside the range of simulated values.

in the random subsets of 100 sequences used for inference. The observed value of Tajima's *D* for each locus (and all loci combined) is recorded in Table 5 for each of the three meningococcal isolate collections.

Significance testing was undertaken using 10,000 simulations with $\hat{\theta}_{Mc}$ and the composite likelihood estimate of ρ. For each simulation *H* or *D* was calculated, producing null distributions for the two statistics[3]. From this the probability of observing such extreme values of *H* and *D* by chance was calculated. In Table 5 bold values indicate those that were significant at $p < 0.05$ and bold and underlined values

[3] The null distributions for *H* (Figure 5), and for *D* using the concatenated nucleotide sequence (Table 5), were generated by Gil McVean. The null distributions for *D* for the individual loci were generated by Daniel Wilson. These results have been published: K.A. Jolley, D.J. Wilson, P. Kriz, G. McVean and M.C.J. Maiden (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Molecular Biology and Evolution* **22**: 562-569.

indicate those that were significant at $p < 0.01$. When taken individually, only one of the seven loci (*pdhC*) shows consistent evidence for a departure from the standard neutral model across populations. For the global disease collection, *gdh* also shows a significant departure from the standard neutral model. When the concatenated nucleotide sequence is analysed, there is strong evidence ($p < 0.01$) for a departure from the standard neutral model in the Czech carriage study. Figure 5 shows the null distribution for *H*, with the average observed *H* in random subsets of 100 sequences from the Czech carriage study indicated with an arrow at 50. $H = 50$ was far outside the range of simulated values of *H* under the estimated parameters.

The direction of the deviation of the summary statistics from their null distributions is informative. In every case where *D* is significant it is positive, indicative of population structure. Similarly, *H* was much lower than expected, suggesting the population is more structured than would be expected under the standard neutral model. Having falsified the standard neutral model, exploring the way in which the model is deficient has revealed an excess of genetic structuring in the carriage population. The next step is to propose a refined model, fit the model and criticise it in a similar manner. In Chapter 1 various alternatives to the standard neutral model that have been proposed were discussed. However, the difficulties surrounding evolutionary inference, which were addressed using a composite likelihood approach for the standard neutral model, are exacerbated for more complex models with more parameters. Two problems exist. First, the efficiency gains made by the composite likelihood approximation are not likely to be sufficiently great to make computation feasible for models with more parameters and more complex missing data, for example the presence of hidden population structure. Second, the necessary

methodological extensions for incorporating more sophisticated models are not obvious. Amongst the problems is the development of new importance samplers for more complex models, which is not trivial. As a result, the composite likelihood approach is unlikely to feature prominently in a framework of iterative refinement of evolutionary models.

2.3 Approximate Bayesian inference

In modelling gene sequences there are two big problems. The first is that the data is discrete and high-dimensional. For n sequences of length L there are 4^{nL} possible datasets. The second is that sequences are not independent: there is a strong inter-dependency imposed by the underlying ancestral history, which is unknown. Handling the dependency structure is a difficult missing data problem, exacerbated by the fact that the missing data is a tree, which has a complex and discrete state space. In the absence of recombination there are a possible $n(n-1)/2^{n-1}$ coalescent tree topologies underlying a sample of n sequences (Hein *et al.* 2005). The problem is greater in the presence of recombination.

A naïve approach to estimating the parameters Θ of some evolutionary model M would be

Algorithm A – rejection sampling

A1. Propose Θ from some distribution $f(\Theta)$.

A2. Simulate data $\mathbf{H}'$ from the model M with parameters Θ.

A3. Accept Θ if $\mathbf{H}' = \mathbf{H}$; return to step 1.

In principal this method produces the posterior probability of the parameter given the data

$$f(\Theta \mid \mathbf{H}) = f(\mathbf{H} \mid \Theta) f(\Theta) / f(\mathbf{H}), \quad (18)$$

where $f(\mathbf{H} \mid \Theta)$ is the likelihood, that cannot be directly calculated, and $f(\Theta)$ is a prior distribution on the parameters. In a coalescent framework, step A2 is easy because data can be readily simulated. But for the first of the reasons detailed above, the acceptance probability in step 3 is essentially zero.

However, if there exist summaries of the full data **H** that contain all the information useful for inference under *M* then the state space of **H** can be massively reduced to perhaps a small number of statistics. This is the problem of statistical sufficiency. If a small number of sufficient or approximately sufficient statistics can be chosen then the acceptance probability in step 3 might no longer be negligible. Bayesian inference using summary statistics has received renewed attention in genetics recently (Tavaré *et al.* 1997; Fu and Li 1997; Weiss and von Haeseler 1998; Pritchard *et al.* 1999; Beaumont *et al.* 2002; Marjoram *et al.* 2003), and the fundamental simplicity of simulating from the model makes it an attractive option for understanding the evolution of natural populations. In addition, there are several advantages to the Bayesian methodology. Previous summary statistic methodologies proceeded by comparing the observed statistics to their distribution under a null model, which is a statistically inefficient and inflexible approach, particularly in complex genetic problems with many nuisance parameters including the unknown genealogy itself. By contrast Bayesian methods are statistically efficient, there is a natural interpretation to the posterior distribution, models can be compared quantitatively and nuisance parameters are dealt with by integration (Beaumont *et al.* 2002).

2.3.1 MCMC without likelihoods

The approach used here is based on the Markov chain Monte Carlo (MCMC, see for example O'Hagan and Forster [2004]) without likelihoods of Marjoram *et al.* (2003), with some modifications drawing mainly from the work of Beaumont *et al.* (2002). MCMC is a method for obtaining the posterior density of the parameters Θ given the data S (where S indicates that we are using a summary of the haplotypes **H**). Initially a value of Θ_0 is chosen, typically from the prior $f(\Theta)$. The following standard Metropolis-Hastings algorithm is then repeated many times

Algorithm B – Metropolis-Hastings MCMC

B1. Propose Θ' from a kernel $K(\Theta \to \Theta')$, which is usually dependent on the current state $\Theta = \Theta_i$.

B2. With probability $\alpha = \min\left\{1, \frac{f(S \mid \Theta')}{f(S \mid \Theta)} \frac{f(\Theta')}{f(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\}$ the proposal is accepted in which case let $\Theta_{i+1} = \Theta'$, otherwise $\Theta_{i+1} = \Theta_i$.

B3. Increment i by 1.

The stationary distribution of the chain is $f(\Theta \mid S)$, independently of the initial value Θ_0, although the variance in the density estimated from a finite number of iterations of the chain, which might be denoted $\hat{f}(\Theta \mid S)$, can be reduced by removing iterations from the beginning of the chain, known as the burn-in.

There is an obvious problem with performing MCMC in coalescent models: the likelihood $f(S \mid \Theta)$ is unknown. The approach of Marjoram *et al.* (2003) circumvents the need to calculate the likelihood explicitly

Algorithm C – MCMC without likelihoods

C1. Propose Θ' from a kernel $K(\Theta \to \Theta')$, where $\Theta = \Theta_i$ is the current state.

C2. Simulate data S' from the model M with parameters Θ'.

C3. If $S' = S$ then with probability $\alpha = \min\left\{1, \frac{f(\Theta')}{f(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\}$ the proposal is accepted in which case $\Theta_{i+1} = \Theta'$, otherwise $\Theta_{i+1} = \Theta_i$.

C4. Increment i by 1.

Marjoram *et al.* (2003) show that the stationary distribution of this chain is $f(\Theta \mid S)$. In step 3, if S has a continuous state space and/or is multidimensional, then S' will equal S very rarely. Thus step 3 can be re-formulated

C3. If $\partial(S', S) \le \varepsilon$ then with probability $\alpha = \min\left\{1, \frac{f(\Theta')}{f(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\}$ the proposal is accepted in which case $\Theta_{i+1} = \Theta'$, otherwise $\Theta_{i+1} = \Theta_i$.

The function $\partial(S', S)$ defines a distance between the observed and simulated data, and ε is a predetermined tolerance. The stationary distribution for this chain is $f(\Theta \mid \delta(S', S) \le \varepsilon)$, which for small ε is hopefully close to $f(\Theta \mid S)$.

The method used here makes two modifications to this scheme. The first is to up-weight the acceptance probability according to the size of $\partial(S', S)$, causing the Markov chain to spend more time closer to S. This is done by treating the distance as a random variable with some distribution that is peaked at zero. The approach is general in that any distributional form can be used. The second is to use local likelihood conditional density estimation (Loader 1996) to estimate $f(\Theta \mid S)$. The benefit of the first modification is to focus the joint density $f(S', \Theta \mid S)$ around the

observed value of S which should aid the precision of the conditional density estimation.

In summarising the data **H** with a summary S that is (almost certainly) not sufficient, an additional, artificial, layer of uncertainty is introduced. The justification for this is to facilitate inference; inference directly on **H** is too hard. Introducing a tolerance ε within which simulated values of S' are treated as equivalent to S is analogous to adding a second, artificial layer of ignorance. Ignorance, or uncertainty, is usually modelled using random variables in probability. The rectangular tolerance region $\partial(S', S) \leq \varepsilon$ is directly analogous to treating the observed summary S as though it were measured with uniform error around a true (unobserved) value X. The likelihood of the observed summary S is conditional only on X

$$f(S \mid X) \propto \begin{cases} 1 & \text{if } \delta(X, S) \leq \varepsilon \\ 0 & \text{otherwise} \end{cases}. \qquad (19)$$

Typically, in one dimension $S \sim U(X - \varepsilon, X + \varepsilon)$. This idea leads to a more general formulation of the method of Marjoram *et al.* (2003) with an arbitrary distribution for $f(S \mid X)$. Because X is unknown, it can be estimated using MCMC to obtain

$$f(X, \Theta \mid S) \propto f(S \mid X) f(X \mid \Theta) f(\Theta).$$

The following algorithm produces a Markov chain with stationary distribution $f(X, \Theta \mid S)$.

Algorithm D – Modified MCMC without likelihoods

D1. Propose Θ' from a kernel $K(\Theta \rightarrow \Theta')$, where $\Theta = \Theta_i$ is the current state.

D2. Simulate X' from the model M with parameters Θ'.

D3. With probability $\alpha = \min\left\{1, \frac{f(S \mid X')}{f(S \mid X)} \frac{f(\Theta')}{f(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\}$ the proposal is accepted in which case $(X_{i+1}, \Theta_{i+1}) = (X', \Theta')$, otherwise $(X_{i+1}, \Theta_{i+1}) = (X_i, \Theta_i)$.

D4. Increment *i* by 1.

Proof. In steps 1 and 2 a new pair (X', Θ') are proposed using the kernel $K(X, \Theta \to X', \Theta') = K(\Theta \to \Theta') K(X \to X' \mid \Theta')$, where $K(X \to X' \mid \Theta')$ is proportional to $f(X' \mid \Theta')$, which is the likelihood from the model *M* with parameters Θ'. Therefore the acceptance probability is (Metropolis 1953; Hastings 1970)

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{f(S \mid X', \Theta')}{f(S \mid X, \Theta)} \frac{f(X', \Theta')}{f(X, \Theta)} \frac{K(X', \Theta' \to X, \Theta)}{K(X, \Theta \to X', \Theta')}\right\} \\
&= \min\left\{1, \frac{f(S \mid X')}{f(S \mid X)} \frac{f(X' \mid \Theta') f(\Theta')}{f(X \mid \Theta) f(\Theta)} \frac{K(\Theta' \to \Theta) K(X' \to X \mid \Theta)}{K(\Theta \to \Theta') K(X \to X' \mid \Theta')}\right\} \\
&= \min\left\{1, \frac{f(S \mid X')}{f(S \mid X)} \frac{f(\Theta')}{f(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\}.
\end{aligned}$$

■

Any arbitrary distribution can be used to model the measurement error $f(S \mid X)$. When the uniform distribution of Equation 19 is used, the method is equivalent to that of Marjoram *et al.* (2003). The second modification to their method follows naturally having obtained a joint posterior distribution $f(X, \Theta \mid S)$. Local linear density estimation (Loader 1996) is used to estimate the conditional density $f(\Theta \mid X = S)$.

Obtaining the joint posterior $f(X, \Theta \mid S)$ might be referred to as the approximate Bayesian computation (ABC) step, and estimating $f(\Theta \mid X = S)$ might be referred to as the conditional density estimation (CDE) step. The benefit of algorithm *D* is that a

normal or double exponential distribution centred around X can be used to model the measurement error in the ABC step, so that the joint density $f(X,\Theta \mid S)$ is focused around $X = S$, which ought to improve the precision of the CDE step. There is some evidence to suggest that basing $f(S \mid X)$ on the Epanechnikov kernel would provide the most efficient estimation for $f(\Theta \mid X = S)$ (Mark Beaumont, personal communication).

2.3.2 Fitting the standard neutral model

Algorithm *D* states the method in general terms, but in any specific MCMC application the proposed moves and auxiliary variables must be designed to exploit the structure of the particular model. The primary objects of inference were the population mutation rate θ, the transition:transversion ratio κ (Kimura's [1980] two parameter model was used), and the population recombination rate ρ. In addition, the data were augmented by the genealogical tree G. The dependence structure of the model was

$$\begin{aligned} f(\mathbf{X},\theta,\kappa,\rho,G \mid S) &\propto f(S \mid \mathbf{X},\theta,\kappa,\rho,G) f(\mathbf{X},\theta,\kappa,\rho,G) \\ &= f(S \mid \mathbf{X}) f(\mathbf{X} \mid \theta,\kappa,G) f(G \mid \rho) f(\theta) f(\kappa) f(\rho) \end{aligned} \qquad (20)$$

where S are the observed summary statistics, assumed to be measured from the (unobserved) haplotypes $\mathbf{X}$ with some error given by $f(S \mid \mathbf{X})$, $f(\mathbf{X} \mid \theta,\kappa,G)$ is the likelihood of the haplotypes given by the mutation model (Kimura 1980), $f(G \mid \rho)$ is the coalescent likelihood of the genealogy (Griffiths and Marjoram 1997), and $f(\theta) f(\kappa) f(\rho)$ are the priors.

Three summary statistics were chosen by performing preliminary simulations in which the correlation between a large number of potential summary statistics and the parameters was recorded. The statistics were chosen to be orthogonal in an informal sense. That is, each statistic was chosen to be strongly correlated with one parameter, but not the other two. The benefit of choosing the statistics this way is that when an update to a single parameter is proposed, only one summary statistic is strongly affected, so the move is in a sense more local, and the acceptance probability is increased. The chosen statistics were the logarithm of the average number of pairwise differences $\log(\bar{\pi})$ which was strongly correlated with θ, the log-odds of $(\bar{\pi}_{Ts}/\bar{\pi})$, $\text{logit}(\bar{\pi}_{Ts}/\bar{\pi})$ which was strongly correlated with κ, and the correlation $\text{cor}(r^2, d)$ between a measure of LD, r^2, and physical distance, d, which was strongly correlated with ρ. $\bar{\pi}_{Ts}$ is the average number of pairwise transitions, and the transformation

$$\text{logit}(\bar{\pi}_{Ts}/\bar{\pi}) = \frac{\log(\bar{\pi}_{Ts}/\bar{\pi})}{\log(1-\bar{\pi}_{Ts}/\bar{\pi})}$$

was used to remove the correlation between $\bar{\pi}_{Ts}$ and θ. For clarity, each of these summary statistics is treated as a function of the haplotypes $\mathbf{X}$, such that $s_1(\mathbf{X}) = \log(\bar{\pi}), s_2(\mathbf{X}) = \text{logit}(\bar{\pi}_{Ts}/\bar{\pi}), s_3(\mathbf{X}) = \text{cor}(r^2, d)$, and the observed summary statistics are

$$S = (s_1(\mathbf{H}), s_2(\mathbf{H}), s_3(\mathbf{H}))$$

where $\mathbf{H}$ are the observed haplotypes. The measurement error is modelled as

$$f(S \mid \mathbf{X}) = f(S_1 \mid \mathbf{X}) f(S_2 \mid \mathbf{X}) f(S_3 \mid \mathbf{X}),$$

where

$$\begin{aligned} S_1 &\sim N(s_1(\mathbf{X}), \sigma_1) \\ S_2 &\sim N(s_2(\mathbf{X}), \sigma_2) \\ S_3 &\sim N(s_3(\mathbf{X}), \sigma_3). \end{aligned}$$

Equation 20 suggests the type of MCMC moves that might be made. Changes to θ or κ require the haplotypes **X** to be updated, but not the genealogy *G*. Changes to ρ also requires the genealogy to be updated. In principal, neither of these statements is strictly true because

$$\frac{f(\mathbf{X} \mid \theta', \kappa')}{f(\mathbf{X} \mid \theta, \kappa)}$$

and

$$\frac{f(G \mid \rho')}{f(G \mid \rho)}$$

are inexpensive to calculate, but moves of this type were not found to help mix the Markov chain. The following MCMC moves were implemented.

2.3.2.1 Update θ

The population mutation parameter is updated so that

$$\log(\theta') \sim N(\log(\theta), \zeta_1).$$

Haplotypes $\mathbf{X}'$ are then simulated from $f(\mathbf{X}' \mid \theta', \kappa, G)$, and $s(\mathbf{X}')$ is calculated. The acceptance probability is

$$\alpha = \min\left\{1, \frac{f(S \mid \mathbf{X}')}{f(S \mid \mathbf{X})} \frac{f(\theta')}{f(\theta)} \frac{K(\theta' \to \theta)}{K(\theta \to \theta')}\right\}.$$

In the implementation used for analysis, an improper prior on $\log(\theta)$ was used, so

$$\alpha = \min\left\{1, \frac{f(S \mid \mathbf{X}')}{f(S \mid \mathbf{X})}\right\},$$

and $\zeta_1 = 2$.

2.3.2.2 Update κ

The transition:transversion ratio is updated so that

$$\log(\kappa') \sim N(\log(\kappa), \zeta_2).$$

Haplotypes $\mathbf{X}'$ are then simulated from $f(\mathbf{X}' \mid \theta, \kappa', G)$, and $s(\mathbf{X}')$ is calculated. The acceptance probability is

$$\alpha = \min\left\{1, \frac{f(S \mid \mathbf{X}')}{f(S \mid \mathbf{X})} \frac{f(\kappa')}{f(\kappa)} \frac{K(\kappa' \to \kappa)}{K(\kappa \to \kappa')}\right\}.$$

In the implementation used for analysis, an improper prior on $\log(\kappa)$ was used, so

$$\alpha = \min\left\{1, \frac{f(S \mid \mathbf{X}')}{f(S \mid \mathbf{X})}\right\},$$

and $\zeta_2 = 2$.

2.3.2.3 Update ρ

A proposal distribution for ρ' of the same form as for θ' and κ' was trialled, but led to poor mixing. Instead an independence sampler was found to work well. The population recombination rate is updated so that that ρ' is drawn from the prior $f(\rho')$, which must be a proper distribution.

A new genealogy G' and haplotypes $\mathbf{X}'$ are then simulated from $f(G' \mid \rho')$ and $f(\mathbf{X}' \mid \theta, \kappa, G')$, and $s(\mathbf{X}')$ is calculated. The acceptance probability is

$$\alpha = \min\left\{1, \frac{f(S \mid \mathbf{X}')}{f(S \mid \mathbf{X})}\right\}.$$

In the implementation used for analysis, the prior for $\log(\rho) \sim U(-10, 2)$.

2.3.3 Parameter estimates

The population mutation rate, transition:transversion ratio and population recombination rates were estimated for each of the seven housekeeping loci for the Czech carriage population. The hyperparameters for the model of measurement error (σ_1, σ_2 and σ_3) was chosen by running pilot analyses. For each summary statistic, the choice of hyperparameter reflects a balance between good mixing of the Markov chain and focusing the posterior density close to $s(\mathbf{X}) = S$. Choosing a small σ_i will penalise simulated datasets $\mathbf{X}$ whose summary statistics do not closely resemble S_i, causing a tight posterior density around $s_i(\mathbf{X}) = S_i$. Concentrating the density around S_i improves precision in the CDE step. However, the Markov chain may fail to mix well, or converge at all, if the resultant acceptance probabilities are too low. On the other hand, choosing a large σ_i improves mixing because a much greater range of $s_i(\mathbf{X})$ is accepted. Too large a σ_i and the chain is essentially no longer conditioned on the data S_i. The posterior will resemble the prior, and the posterior density will not necessarily be concentrated around $s_i(\mathbf{X}) = S_i$, making the CDE step unreliable. Worse still, use of an improper prior means that the posterior cannot converge at all, and the chain resembles a random walk. In the analyses that follow, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.25$ and $\sigma_3^2 = 0.005$ were found to work, although there was some flexibility. Each Markov chain was run for 100,000 iterations.

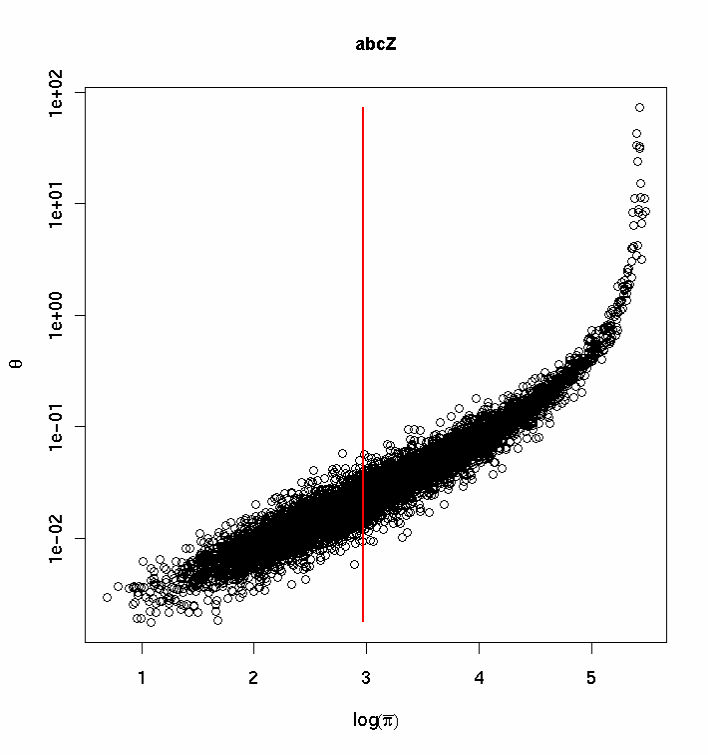


Figure 6 Joint posterior of $f(s_1(\mathbf{X}),\theta|S)$ for *abcZ*, with θ on a log scale. The red line indicates the observed value $S_1 = 2.98$. Locfit (Loader 1996) is used to estimate the conditional density of $f(\theta| s_1(\mathbf{X}) = S_1,S_2,S_3)$ along this line. See Figure 7.

Figure 6 is a scatterplot of the posterior of $f(s_1(\mathbf{X}),\theta \mid S)$ for *abcZ*. The red line indicates the observed value of the statistic S_1, which is $\log(\bar{\pi}) = 2.98$. The relationship between $s_1(\mathbf{X})$ and θ appears to be log-linear except for high values of θ, where $s_1(\mathbf{X})$ plateaus towards its maximum of $\log(L) = 6.07$ as the sequence becomes saturated with mutations. The accuracy of any method that computes the posterior $f(\theta \mid \delta(s_1(X), S_1) \le \varepsilon)$ obviously depends on the width of the tolerance ε. However, choice over ε is determined by pragmatic considerations. Conditional density estimation at $s_1(\mathbf{X}) = S_1$ is equivalent to obtaining the optimal tolerance of $\varepsilon = 0$, within the accuracy of the density estimation. Figure 7 demonstrates how conditioning on $s_1(\mathbf{X}) = S_1$ yields a much tighter posterior on θ. In the results that

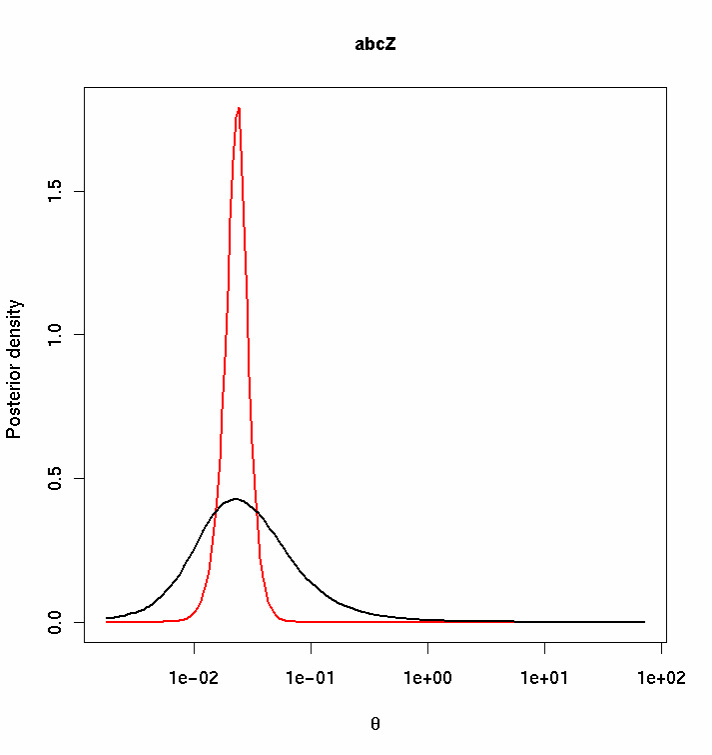


Figure 7 Black line: posterior of $f(\theta|S)$. Red line: posterior of $f(\theta|s_1(\mathbf{X})=S_1,S_2,S_3)$. Both were fit using locfit (Loader 1996). θ is on a log scale. By conditioning on the observed value S_1, a much tighter posterior is obtained.

follow, conditional density estimation is performed jointly for all summary statistics so $s_1(\mathbf{X}) = S_1, s_2(\mathbf{X}) = S_2, s_3(\mathbf{X}) = S_3$, or $s(\mathbf{X}) = S$ for short.

In Table 6 the mean and 95% highest posterior density (HPD) interval is recorded for each parameter θ, κ and ρ. The estimates of θ and ρ are on the same order of magnitude as those estimated using $\hat{\theta}_{Mc}$ and LDhat. The relative magnitude of the estimates among loci is similar, but not the same. Estimates of θ range from 0.0037 for *adk* to 0.0191 for *abcZ*. The largest estimate of $\hat{\theta}_{Mc}$ was 0.0612 for *aroE*. Although *aroE* does not have the highest point estimate, it does have the highest 95% HPD bound (0.0644). Estimates of ρ range from 0.0049 for *aroE* to 0.1686 for *fumC*.

These two loci were at the extremes of the range for the LDhat estimates. Estimates of the transition:transversion ratio κ range from 2.7 for *pgm* to 25.5 for *adk*. No estimates of κ have previously been obtained.

The 95% HPD intervals for some parameters, particularly ρ are wide, the highest upper bound being 4.97. To some extent, the width of the 95% HPD interval is related to the point estimate. Because ρ is constrained to be a positive number, it is natural that as the mean increases the upper bound increases disproportionately. Nevertheless, *adk*, *fumC* and *gdh* have especially high upper bounds (4.34, 4.97 and 1.56 respectively) compared to the point estimates and the upper bounds for the other loci. For *adk* and *gdh* this can be explained in part by the low mutation rates (estimated at 0.0037 and 0.0054 respectively) which strictly limits the information available for inference on ρ. The wide credible intervals might be a penalty for using a small number of summaries of the data for inference. However, the credible intervals cannot be compared to the confidence intervals from LDhat because none are produced. The composite likelihood curve cannot produce reliable estimates of uncertainty because by assuming independence between pairs of sites, the data are assumed to be much more informative than they really are. Obtaining meaningful credible intervals is one of the benefits of the Bayesian inference method used here. Further investigation into the summary statistics used for inferring ρ might be necessary to find a more sensitive statistic or combination of statistics. For example, Wall (2000) used H and R_m to estimate ρ in a rejection sampling setting.

Table 6 Posterior mean (and 95% HPD) for meningococcal evolutionary parameters

Locus	$\theta \times 10^3$	κ	$\rho \times 10^3$	ρ / θ	Relative rate of diversification
abcZ	19.1	18.9	43.8	2.3	57.5
	(9.1, 36.4)	(7.2, 51.3)	(3.9, 335.9)	(0.2, 21.4)	(5.1, 533.1)
adk	3.7	25.5	172.1	50.5	242.9
	(1.7, 6.7)	(2.7, 561.9)	(0.6, 4344.1)	(0.3, 1409.0)	(1.4, 6783.1)
aroE	13.3	2.9	4.9	0.4	14.0
	(2.6, 64.4)	(0.6, 19.8)	(0.3, 33.2)	(0.0, 4.8)	(0.9, 177.8)
fumC	10.1	11.8	168.6	17.4	187.3
	(5.7, 16.7)	(4.5, 37.8)	(0.7, 4970.9)	(0.1, 685.2)	(0.8, 7383.3)
gdh	5.4	16.4	50.8	9.6	75.4
	(3.0, 9.5)	(5.2, 67.9)	(0.8, 1559.8)	(0.1, 329.3)	(1.1, 2577.9)
pdhC	17.3	6.5	25.0	1.5	38.5
	(8.0, 36.8)	(3.3, 13.9)	(1.5, 222.8)	(0.1, 14.8)	(2.2, 389.6)
pgm	12.6	2.7	5.0	0.4	10.5
	(4.2, 34.0)	(1.3, 6.0)	(0.8, 59.5)	(0.0, 5.9)	(1.1, 145.9)

Also shown in Table 6 are the mean and 95% HPD intervals for the posteriors on ρ/θ. The point estimates are somewhat higher than those estimated using LDhat (Table 4) differing by a factor of 0.9 for *adk* to 27.6 for *adk*. However, this is mainly a result of the differences in the estimates of θ and ρ marginally, and not caused by the crude estimation of ρ/θ as the ratio of the marginal point estimates (Table 4). The relative rates at which recombination and mutation cause diversification are also estimated to be higher by the Bayesian method than by LDhat (Table 6). The estimates range from 10.5 for *pgm* to 242.9 for *adk*. LDhat estimated *adk* to have the second lowest relative rate (8.8). For these estimates the average tract length estimated by LDhat of 1,100 bp was used. In principal the average tract length could be estimated by the Bayesian method using the correlation between LD and a Bernoulli variable recording whether the sites are at the same or different loci.

2.3.4 Bayesian cross-validation

Any interpretation of the parameter estimates is obviously contingent upon the adequacy of the model, and there are various ways to perform model criticism in a Bayesian framework. In this section I will use the method of cross-validation to evaluate the adequacy of the standard neutral model. In Chapter 5 posterior predictive *p*-values are used for goodness-of-fit testing. However, an informal indication of the fit of a model can come directly from the Markov chain. Poor mixing can be a signal, not just of a poorly designed MCMC scheme, but also of a dataset that does not fit the model. In the context of the ABC-CDE method, difficulty in getting the posterior density concentrated around $s(\mathbf{X}) = S$ can be a symptom of a poor model fit, and this informal diagnostic can be understood from the perspective of cross-validation.

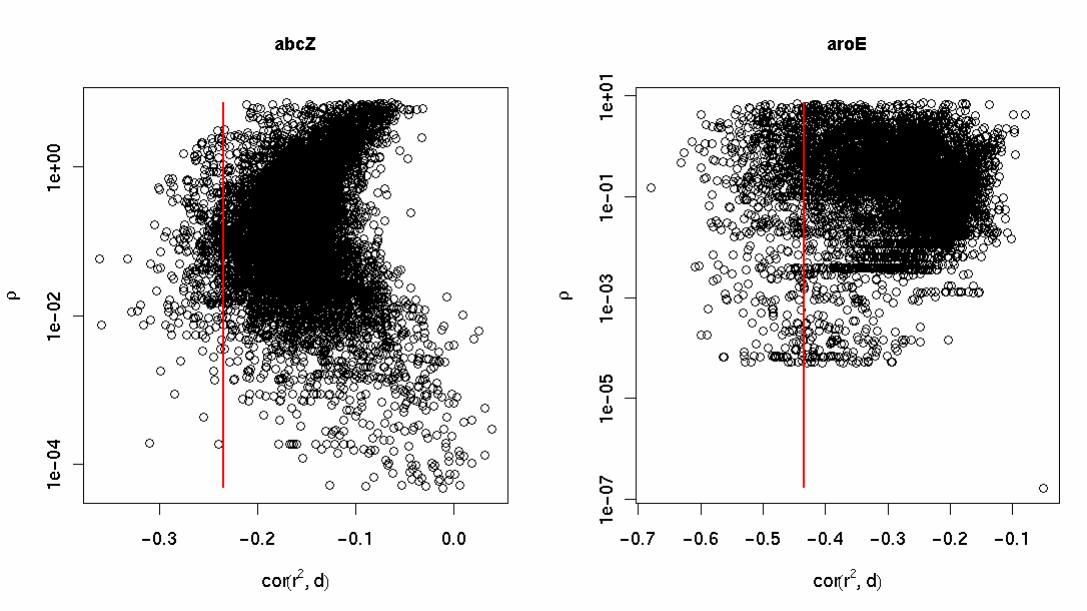


Figure 8 Scatterplots of $f(s_3(\mathbf{X}),\rho|S)$ for *abcZ* and *aroE* (ρ is on a log scale). The observed values of S_3 are marked with red lines. For *abcZ* the chain has mixed well, although the density is concentrated at smaller values of $s_3(\mathbf{X})$ than that observed. For *aroE* the observed value is yet more extreme, and the chain shows some sign of problems mixing. The quality of CDE may be affected.

In Figure 6 the posterior density $f(s_1(\mathbf{X}),\theta \mid S)$ is centred around the observed value of $S_1 = 2.98$. The region around S_1 is well-sampled, so CDE is likely to be accurate. Contrast that with Figure 8, which shows the posterior density $f(s_3(\mathbf{X}),\rho \mid S)$ for the same locus (*abcZ*, left hand graph). The density is not so well centred around $S_3 = -0.235$ (red line), although the area is probably sufficiently well-sampled for accurate CDE. However, for *adk* (right hand graph), which has an even more extreme observed value of $S_3 = -0.434$ (red line), the chain shows some sign of not mixing well, and the area around S_3 is not well-sampled. Obviously this will affect the quality of CDE. No amount of tweaking the hyperparameters σ_1, σ_2 and σ_3, or the parameters of the proposal distributions ζ_1 and ζ_2 appeared to be able to make *aroE* mix as well as

abcZ. Furthermore, the problem was confined to the plot of ρ on $s_3(\mathbf{X})$, and not the other parameter-statistic pairs. Locus *pgm* suffered similar problems to *aroE*. The problem was that datasets $\mathbf{X}$ were rarely being simulated for which $s_3(\mathbf{X})$ was as extreme as the observed value S_3. Informally speaking, this suggests that the data are not well described by the model.

Bayesian cross-validation is a formal technique for model criticism (see for example, O'Hagan and Forster 2004), and helps explain the problems seen in Figure 8. Cross-validation is based on dividing the data into two parts, one part that is used for inference (x_f) and the other part that is used for model criticism (x_c). If the model is a good fit, then x_c will be well-supported in the predictive distribution conditional on x_f. If x_c are unlikely conditional on x_f then there is a problem. The predictive distribution of x_c given x_f is

$$f(x_c \mid x_f) = \int f(x_c \mid x_f, \Theta) f(\Theta \mid x_f) \mathrm{d}\Theta . \qquad (21)$$

In ABC, the data can be partitioned into summary statistics used for inference and summary statistics used for model criticism. To address the problems noted in Figure 8, S_1 and S_2 were used for inference and S_3 for model criticism.

Table 7 Cross-validation for standard neutral model

Locus	p^*	p
abcZ	0.004	0.005
adk	0.120	0.116
aroE	0.000	0.000
fumC	0.683	0.687
gdh	0.018	0.015
pdhC	0.001	0.002
pgm	0.000	0.000

Modifying the MCMC scheme to perform cross-validation is straightforward by removing the conditioning on the statistic(s) in question, in this case S_3. Because the dimensionality of the data is also reduced, the chain takes less time to run. As a diagnostic of model fit, the predictive probability of observing $s_3(\mathbf{X})$ as extreme as S_3 was calculated as

$$p^* = \int_{-\infty}^{S_3} f\left(s_3(\mathbf{X}) = u \mid S_1, S_2\right) \mathrm{d}u, \tag{22}$$

$$p = \int_{-\infty}^{S_3} f\left(s_3(\mathbf{X}) = u \mid s_1(\mathbf{X}) = S_1, s_2(\mathbf{X}) = S_2\right) \mathrm{d}u, \tag{23}$$

and the *p*-values were made two-tailed in the usual way. p^* can be estimated directly from the Markov chain, without CDE, simply as the number of times $s_3(\mathbf{X})$ was as extreme or more extreme as S_3. Equation 23 requires CDE using locfit (Loader 1996). In practice, p^* and p are almost identical. The results are shown in Table 7. Cross validation suggests that there are problems with the model. The predictive probability

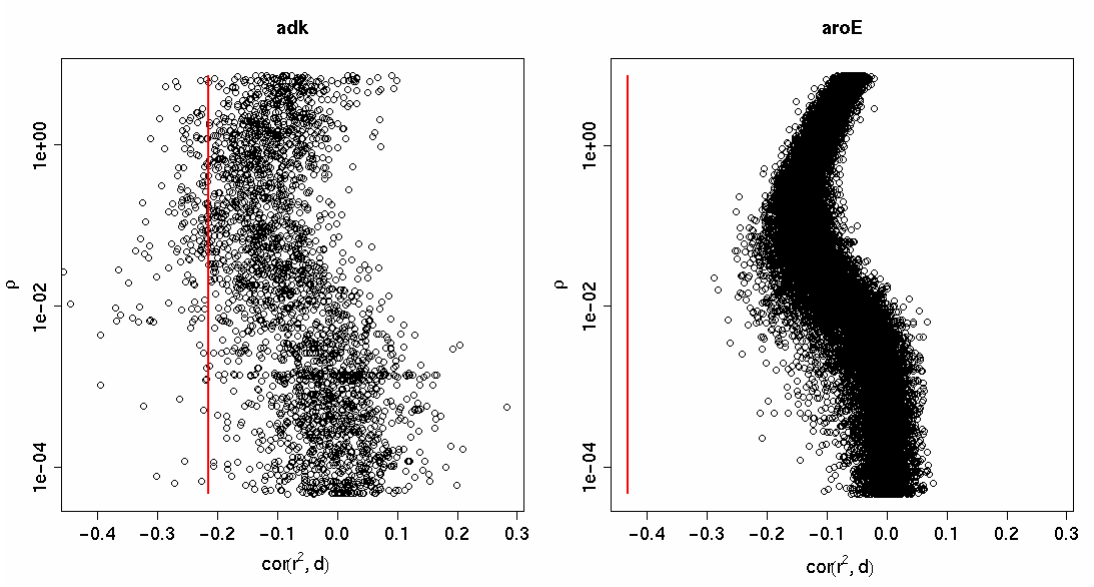


Figure 9 Cross-validation reveals discrepancies between the observed $S_3 = \mathrm{cor}(r^2,d)$ and that predicted by a model fit using S_1 and S_2. For loci *adk* and *aroE*, $f(s_3(\mathbf{X}),\rho|S_1,S_2)$ is plotted, with the observed value of S_3 indicated by the red line. For *adk*, $p^* = 0.120$, whereas for *aroE*, $p^* = 0.000$ (see Table 7 and text).

of S_3 given S_1 and S_2 is less than 0.05 for all but two of the loci (*adk* and *fumC*). What this means is that for the inferred values of θ and κ (about which S_1 and S_2 are informative), the model rarely predicts values of S_3 as extreme as observed, when the values of ρ are taken from the flat prior which was $U(-10,2)$ on $\log(\rho)$. This is illustrated by Figure 9, which shows $f(s_3(\mathbf{X}),\rho \mid S_1,S_2)$ for *adk* and *aroE*. The red line indicates the observed value of S_3, which is within the range of $s_3(\mathbf{X})$ sampled for *adk*, but well outside the range sampled for *aroE*. Because values of ρ are taken from the prior, the prior will have an important effect on the conclusions of cross-validation. However, for *aroE* it is clear that no choice of prior would have changed the conclusion that the model does not predict values of S_3 as extreme as observed.

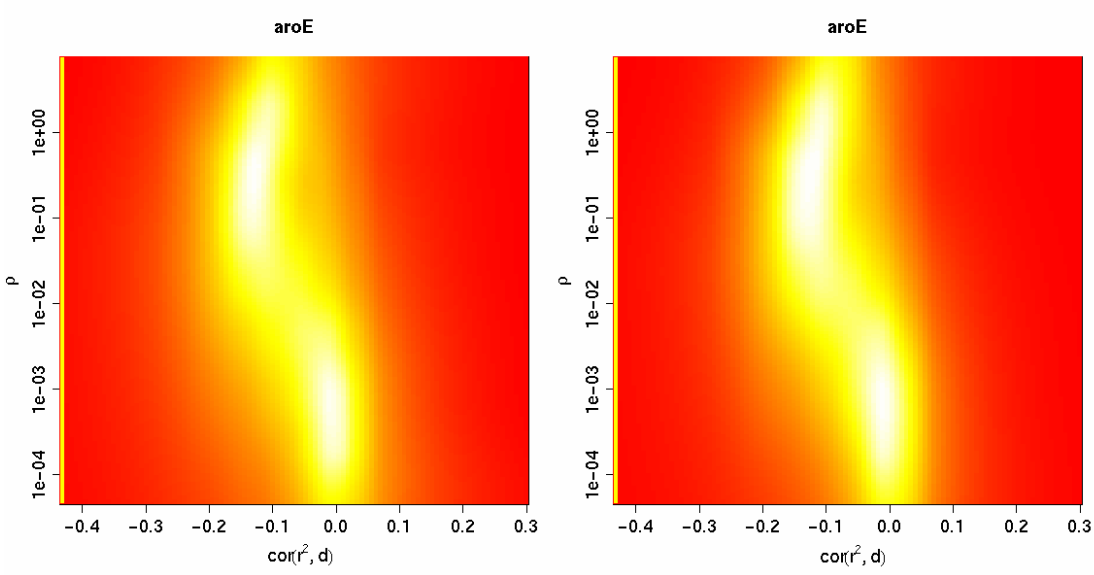


Figure 10 Locfit has been used to estimate $f(s_3(\mathbf{X}),\rho|S_1,S_2)$ and $f(s_3(\mathbf{X}),\rho|\ s_1(\mathbf{X})=S_1,\ s_2(\mathbf{X})=S_2)$ (left and right images respectively) for *aroE*. Note that ρ is on a log scale. More intense colours (closer to white) indicate higher posterior density. The observed value of S_3 is indicated with a yellow line on the far left of each image.

Owing to the fact that S_1 and S_2 were chosen to be informative about θ and κ, but not ρ, conditioning on $s_1(\mathbf{X})=S_1$ and $s_2(\mathbf{X})=S_2$ using CDE barely alters the joint posterior of $s_3(\mathbf{X})$ and ρ. This is illustrated by Figure 10, in which locfit has been used to estimate the joint posterior of $s_3(\mathbf{X})$ and ρ marginal to $s_1(\mathbf{X})$ and $s_2(\mathbf{X})$ (left hand image) and conditional upon $s_1(\mathbf{X})=S_1$ and $s_2(\mathbf{X})=S_2$ (right hand image) for *aroE*. In each image, the observed value of S_3 is indicated with a yellow line, which falls well outside the density in both cases. This observation is reflected in the fact that p^* and p are almost identical for all loci (Table 7). Figures 9 and 10 illustrate the ability of locfit to estimate joint densities. The left hand image in Figure 10 is a density estimate of the right hand image in Figure 9. In the density plot, more intense colours (closer to white) correspond to higher posterior density. The density plot (Figure 10, left) is more informative than the scatterplot (Figure 9, right), because in

areas of high density the scatterplot becomes saturated, whereas the density plot does not.

Cross-validation of S_3 reveals the reason for the difficulties mixing the MCMC chain for *aroE* and *pgm*. When there is little support for S_3 in the predictive distribution of $s_3(\mathbf{X})$, the distance between S_3 and $s_3(\mathbf{X}')$ for simulated datasets $\mathbf{X}'$ will be large, and as a result the acceptance probability small. Therefore it will be more difficult to perform inference on datasets that are poorly described by the model because of lower acceptance probabilities in the Markov chain. If the adequacy of the model is in question (which surely it is for any basic model), then the primary purpose of estimating the model parameters is to perform goodness-of-fit testing. Biologically meaningful interpretation of the parameters is contingent upon the adequacy of the model, and if it can be shown that the model is a bad fit, then the utility of parameter estimates *per se* is diminished. If there is difficulty in getting the Markov chain to mix, particularly if a single summary statistic is affected, then cross-validation is a useful method of model criticism because it may reveal that the predictive distribution of the observed statistic is not well supported by the model. The advantage of cross-validation is that it is a formal model criticism technique, in contrast to the informal observation of poor mixing which might be symptomatic of a number of underlying problems.

2.4 Refining the model

Regardless of the method of inference, be it composite likelihood or approximate Bayesian computation, the central conclusion that patterns of meningococcal genetic

diversity cannot be explained by the standard neutral model is unaffected. Whilst that conclusion does not have to question the validity of the coalescent as the basic starting point for evolutionary inference, it does mean that for understanding meningococcal evolution, a refinement to the coalescent is required.

Model criticism techniques have revealed that there is an excess of genetic structuring in meningococcal populations. The observed number of sequence types (STs) is too high for the estimated rate of recombination, and there appears to be a dearth of low frequency allelic variants, indicative of long-term population subdivision. The correlation between LD and physical distance is too strong for five of the seven housekeeping loci studied, implying that LD decays more deterministically than expected under the standard neutral model. This may also reflect the existence of population structure (Pritchard and Przeworski 2001). Together, these results suggest that any refinement to the standard neutral model must incorporate some degree of population structuring, but the exact formulation of that structure, and the cause, is unclear. For that reason, the next step is to propose a revised model, fit the model, and criticise it.

A process of iterative refinement of the evolutionary model is, in my opinion, essential to furthering the understanding of meningococci population biology. The coalescent provides a common thread for refinement of the model. In the next chapter, I will fit the neutral microepidemic model of Fraser *et al.* (2005) using a modification to the coalescent. The conclusions are somewhat different to those found by fitting a multinomial distribution to the observed allele frequencies (Fraser *et al.* 2005). The importance of geographic structuring and the relationship between carried and

disease-causing populations of meningococci is also examined using a variety of statistic models. Together these suggest what the next refinement to a coalescent model of meningococcal evolution might be.

Chapter 3

Genetic structuring in *Neisseria meningitidis*

Meningococcal populations exhibit greater genetic structuring than is expected under a standard neutral model of evolution, as can be seen using the inference and model criticism techniques of Chapter 2. In this chapter I investigate the nature of genetic structuring in meningococci. I begin by using approximate Bayesian computation with conditional density estimation (ABC-CDE) to criticise the neutral microepidemic model of Fraser *et al.* (2005), in which meningococci evolve according to the standard neutral model, but structuring is imposed by biased sampling. I then use analysis of molecular variance (AMOVA) and Mantel tests to quantify the extent of geographic structuring in meningococcal populations sampled from within the same country and between different countries. The role of host age in structuring carriage populations is investigated and patterns of genetic diversity are compared between school and military institutions in Bavaria, Germany. I use the same techniques to compare patterns of genetic diversity in disease-causing and carried meningococci, and assess the extent of overlap between these populations.

3.1 Neutral microepidemic model

As discussed in Chapter 1 (section 1.2.4), the neutral microepidemic model is used (Fraser *et al.* 2005) to explain the observed excess of homozygosity in the Czech carriage study (Jolley *et al.* 2000). Homozygosity can be calculated as the proportion of pairs of isolates that are identical at all seven MLST loci. Figure 1 shows the allelic

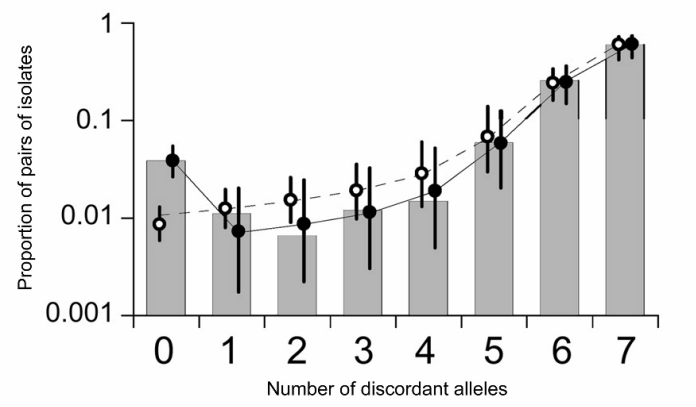


Figure 1 Allelic mismatch distribution for Czech carriage study (grey bars). The horizontal axis shows the number of loci at which a pair of isolates can differ (up to 7 for MLST), and the vertical axis the proportion of pairs that differ at that number of loci. Open circles show the fit under the standard neutral model, and the filled circles show the fit under the neutral microepidemic model. Source: Fraser *et al.* (2005).

mismatch distribution in the Czech carriage study, where each bar represents the number of pairs of isolates that differ at the stated number of loci. Individuals that differ at none of the seven loci are said to be homozygous.

In the neutral microepidemic model, biased sampling causes an excess of homozygosity. The population is thought to be made up of many microepidemics, which comprise short transmission chains within the host population. In the model n_c of these microepidemics are repeatedly sampled, contributing an average of $\bar{\sigma}$ isolates each. The total number of isolates, n_o say, that come from over-sampled microepidemics is modelled as a Poisson random variable with parameter $n_c\bar{\sigma}$, but truncated at n, the total sample size, and the joint distribution of the number of isolates sampled from each microepidemic conditional on n_o and n_c is symmetric multinomial.

Fraser *et al.* (2005) use maximum likelihood to fit the allelic mismatch distribution to a multinomial distribution with parameters $(p_0, p_1, \ldots, p_7)$ where p_i is the expected proportion of pairs of isolates differing at *i* loci under the infinite alleles model in a standard neutral population (Kimura 1968). The p_i are a function of the per-locus mutation rate θ and between-locus recombination rate ρ. The open circles in Figure 1 show the fit of the standard neutral model. To fit the neutral microepidemic model (filled circles, Figure 1), Fraser *et al.* (2005) matched p_0 exactly to the observed homozygosity using an extra free parameter h_e, which represents the excess homozygosity. To estimate n_c and $\bar{\sigma}$, they conducted simulations using $\hat{\theta}$, $\hat{\rho}$ and $\hat{h}_e$; n_c and $\bar{\sigma}$ were resolved by matching the observed and expected number of STs, subject to the constraint that $h_e = n_c\bar{\sigma}^2/(n(n-1))$ (Christophe Fraser, personal communication). This constraint arises by considering that for each cluster of size σ there are an additional $\sigma(\sigma-1)/2$ identical pairs of isolates, so there are approximately $n_c\bar{\sigma}^2/2$ extra identical pairs in total. The mutation rate, recombination rate, number of clusters and average cluster size were estimated to be $\theta = 10.2$, $\rho = 13.6$, $n_c = 9$ and $\bar{\sigma} = 13.1$ respectively.

Despite the advantages of fitting an explicit statistical model, there are a number of difficulties with the analysis. As discussed in Chapter 1, the frequencies of each class in the mismatch distribution are not independent, so the multinomial distribution is inappropriate. Using only the mismatch distribution discards a great deal of information about patterns of genetic diversity in the bacterial population. Indeed, the allele numbers of seven MLST loci probably contain much less information than the electromorph numbers of 20 MLEE loci. The infinite alleles model is also not

appropriate to apply to the MLST loci because, as was shown in Chapter 2, recombination causes diversification within a locus approximately ten times faster than mutation. Therefore mutation will be credited with causing much of the diversification that is due to recombination. I used ABC-CDE to fit a coalescent formulation of the neutral microepidemic model to each locus individually, and the adequacy of the model was assessed, by cross-validation, using the same summaries of nucleotide diversity that were used for the standard neutral model in section 2.3.4.

3.1.1 Coalescent formulation of the microepidemic model

To conduct simulations for ABC-CDE, a coalescent version of the neutral microepidemic model was formulated. The model is essentially a standard neutral coalescent, with over-sampling at the tips. Simulations are performed as follows.

1. Draw n_o from a Poisson distribution with parameter $n_c\bar{\sigma}$ truncated at n.
2. Draw the sizes of the microepidemics, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_{n_c})$ from a symmetric multinomial distribution, where $\sum_i \sigma_i = n_o$.
3. Simulate a standard neutral coalescent genealogy for a sample of size $n - n_o + n_c$, and superimpose mutations as described in section 2.2.3.
4. Of those sequences, choose n_c uniformly at random without replacement to form the microepidemic clusters. For each microepidemic i include the chosen sequence σ_i times in the final sample.
5. The remaining $n - n_o$ sequences are included once each in the final sample.

3.1.2 Approximate Bayesian inference

To fit the neutral microepidemic model, ABC-CDE inference was performed on the mutation rate θ, transition:transversion ratio κ and recombination rate ρ using Kimura's (1980) two-parameter mutation model and the coalescent microepidemic model described in the previous section. The number n_c and average size $\bar{\sigma}$ of the microepidemics were not estimated; instead the values of $n_c = 9$ and $\bar{\sigma} = 13.1$ were taken from Fraser *et al.* (2005). To estimate the parameters, the same three statistics as used for the standard neutral model (section 2.3.2) were used: $\log(\bar{\pi})$, $\text{logit}(\bar{\pi}_{Ts} / \bar{\pi})$ and $\text{cor}(r^2, d)$. The Markov chains showed signs of problems mixing, indicative of model misspecification. As a result, cross-validation was performed (see section 2.3.4) by obtaining the predictive distribution of $\text{cor}(r^2, d)$ conditional upon $\log(\bar{\pi})$ and $\text{logit}(\bar{\pi}_{Ts} / \bar{\pi})$.

The results are shown in Table 1 for each of the two-tailed cross-validation *p*-values p^* and p (Equations 22 and 23, section 2.3.4). The two *p*-values are almost identical, and the results are similar to those for the standard neutral model. The predictive probability of $\text{cor}(r^2, d)$ conditional on $\log(\bar{\pi})$ and $\text{logit}(\bar{\pi}_{Ts} / \bar{\pi})$ is less than 0.05 for five of the seven loci. Of the other two, $p = 0.241$ for *fumC*, and $p = 0.051$ for *adk*, which is marginal. For these two loci, the *p*-values are considerably lower under the neutral microepidemic model than under the standard neutral model (Chapter 2, Table 7).

Table 1 Cross-validation for microepidemic model

Locus	p^*	p
abcZ	0.004	0.003
adk	0.051	0.051
aroE	0.000	0.000
fumC	0.241	0.240
gdh	0.019	0.019
pdhC	0.003	0.003
pgm	0.000	0.000

In summary, the neutral microepidemic model does not appear to provide a better fit to the data than the standard neutral model; if anything it is worse. Whilst the parameters n_c and $\overline{\sigma}$ were not estimated here, it seems unlikely that the microepidemic model can explain the patterns of genetic diversity observed in meningococcal populations. One might be tempted to think that reducing complex multilocus nucleotide sequence data to an allelic mismatch distribution forfeits the power to reject the model. Much information is discarded by reducing the sequence data to an allelic mismatch distribution, but the microepidemic model was rejected here using only three statistics. Arguably, the allelic mismatch distribution contains less information because the frequencies of the mismatch classes are highly correlated, but regardless of the number of statistics used for inference, thorough model criticism is essential for learning about the evolutionary history of the population. Genetic structuring in meningococci cannot be explained by a simple model of sampling bias, and in the rest of this chapter I evaluate the extent of

geographic structuring within and between European countries, the role of host age-structure and differences in the composition of meningococcal populations between schools and military institutions within Germany, and the extent of overlap between European populations of disease-causing and carried meningococci. In the next section I describe the methods used for these analyses, analysis of molecular variation (AMOVA) and the Mantel test.

3.2 Analysing population structure

Standard methods are used in the rest of this chapter for analysing population structure: analysis of molecular variance (AMOVA; Excoffier *et al.* 1992), the Mantel test (Mantel 1967) and logistic regression. I will describe AMOVA and the Mantel test, and make a straightforward extension to AMOVA to allow a two-way design which is used later (section 3.5) for resolving the effects of geography and propensity to cause disease.

3.2.1 Analysis of molecular variance

Introduced by Excoffier *et al.* (1992), AMOVA uses the number of pairwise differences between isolates to define Euclidean distance, on which analysis of variance (ANOVA; Fisher 1925) is performed. Using AMOVA the following random effects model is fitted

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \mathbf{A}_i + \boldsymbol{\varepsilon}_{ij}, \qquad (1)$$

where $\mathbf{Y}_{ij}$ is the jth haplotype from population i, $\boldsymbol{\mu} + \mathbf{A}_i$ is a notional mean haplotype for population i, $\mathbf{A}_i$ being a realisation of a random variable $\mathbf{A}$ whose expectation is

zero and variance is σ_A^2, and $\boldsymbol{\varepsilon}_{ij}$ is the deviation of the *j*th gene sequence from the mean haplotype for population *i*, such that $\boldsymbol{\varepsilon}_{ij}$ is a realisation of an independent random variable $\boldsymbol{\varepsilon}$ whose expectation is zero and variance is σ_E^2. The model parameters are the variance components σ_A^2 and σ_E^2, which are estimated by decomposing the total genetic variance in the population into that which is due to differences between populations and the residual within-population variance. F_{ST}, which is a function of the variance components, is interpreted as the correlation of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the whole species.

$$F_{ST} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \tag{2}$$

An explicit assumption of the model is that σ_E^2 is the same for all populations, so F_{ST} represents an average over populations.

As in ANOVA, variance is measured in terms of sums of squares. Denote SS_T the total sum of squares, SS_A the sum of squares between populations and SS_E the residual sum of squares, or the sum of squares within populations.

$$SS_T = SS_A + SS_E$$

The sums of squares would normally be computed as

$$SS_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{i\cdot} \right)^2 \tag{3}$$

and

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{\cdot\cdot} \right)^2 \tag{4}$$

where *k* is the number of populations, n_i is the sample size of population *i*, $\overline{\mathbf{Y}}_{i\cdot}$ is the mean haplotype in population *i* and $\overline{\mathbf{Y}}_{\cdot\cdot}$ is the mean haplotype for the total population.

For genetic data that is discrete and multidimensional, it is not immediately obvious how to define an average. However, Equations 3 and 4 can be rewritten so that the sums of squares are defined by the distance between pairs of haplotypes

$$SS_E = \sum_{i=1}^{k} \frac{1}{2n_i} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \left(\mathbf{Y}_{ij} - \mathbf{Y}_{ij'} \right)^2 \tag{5}$$

and

$$SS_T = \frac{1}{2n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \sum_{i'=1}^{k} \sum_{j'=1}^{n_{i'}} \left(\mathbf{Y}_{ij} - \mathbf{Y}_{i'j'} \right)^2 , \tag{6}$$

where n is the total sample size. A natural measure of genetic distance is the number of nucleotides that differ between a pair of sequences, π. As in a standard random effects ANOVA (see for example Sokal and Rohlf 1995), the variance components are estimated using the method of moments

$$\begin{aligned} E(MS_A) &= \frac{\sigma_A^2}{k-1} \left(n - \frac{1}{n} \sum_{i=1}^{k} n_i^2 \right) + \sigma_E^2 \\ E(MS_E) &= \sigma_E^2, \end{aligned} \tag{7}$$

where *MS* is the mean square and (in Equation 8, below) *DF* are the degrees of freedom. Because the variance components are estimated by the method of moments, it is possible to obtain negative F_{ST} in the absence of population structure (Excoffier *et al.* 1992). The test statistic for determining the significance of population differentiation is

$$F = \frac{MS_A}{MS_E} = \frac{SS_A DF_E}{SS_E DF_A}, \tag{8}$$

where $DF_A = k - 1$ and $DF_E = n - k$. The null distribution of F is not the ratio of two chi-squared random variables, but is determined by permutation of haplotypes amongst the populations. In all analyses of molecular variance in this chapter, 1,000 permutations were used to obtain the *p*-value. AMOVA is implemented in Arlequin version 2.000 (Schneider *et al.* 2000).

3.2.1.1 Two-way AMOVA

Here I make a straightforward extension to AMOVA to allow for two crossed factors in an unbalanced design (i.e. different sample sizes for each combination of factors; see for example McCullagh and Nelder [1989]). In section 3.5 the two factors will be country and host disease status. The extended model can be written

$$\mathbf{Y}_{ijk} = \boldsymbol{\mu} + \mathbf{A}_i + \mathbf{B}_j + \boldsymbol{\varepsilon}_{ijk}, \tag{9}$$

where $\mathbf{Y}_{ijk}$ is the kth haplotypes from country i and disease status j, $\mathbf{A}_i$ is the random effect of country i as before, $\mathbf{B}_j$ is the independent random effect of disease status j, which is a realisation of a random variable $\mathbf{B}$ whose expectation is zero and variance is σ_B^2, and $\boldsymbol{\varepsilon}_{ijk}$ is the deviation of the kth sequence from the mean haplotype for that country and disease status, $\boldsymbol{\mu} + \mathbf{A}_i + \mathbf{B}_j$, which is a realisation of an independent random variable $\boldsymbol{\varepsilon}$ whose expectation is zero and variance is σ_E^2.

The total sum of squares becomes

$$SS_T = SS_A + SS_B + SS_E,$$

where SS_A is the sum of squares between countries, SS_B is the sum of squares for disease status and SS_E is the residual sum of squares. The sums of squares are calculated as

$$SS_T = \frac{1}{2n}\sum_{i=1}^{k_A}\sum_{j=1}^{k_B}\sum_{k=1}^{n_{ij}}\sum_{i'=1}^{k_A}\sum_{j'=1}^{k_B}\sum_{k'=1}^{n_{i'j'}}\left(\mathbf{Y}_{ijk} - \mathbf{Y}_{i'j'k'}\right)^2 \tag{10}$$

$$SS_A = SS_T - \sum_{i=1}^{k_A}\frac{1}{2n_i}\sum_{j=1}^{k_B}\sum_{k=1}^{n_{ij}}\sum_{j'=1}^{k_B}\sum_{k'=1}^{n_{ij'}}\left(\mathbf{Y}_{ijk} - \mathbf{Y}_{ij'k'}\right)^2, \tag{11}$$

and $$SS_E = \sum_{i=1}^{k_A}\sum_{j=1}^{k_B}\frac{1}{2n_{ij}}\sum_{k=1}^{n_{ij}}\sum_{k'=1}^{n_{ij}}\left(\mathbf{Y}_{ijk} - \mathbf{Y}_{ijk'}\right)^2, \quad (12)$$

where k_A is the number of countries, k_B is the number of disease statuses, $n_i = \sum_j n_{ij}$ and n_{ij} is the sample size of country i, disease status j. The expected mean squares are

$$E(MS_A) = \sigma_E^2 + z_A\sigma_A^2,$$
$$E(MS_B) = \sigma_E^2 + z_B\sigma_B^2, \quad (13)$$

and $$E(MS_E) = \sigma_E^2,$$

where $$z_A = \frac{1}{k_A - 1}\left(n - \sum_{j=1}^{k_B}\frac{\sum_{i=1}^{k_A} n_{ij}^2}{\sum_{i=1}^{k_A} n_{ij}}\right), \quad (14)$$

and $$z_B = \frac{1}{k_B - 1}\left(n - \sum_{i=1}^{k_A}\frac{\sum_{j=1}^{k_B} n_{ij}^2}{\sum_{j=1}^{k_B} n_{ij}}\right), \quad (15)$$

from which the variance components can be estimated by the method of moments.

For each effect, country and disease status, F_{ST} can be calculated so that

$$F_{ST}^{(A)} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_E^2}$$
$$F_{ST}^{(B)} = \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2 + \sigma_E^2}, \quad (16)$$

which provides a natural way to interpret the parameters. Significance testing is performed for each random effect using the test statistics

$$F^{(A)} = \frac{MS_A}{MS_E} = \frac{SS_A DF_E}{SS_E DF_A}$$
$$F^{(B)} = \frac{MS_B}{MS_E} = \frac{SS_B DF_E}{SS_E DF_B}, \quad (17)$$

where the null distribution for $F^{(A)}$ is found by permuting haplotypes amongst countries but not disease status and vice versa for $F^{(B)}$. $DF_A = k_A - 1$, $DF_B = k_B - 1$ and $DF_E = n - k_A - k_B + 1$. To implement the two-way AMOVA I wrote pilot program in Maple and then re-wrote it in C++ for the full analyses.

3.2.2 Mantel test

The Mantel test (Mantel 1967) is a test for correlation between two distance matrices, say **A** and **B**. In this chapter, the two distances will be geographic distance and pairwise genetic distance. The test statistic is the correlation coefficient

$$r = \frac{\mathrm{cov}(\mathbf{A},\mathbf{B})}{\sqrt{\mathrm{var}(\mathbf{A})\mathrm{var}(\mathbf{B})}} = \frac{n^2\sum_{i,j} a_{ij}b_{ij} - \sum_{i,j} a_{ij}\sum_{i,j} b_{ij}}{\sqrt{\left[n^2\sum_{i,j} a_{ij}^2 - \left(\sum_{i,j} a_{ij}\right)^2\right]\left[n^2\sum_{i,j} b_{ij}^2 - \left(\sum_{i,j} b_{ij}\right)^2\right]}}. \quad (18)$$

Under the null hypothesis, $r = 0$. A null distribution for the test statistic is obtained by permuting haplotypes amongst the populations. The sample size of each population remains constant. During permutation, the only part of Equation 18 that changes is

$$\sum_{i,j} a_{ij}b_{ij}\,, \quad (19)$$

allowing faster computation. I implemented the Mantel test in a C++ program.

3.3 Geographic structuring in Europe

In this section I use AMOVA to test for significant differentiation in populations of carried meningococci, firstly between towns in the Czech Republic (Jolley *et al.*

Figure 2 Map of the Czech Republic. The sampling locations of the Czech carriage study (Jolley *et al.* 2000) are indicated in red.

2000), and then between the European countries of the Czech Republic, Greece and Norway (Yazdankhah *et al.* 2004).

3.3.1 Structuring within the Czech Republic

Figure 2 shows the locations in the Czech Republic from which 217 isolates were collected from healthy carriers in 1993 (Jolley *et al.* 2000): Prague (2 isolates), České Budejovice (87), Hradec Králové (3), Kutna Hora (1), Plzeň (56), Olomouc (64) and Opava (3). Of these sampling locations, some have a very small number of sequences. It was necessary to pool some locations to obtain sufficient power to detect population

Table 2 Partitions of the Czech carriage study used for AMOVA

Bipartite		
	Region A	Prague, Plzeň, Hradec Králové, České Budejovice and Kutna Hora (149 isolates)
	Region B	Olomouc and Opava (67 isolates)
Tripartite		
	Region A	Prague, Plzeň and Hradec Králové (61 isolates)
	Region B	České Budejovice and Kutna Hora (88 isolates)
	Region C	Olomouc and Opava (67 isolates)
Quadripartite		
	Region A	Prague and Hradec Králové (5 isolates)
	Region B	České Budejovice and Kutna Hora (88 isolates)
	Region C	Olomouc and Opava (67 isolates)
	Region D	Plzeň (56 isolates)

Table 3 Evidence for population structure in the Czech carriage study

Locus	Bipartite		Tripartite		Quadripartite		Septempartite	
	F_{ST}	p	F_{ST}	p	F_{ST}	p	F_{ST}	p
abcZ	-0.005	0.685	0.000	0.434	0.016	0.057	0.006	0.232
adk	0.010	0.104	0.004	0.249	0.015	0.051	0.009	0.217
aroE	0.000	0.381	0.007	0.167	0.007	0.209	0.006	0.279
fumC	0.005	0.193	0.017	0.014	**0.023**	**0.006**	0.019	0.052
gdh	0.004	0.239	0.005	0.202	0.008	0.173	0.001	0.415
pdhC	-0.002	0.502	0.002	0.322	0.010	0.109	0.002	0.379
pgm	0.006	0.138	0.002	0.343	0.018	0.028	0.011	0.132

Enboldened entries indicate significance at $p < 0.0073$.

subdivision. From the map there is no obvious way to partition the locations, so a number of divisions were analysed, shown in Table 2. Note that the sampling location was unknown for one of the isolates, which is left out of these analyses.

Analysis was performed on each partition of the data, yielding an average F_{ST} and p-value for each partition. In the septempartite analysis each location was taken as a separate population, despite the small sample sizes. The results are shown in Table 3. To correct for multiple comparisons amongst the seven loci the Bonferroni correction was applied so that p-values less than or equal to

$$\alpha = 1 - \exp\{\log(1 - 0.05)/n\} \tag{20}$$

were taken as evidence for significant structuring. For $n = 7$ loci, $\alpha = 0.0073$. Bonferroni is conservative, especially since the loci may not give independent accounts of the population structure due to correlation between the genealogies for

different loci. However, the *p*-values in Table 3 show that the conclusions would be little affected even if the Bonferroni correction were not applied.

Whichever way the sampling locations are partitioned there is no convincing evidence for population subdivision. The only significant *p*-value at $\alpha = 0.0073$ occurred at the *fumC* locus for the quadripartite division of the sampling locations. This gave the maximum value of F_{ST} for any locus for any partition (0.023). The minimum F_{ST} was -0.005, which occurs because the variance components are estimated by the method of moments (Equation 7). This value is best interpreted as $F_{ST} = 0$.

3.3.2 Differentiation between European countries

From Figure 2, the greatest distance between sampling locations in the Czech carriage study was 200 miles (Plzeň to Opava), suggesting that on this sort of scale mixing of meningococci occurs sufficiently quickly to eradicate any signal of population structure. In this section I use AMOVA to investigate the degree of differentiation between carried meningococci sampled from three different European countries: the Czech Republic, Greece and Norway (Yazdankhah *et al.* 2004). Norway is separated from the Czech Republic by 694 miles, including the Baltic Sea, and the Czech Republic is separated from Greece by 953 miles (see Figure 3; distances measured from capital to capital).

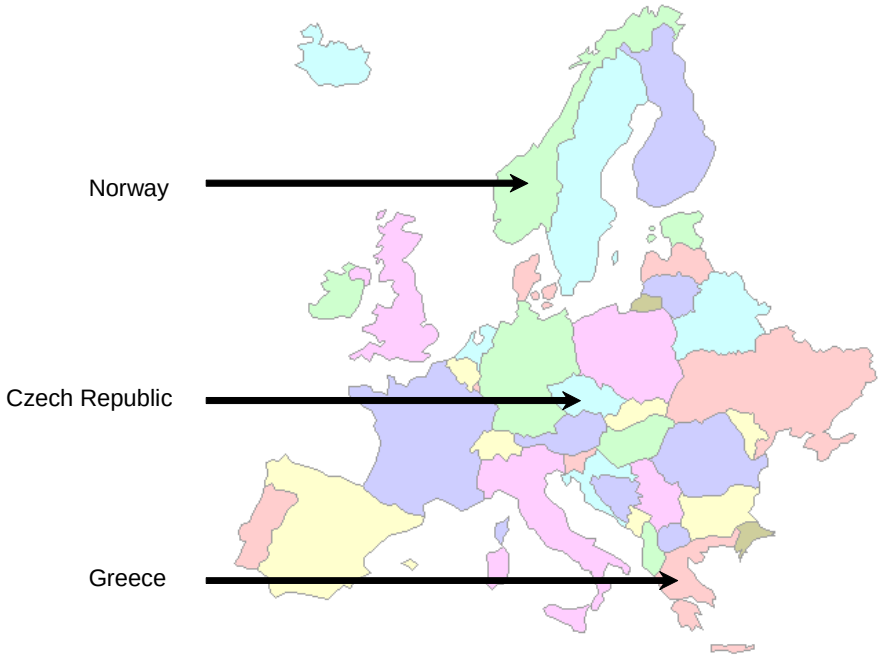


Figure 3 Map of Europe showing the location of the three countries sampled by Yazdankhah *et al.* (2004): Norway, Czech Republic and Greece. Distances between countries were measured from capital to capital (http://www.wcrl.ars.usda.gov/cec/java/lat-long.htm).

A total of 353 isolates were sampled from healthy carriers: 112 from the Czech Republic in 1994 and 1996, 88 from Greece in 1999, and 153 from Norway in 1991 and 1996. For the analysis, the isolates were grouped simply according to the country of origin. Table 4 shows the results. For each locus AMOVA was used to calculate an average F_{ST} between the three countries and a *p*-value. Pairwise F_{ST} is also calculated for each pair of countries. There was strong evidence for genetic differentiation between the carried meningococci in the three countries ($p < 0.0005$ for all loci), with F_{ST} ranging from 0.022 for *pgm* up to 0.071 for *aroE*. The lowest value of $F_{ST} = 0.022$ was higher than any non-significant F_{ST} for sampling locations within the Czech Republic (Table 3).

Table 4 Evidence for differentiation between carried meningococci in Europe

Locus	F_{ST}	p	pairwise F_{ST}		
			CR *vs* G	CR *vs* N	G *vs* N
abcZ	0.024	< 0.0005	0.040	0.012	0.026
adk	0.042	< 0.0005	0.026	0.072	0.014
aroE	0.071	< 0.0005	0.109	0.053	0.065
fumC	0.027	< 0.0005	0.024	0.029	0.024
gdh	0.043	< 0.0005	0.082	0.027	0.035
pdhC	0.041	< 0.0005	0.063	0.046	0.014
pgm	0.022	< 0.0005	0.044	0.000*	0.035

CR = Czech Republic, G = Greece, N = Norway. * not significant at $p = 0.360$. All other pairwise F_{ST} are significant at $p < 0.05$.

The pairwise F_{ST} estimates exhibited a greater range of values ($F_{ST} = 0.000$ between the Czech Republic and Norway for *pgm* up to $F_{ST} = 0.109$ between the Czech Republic and Greece for *aroE*). All pairwise F_{ST} estimates were significant at $\alpha = 0.05$ except for $F_{ST} = 0.000$ between the Czech Republic and Norway for *pgm*. However, these pairwise *p*-values were not used to determine the significance of population differentiation, nor are they displayed in Table 4 because each pairwise comparison has lower power than the joint three-way comparison. Nevertheless, it is interesting that whilst the average pairwise F_{ST} across loci was smaller for CR *vs* N (using the notation of Table 4; $F_{ST} = 0.040$) than for CR *vs* G ($F_{ST} = 0.055$), the former being separated by 694 miles as opposed to 953 miles for the latter, it was smallest of all for G *vs* N ($F_{ST} = 0.030$) which are separated by 1619 miles. This is consistent with the

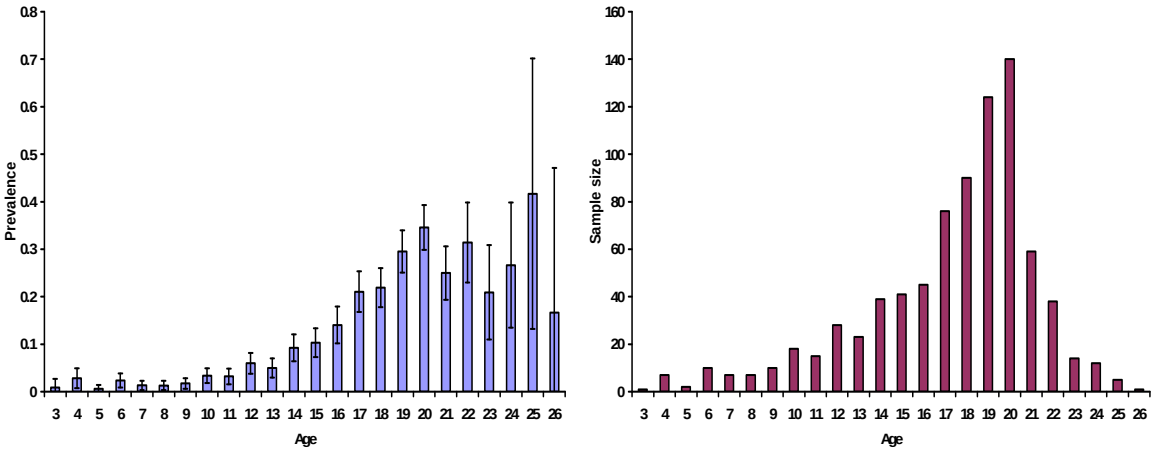


Figure 4 Left: Estimated prevalence for each age group in the Bavarian carriage study (Claus *et al.* 2005). The error bars show the approximate 95% confidence interval. Right: Sample size for each age group in the Bavarian carriage study.

average number of pairwise differences between isolates from the three pairs of countries: $\overline{\pi} = 16.0$ for CR *vs* N, $\overline{\pi} = 26.0$ for CR *vs* G and $\overline{\pi} = 14.3$ for the most distant countries G *vs* N. It is difficult therefore to come to the conclusion that genetic differentiation is due to a simple process of isolation by distance. The observed patterns of pairwise F_{ST} and $\overline{\pi}$ suggest that transmission routes are complex on the continental scale.

3.4 Meningococcal population structure in Bavaria

Bavaria is a region of Germany roughly the same size as the Czech Republic, spanning approximately 200 miles at the widest point. A carriage study was carried out in Bavaria in the winter of 1999-2000, in which 822 isolates were sampled from healthy carriers at schools and military institutions across the region (Claus *et al.* 2005). I used these sequences to analyse the role of host age-structure, geography and institution type on genetic structure in the meningococcal carriage population.

Table 5 Role of host age in meningococcal population structure

Locus	By age		By age group	
	F_{ST}	*p*	F_{ST}	*p*
abcZ	0.008	0.028	0.003	0.102
adk	-0.002	0.634	0.000	0.497
aroE	0.007	0.069	0.003	0.082
fumC	0.007	0.028	0.002	0.167
gdh	0.001	0.379	0.002	0.186
pdhC	0.003	0.159	0.001	0.296
pgm	0.006	0.102	0.004	0.049
Concatenated	0.004	0.089	0.003	0.047

3.4.1 Role of host age-structure

Figure 4 shows the estimated prevalence for each age group 3-26. The results are comparable to those of Cartwright *et al.* (1987; Chapter 1 Figure 5). In the chart the error bars indicates the approximate 95% confidence interval. Prevalence increases rapidly during the teenage years, peaking in the early twenties. Figure 4 also shows that the sample size from each age group roughly mirrors the prevalence. Isolates were sampled from carriers attending kindergarten (years 3-6), primary school (6-11), secondary school (10-17) and high school (15-21), and military recruits from six barracks (years 18-26). To determine whether host age plays any role in shaping meningococcal population structure I performed two analyses using AMOVA, one in

which each age was treated separately and one in which ages were pooled into the following groups: 3-9, 10-14, 15-19 and 20+.

The results of the analyses on each of the seven housekeeping loci, as well as the concatenated nucleotide sequence, are shown in Table 5. When the meningococcal population is analysed by host age (rather than host age group), F_{ST} ranges from -0.002 for *adk* up to 0.008 for *abcZ*. Both *fumC* and *abcZ* have significant *p*-values at the $\alpha = 0.05$ level ($p = 0.028$ for both), but this is not significant when correcting for multiple comparisons ($\alpha = 0.0073$ for seven loci). $F_{ST} = 0.004$ for the concatenated sequence is not significant at $\alpha = 0.05$, casting further doubt on the significance of host age on meningococcal population structure. When the meningococcal population is analysed by host age group (rather than host age), F_{ST} ranges from 0.000 for *adk* up to 0.004 for *pgm*. Of the *p*-values, *pgm* and the concatenated sequence are marginally significant at $\alpha = 0.05$ ($p = 0.049$ and 0.047 respectively). Taken together, there is no strong evidence that genetically distinct meningococcal populations circulate amongst different host age groups.

3.4.2 Geographic differentiation

The isolates were sampled from schools in eleven towns: Coburg (38 isolates), Weiden (9), Passau (47), Rottal-Inn (47), München (75), Oberallgäu (20), Augsburg

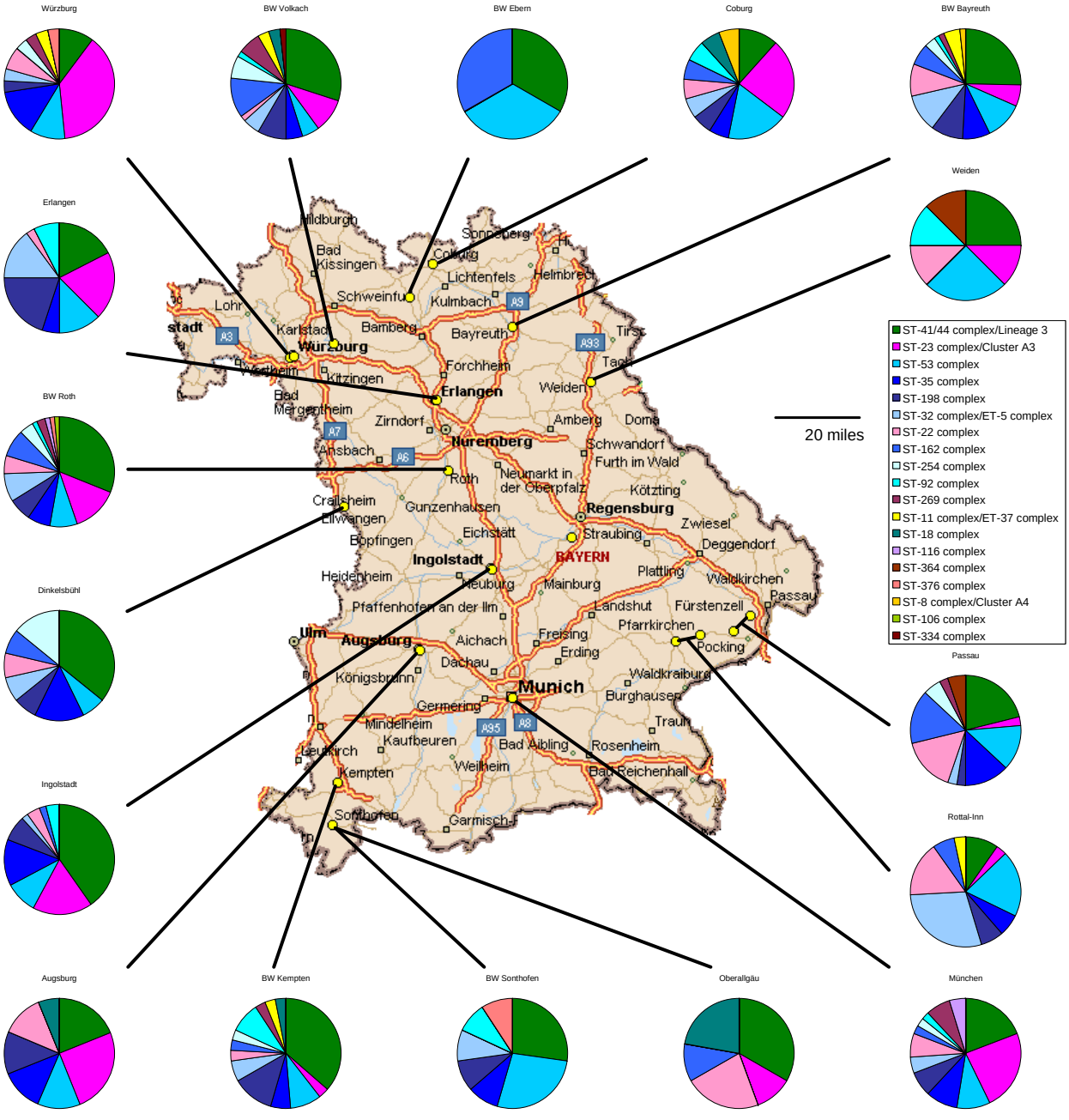


Figure 5 Genetic composition of meningococcal samples according to sampling location in Bavaria (Claus *et al.* 2005). Clonal complexes are colour-coded (see key). The principal clonal complexes are ST-41/44 complex (dark green), ST-23 complex (magenta) and ST-53 complex (sky blue). ST-11 complex, which is hyperinvasive, is coloured yellow.

(23), Ingolstadt (61), Dinkelsbühl (17), Erlangen (57) and Würzburg (47); and from recruits at six military barracks: Volkach (96 isolates), Ebern (10), Bayreuth (95), Sonthofen (21), Kempten (50) and Roth (109). In Figure 5 the sampling locations are highlighted on a map of Bavaria. For each sampling location a pie chart breaks down the genetic constitution of the sample according to clonal complex. The clonal complex assignment is used only for the purpose of comparing genetic profiles amongst the sampling locations; the ST-41/44 complex is often the most common, although ST-23 complex and ST-53 complex meningococci are also well-represented. The pie charts show that there is some difference in genetic constitution between sampling locations, but the effect is difficult to quantify. To determine the strength and nature of geographical differentiation between sampling locations in Bavaria I conducted analyses using AMOVA and the Mantel test.

3.4.2.1 Evidence for population structure

Using the concatenated nucleotide sequence, the average F_{ST} between sampling locations in Bavaria was 0.007, which is small, but very highly significant ($p < 0.0005$). To assess the nature of allele-sharing between sampling locations I performed AMOVA using different measures of pairwise genetic distance. In addition to defining genetic distance between a pair of sequences as the number of nucleotide mismatches, I also utilised the number of allelic mismatches. For an individual locus the number of allelic mismatches is simply 0 if the genes are identical or 1 if they are different. For the concatenated sequence the number of allelic mismatches ranges from 0 to 7. Finally, I utilised a definition of genetic distance that records simply whether the sequences are identical across all loci (0) or not (1). This definition applied only to the concatenated sequence. The results of AMOVA using the three

Table 6 Evidence for geographic differentiation in the Bavarian carriage study

Locus	Definition of pairwise genetic distance					
	No. nucleotide mismatches		No. allelic mismatches		Identity of entire sequence	
	F_{ST}	p	F_{ST}	p	F_{ST}	p
abcZ	**0.012**	**0.002**	**0.011**	**0.000**	-	-
adk	0.006	0.055	**0.011**	**0.004**	-	-
aroE	0.005	0.092	0.006	0.011	-	-
fumC	0.003	0.158	**0.008**	**0.000**	-	-
gdh	**0.010**	**0.004**	**0.013**	**0.000**	-	-
pdhC	**0.015**	**0.000**	**0.013**	**0.000**	-	-
pgm	-0.001	0.539	**0.007**	**0.005**	-	-
Concatenated	**0.007**	**0.000**	**0.010**	**0.000**	**0.008**	**0.000**

Enboldened entries indicate significance at $\alpha = 0.0073$.

definitions of genetic distance are shown in Table 6, for each locus and the concatenated sequence. The results are described below according to the definition of genetic distance.

Number of nucleotide mismatches

Whereas the concatenated nucleotide sequence provided evidence of significant population differentiation ($p = 0.007$), F_{ST} at individual loci varied from -0.001 for *pgm* up to 0.015 for *pdhC*. Population structuring was significant for *abcZ*, *gdh* and *pdhC* ($p = 0.002$, 0.004 and < 0.0005 respectively). There are a number of possible explanations. The most mundane is that analyses of individual loci have lower power

to detect significant population structure. However, a locus-specific effect such as balancing selection might lead to real differences in genetic differentiation between loci. Alternatively, it may be that at the nucleotide level genetic structuring is relatively weak, by which I mean that all populations share the same nucleotide polymorphisms, but that genetic differentiation is reflected in the combinations of those polymorphisms that appear in each population. An allele can be thought of as a particular combination of polymorphic nucleotides in a gene, and by defining genetic distance in terms of allelic mismatches, differences between the particular combinations of nucleotide polymorphisms are emphasised.

Number of allelic mismatches

In Table 6 the concatenated sequence exhibits strong evidence for limited population differentiation ($F_{ST} = 0.010$, $p < 0.0005$). In additional to the individual loci that exhibited significant population differentiation for the nucleotide mismatch definition of genetic distance (*abcZ*, *gdh* and *pdhC*), three of the four other loci also show evidence for significant differentiation at the allele level ($F_{ST} = 0.011$, $p = 0.004$, $F_{ST} = 0.008$, $p < 0.0005$ and $F_{ST} = 0.007$, $p = 0.005$ for *adk*, *fumC* and *pgm* respectively). That there is convincing evidence for population structure at six out of seven loci when genetic distance is measured as the number of allelic mismatches is consistent with the idea that populations share the same nucleotide polymorphisms, but differ in the particular combinations of those nucleotide polymorphisms that are circulating. In such a scenario, population structure at the allelic level can be explained by relatively recent recombination events, which generate novel alleles from an existing pool of shared nucleotide polymorphisms; in Chapter 2 (Tables 4 and

6) it was shown that recombination events cause genetic diversification an order of magnitude faster than *de novo* mutation.

Identity of entire sequence

The sequence type can be thought of as a particular combination of alleles across the seven loci, and to that extent measuring genetic distance as the identity or non-identity of the entire sequence, or equivalently the sequence type (ST), emphasises differences in the particular combinations of alleles between populations. Table 6 shows that there was strong evidence ($F_{ST} = 0.008$, $p < 0.0005$) for weak differentiation at the level of the ST between sampling locations in Bavaria.

3.4.2.2 Evidence for isolation by distance

If population structure within Bavaria can be explained by isolation by distance then there ought to be a positive correlation between geographic distance and genetic distance. Table 7 shows the correlation coefficient, *r*, between geographic distance and the three measures of genetic distance for each of the seven loci and the concatenated sequence. The *p*-values are obtained by using the Mantel test, in which isolates are permuted between sampling locations to obtain a null distribution for the correlation coefficient.

Table 7 Evidence for isolation by distance in the Bavarian carriage study

Locus	Definition of pairwise genetic distance					
	No. nucleotide mismatches		No. allelic mismatches		Identity of entire sequence	
	r	*p*	*r*	*p*	*r*	*p*
abcZ	-0.010	0.876	0.008	0.022	-	-
adk	-0.028	0.963	0.002	0.410	-	-
aroE	-0.009	0.796	0.007	0.119	-	-
fumC	0.007	0.301	0.009	0.121	-	-
gdh	0.006	0.218	**0.017**	**0.003**	-	-
pdhC	0.000	0.492	**0.017**	**0.003**	-	-
pgm	**0.030**	**0.003**	**0.015**	**0.004**	-	-
Concatenated	-0.002	0.599	**0.016**	**0.003**	**0.017**	**0.001**

Enboldened entries indicate significance at $\alpha = 0.0073$.

When pairwise genetic distance is defined as the number of nucleotide mismatches, there is little evidence for isolation by distance. For the concatenated sequence $r = -0.002$ ($p = 0.599$), which is not even a positive correlation, and is certainly not significant. For the individual loci, there is no evidence for isolation by distance except at *pgm* ($r = 0.030$, $p = 0.003$). For this definition of genetic distance, loci *abcZ*, *gdh* and *pdhC* showed significant population structure according to AMOVA (Table 6), but *pgm* did not. As before, the absence of strong evidence may reflect a lack of statistical power or the biologically more interesting explanation that particular nucleotide polymorphisms are not geographically structured, but their particular combinations are.

Such a hypothesis is supported by the results of the Mantel test using the number of allelic mismatches as the measure of genetic distance. Now three of the seven loci ($r = 0.017$, $p = 0.003$, $r = 0.017$, $p = 0.003$ and $r = 0.015$, $p = 0.004$ for *gdh*, *pdhC* and *pgm* respectively) exhibit significant evidence for isolation by distance, as does the concatenated sequence ($r = 0.016$, $p = 0.003$). The third definition of genetic distance, identity or non-identity of the ST, also yields evidence for isolation by distance ($r = 0.017$, $p = 0.001$). These results suggest that alleles and STs, which can be thought of as particular combinations of nucleotide polymorphisms and alleles respectively, are shared more rapidly between geographically proximate locations, resulting in genetic isolation by distance.

Table 8 Patterns of genetic structuring between schools and military institutions

Locus	Definition of pairwise genetic distance					
	No. nucleotide mismatches		No. allelic mismatches		Identity of entire sequence	
	F_{ST}	r	F_{ST}	r	F_{ST}	r
Schools only						
abcZ	**0.022**	-0.005	**0.020**	**0.032**	-	-
adk	0.014	-0.028	**0.020**	-0.010	-	-
aroE	0.009	0.000	**0.011**	0.016	-	-
fumC	0.010	0.009	**0.018**	0.019	-	-
gdh	0.012	0.006	**0.019**	**0.025**	-	-
pdhC	**0.027**	0.019	**0.023**	**0.040**	-	-
pgm	0.001	0.015	**0.014**	**0.025**	-	-
Concatenated	**0.013**	0.006	**0.018**	**0.028**	**0.013**	**0.028**
Military only						
abcZ	-0.001	0.007	-0.001	-0.004	-	-
adk	-0.002	-0.016	-0.004	0.018	-	-
aroE	-0.003	-0.043	-0.003	0.003	-	-
fumC	-0.003	0.000	-0.002	-0.007	-	-
gdh	0.005	0.015	0.000	0.008	-	-
pdhC	0.002	-0.008	0.000	0.006	-	-
pgm	-0.004	0.034	-0.005	0.003	-	-
Concatenated	-0.002	-0.022	-0.002	0.007	0.001	0.009

Enboldened entries indicate significant p-values at $\alpha = 0.0073$.

3.4.3 Institution type and genetic structure

Analysing genetic differentiation according to institution type in Bavaria revealed some interesting differences between schools and military barracks. AMOVA and Mantel tests were performed on two subsets of the carriage study: one in which schools only were analysed and one in which military barracks only were analysed. The results are shown in Table 8.

Taking a broad overview, the effect of analysing schools on their own is to increase the size of the estimates of F_{ST} and r compared to when schools and military institutions are analysed together (Tables 6 and 7). By contrast, the effect of analysing military barracks on their own is to drastically reduce the estimates of F_{ST} in particular. Estimates of r on the whole are closer to zero, but the effect is weaker and less consistent than the effect on F_{ST}. In terms of significance testing, where α was taken to be 0.0073 to control for multiple comparisons across loci, the effect of analysing schools on their own is to increase the number of loci that report significant population differentiation (as reported by F_{ST}) and isolation by distance (as reported by r). By contrast, the effect of analysing military barracks on their own is to remove all significant results, for both F_{ST} and r, across all loci and the concatenated sequence, for all three measures of genetic distance.

When all institutions are analysed together, AMOVA using the nucleotide mismatch definition of genetic distance reveals evidence of significant population differentiation at *abcZ*, *gdh* and *pdhC* and the concatenated sequence. The Mantel test reveals evidence for significant isolation by distance at *pgm*. For the same definition of genetic distance, when schools only are analysed, only *abcZ*, *pdhC* and the

concatenated sequence have significant F_{ST} and there is no evidence for significant isolation by distance. However, the magnitude of F_{ST} is greater for all loci and the concatenated sequence, and except for *pgm*, all loci and the concatenated sequence have more positive or equal *r*. The smaller number of significant results might be due to lower power resulting from reduced sample sizes when the data are partitioned. By contrast, when military barracks only are analysed all loci and the concatenated sequence have vastly reduced F_{ST}, six out of eight of which are negative. None of the estimates of F_{ST} or *r* are significant.

Using the allelic mismatch definition of genetic distance, when all institutions are analysed together, six out of seven loci and the concatenated sequence have significant F_{ST} and three out of seven loci and the concatenated sequence have significant *r*. When schools only are analysed, all seven loci and the concatenated sequence have significant F_{ST}, and an additional locus (*abcZ*) has significant *r*. Therefore when military barracks are removed from the analysis there is stronger evidence for population differentiation and isolation by distance. Indeed, when military barracks only are analysed, none of the loci or the concatenated sequence exhibits significant F_{ST} or *r*. Similarly, using the identity or non-identity of the entire sequence to define genetic distance, there is evidence for significant population differentiation and isolation by distance when all institutions are analysed together and when schools only are analysed, but not when military barracks only are analysed.

These results are striking in that it appears that whereas there is weak but discernable genetic differentiation amongst schools in different locations caused by isolation by

distance, military barracks in different locations appear to carry a homogeneous population of meningococci. When a simple AMOVA was conducted with two populations: schools versus military barracks, there was not significant evidence for population differentiation in the concatenated sequence using pairwise nucleotide mismatches ($F_{ST} = 0.002$, $p = 0.069$). These results suggest that the meningococci carried by military recruits are a homogenous sample of meningococci from across Bavaria, whereas meningococci carried by school children exhibit local differentiation. Such a conclusion is consistent with what is known about the catchment areas of the two institution types. Schools tend to have small catchment areas, drawing pupils from neighbouring towns, whereas the military barracks have very large catchment areas, drawing recruits from across Bavaria and other parts of Germany.

3.5 Relationship between disease and carriage

In addition to the 353 carriage isolates sampled by Yazdankhah *et al.* (2004) from the Czech Republic, Greece and Norway (see section 3.3.2), 314 disease-causing isolates were sampled from patients in the same three countries: 81 from the Czech Republic in 1994 and 1996, 91 from Greece in 1999 and 2000, and 142 from Norway in 1999 and 2000. These collections represented 37%, close to 100% and 85% of cases sent to the national reference laboratories in the three countries respectively during those years. Table 9 summarises the diversity in the carried and disease-causing meningococci populations from each of the three populations by estimating θ using the observed average number of pairwise differences $\overline{\pi}$ (see section 2.1.1). The estimates in the carriage populations resemble those for the 1993 Czech carriage study

Table 9 Estimates of $\theta \times 10^3$ based on pairwise differences

Locus	Czech Republic		Greece		Norway	
	Carriage	Disease	Carriage	Disease	Carriage	Disease
abcZ	42.93	23.54	49.50	39.58	48.49	38.26
adk	9.11	6.61	8.66	7.09	8.81	8.34
aroE	72.11	33.40	130.41	103.65	100.54	74.77
fumC	18.64	12.62	13.08	15.50	16.96	19.10
gdh	15.51	10.54	15.09	14.60	16.99	12.95
pdhC	47.60	35.37	35.41	30.68	35.89	36.60
pgm	45.47	25.45	35.98	34.68	44.27	29.03

(Jolley *et al.* 2000) obtained by the same method (Chapter 2, Table 1), except perhaps for *aroE* which shows elevated diversity here. Diversity in the disease-causing meningococci is on average slightly lower than in the carriage populations, but certainly on the same order of magnitude. The total number of carriage and disease sequences was 667, and these sequences were analysed using the two-way AMOVA described in section 3.2.1.1 to determine to what extent carriage and disease-causing meningococci represent distinct populations. A two-way AMOVA was used to control for differentiation due to country of origin, which has already been shown to be significant (see section 3.3.2).

Table 10 Summary of the logistic regression of disease status on serogroup and country

Estimates of odds ratios and *p*-values		
Factor	Odds Ratio	
Serogroup† ($p < 0.0005$)	Other baseline	B baseline
Other	1	0.01 (0.00–0.04)
B	75.0 (27.0–208.3)	1
C	377.4 (119.6–1190.7)	5.03 (2.73–9.25)
W135	43.4 (11.4–164.8)	0.58 (0.23–1.46)
Y	11.5 (3.3–39.6)	0.15 (0.07–0.33)
Country‡ ($p < 0.0005$)	Czech Republic baseline	
Czech Republic	1	
Greece	2.65 (1.52–4.63)	
Norway	2.47 (1.49–4.09)	

† Estimate of $\exp(\beta_i)$ for serogroup i. ‡ Estimate of $\exp(\gamma_j)$ for country j. See Equation 21.

Test that all factor coefficients are zero: deviance = 327.592, $p < 0.0005$

Goodness of fit tests	
Method	*p*-value
Pearson	0.233
Deviance	0.274
Hosmer-Lemeshow	0.850
Brown	
General alternative	0.173
Symmetric alternative	0.593

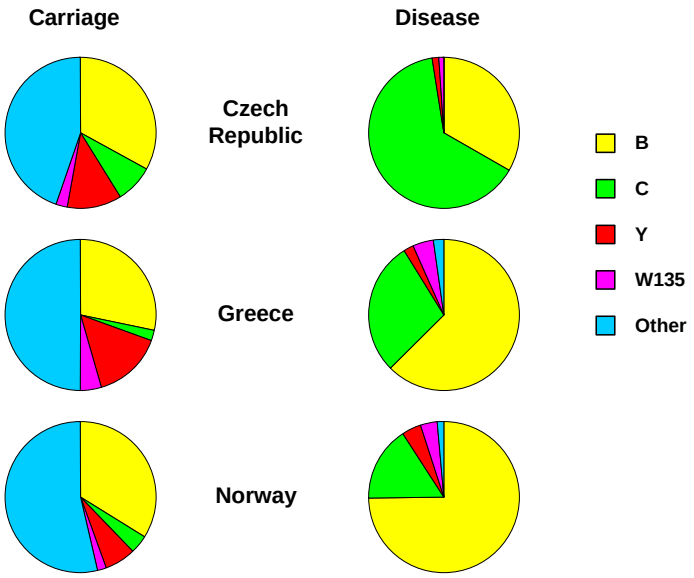


Figure 6 Distribution of serogroups in the Czech, Greek and Norwegian carriage and disease-causing isolate collections (Yazdankhah *et al.* 2004).

To some extent it is already appreciated that there is a genetic determinant to propensity to cause disease which varies between carried and disease-causing meningococci. Binary logistic regression of disease status (disease-causing or carriage isolate) on serogroup and country was performed. In the binary logistic regression the disease status for an isolate from serogroup i and country j is modelled as a Bernoulli random variable with parameter p_{ij}, where the log odds of p_{ij} are a linear function

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha+\beta_i+\gamma_j, \qquad (21)$$

where α is the baseline log odds of disease, $\exp(\beta_i)$ is the multiplicative change to the odds ratio for serogroup i relative to the baseline serogroup and $\exp(\gamma_j)$ is the multiplicative change to the odds ratio for country j relative to the baseline country. For short, $\exp(\beta_i)$ and $\exp(\gamma_j)$ are referred to as the odds ratio for serogroup i and

the odds ratio for country *j* respectively. The parameters are estimated by maximum likelihood, and the results are shown in Table 10. Serogroup, which is genetically determined, has a significant effect on disease ($p < 0.0005$). Serogroups differ in their propensity to cause disease by a factor of up to 377.4 (odds ratio for serogroup C versus other non-B, W135 or Y serogroups). Country was also found to be a significant predictor ($p < 0.0005$), with isolates from Greece and Norway more likely to cause disease. Disease-causing and carried meningococci differ in their serogroup profiles (Figure 6), with serogroups B and C over-represented in disease-causing meningococci. So serogroup, which is a genetic determinant of virulence, is known to vary between populations of disease-causing and carried meningococci. The purpose of the AMOVA was to determine whether populations of disease-causing and carried meningococci differ at housekeeping loci.

Table 11 Two-way AMOVA table for *abcZ*

AMOVA Table

	df	Seq SS	Adj SS	MS	F	σ^2
POP	2	227.62	282.41	141.2	15.85	0.61286
DIS	1	172.24	172.24	172.24	19.334	0.49377
Error	663	5906.4	5906.4	8.9086		8.9086
Total	666	6306.3				10.015

P value for POPULATION < 0.0005

P value for DISEASE < 0.0005

F_{ST} for POPULATION = σ_A^2/σ^2 = 0.0612

F_{ST} for DISEASE = σ_B^2/σ^2 = 0.0493

POPULATION records the country of origin and DISEASE whether the isolate is disease-causing or carried

Table 11 shows the output of the two-way AMOVA program for *abcZ* (see section 3.2.1.1 for details), and Table 12 summarises the results for all loci. The number of nucleotide mismatches was used to define pairwise genetic distance. There is very strong evidence for genetic differentiation, both between isolates sampled from different countries and between carriage and disease-causing isolates ($p < 0.0005$ for both effects for all loci). The effect of country ranges from $F_{ST}^{(A)} = 0.059$ for *pdhC* up to $F_{ST}^{(A)} = 0.111$ for *adk*, indicating that between 6% and 10% of total sequence variation can be attributed to differences between countries. The effect of disease ranges from $F_{ST}^{(B)} = 0.049$ for *abcZ* and *aroE* up to $F_{ST}^{(B)} = 0.101$ for *pgm*, indicating that between 5% and 10% of total sequence variation can be attributed to differences

between carried and disease-causing isolates. $F_{ST}^{(A)}$ and $F_{ST}^{(B)}$ are additive, in the sense that $F_{ST} = F_{ST}^{(A)} + F_{ST}^{(B)}$ is the proportion of the total sequence variation that can be attributed to differences between country or disease. The higher F_{ST}, the greater the proportion of total population variation is explained by these two effects. Together, the two effects explain a similar proportion of genetic variation across loci, ranging from F_{ST} = 0.110 for *abcZ* up to F_{ST} = 0.195 for *fumC*.

Disease-causing isolates are thus genetically distinct from carried isolates at housekeeping loci as well as at the serogroup locus (*cps*). Differentiation between disease-causing and carried meningococci in the same country is of the same order as differentiation between meningococci sampled from different countries. Therefore disease-causing meningococci, which are prevalent at much lower levels (Broome *et*

Table 12 Results of the two-way AMOVA

	Country		Disease	
Locus	$F_{ST}^{(A)}$	*p*	$F_{ST}^{(B)}$	*p*
abcZ	0.061	< 0.0005	0.049	< 0.0005
adk	0.111	< 0.0005	0.074	< 0.0005
aroE	0.074	< 0.0005	0.049	< 0.0005
fumC	0.097	< 0.0005	0.098	< 0.0005
gdh	0.085	< 0.0005	0.077	< 0.0005
pdhC	0.059	< 0.0005	0.093	< 0.0005
pgm	0.074	< 0.0005	0.101	< 0.0005

al. 1986; Caugant *et al.* 1994), are not a random sample of the carriage population at large. There are two possible explanations: (i) disease-causing meningococci constitute a genetically isolated population with limited genetic exchange with the carriage population, so drift causes the housekeeping loci of disease-causing and carried meningococci to diverge, or (ii) particular housekeeping loci in the carriage population are more likely to be associated with the emergence of disease-causing genotypes because they are determinants of virulence, or are closely linked to determinants of virulence. The former explanation, to some extent, contradicts the hypothesis that *N. meningitidis* is an accidental pathogen for which virulence is an evolutionary dead-end (Levin and Bull 1994; Maiden 2002; Stollenwerk *et al.* 2004) because disease-causing populations must persist sufficiently long for drift to cause differentiation. Yet the latter explanation is difficult to reconcile with the consistency of the signal of differentiation across loci. The fact that observed levels of diversity in disease-causing meningococci (Table 9) are almost as high as in the corresponding carriage populations lends support to the idea that virulent meningococci persist alongside carried meningococci with restricted genetic exchange between the two.

3.6 Summary

3.6.1 Causes of structure in meningococcal populations

In the previous chapter the standard neutral model of evolution was fitted to a meningococcal carriage population sampled from the Czech Republic in 1993 (Jolley *et al.* 2000). Using goodness-of-fit testing it was shown that the observed levels of genetic structuring, as measured by the number of unique STs and Tajima's *D* was incongruent with the estimated rates of mutation and recombination. In this chapter a

simple extension of the standard neutral model in which biased sampling causes an excess of identical isolates, known as the neutral microepidemic model, was also shown to inadequately describe observed patterns of genetic diversity in carried meningococci.

In this chapter I have used a variety of standard techniques (AMOVA, Mantel tests and logistic regression) to investigate the causes of genetic structuring in natural populations of meningococci. The results are not always consistent across datasets, emphasising the complexity of meningococcal population biology. Geography plays an important role in structuring meningococcal populations, but genetic differentiation is not always detectable. For example, isolates sampled from different towns across the Czech Republic, a region spanning roughly 200 miles at its widest, did not exhibit significant genetic differentiation when analysed using AMOVA (section 3.3.1). In contrast, isolates sampled from school children in Bavaria, a region of Germany comparable in size to the Czech Republic, showed strong evidence for weak genetic differentiation. Use of the Mantel test showed that this differentiation could be attributed to genetic isolation by distance (section 3.4.2). Why these two areas of Europe should show quite different patterns of geographical structure is unresolved. One could speculate that genetic uniformity within the Czech Republic is indicative of a wave of homogeneous meningococci passing rapidly through the country, whereas in Bavaria the meningococcal population has had time to differentiate locally. Alternatively, rates of transmission may be higher in the Czech Republic for unknown reasons.

Social considerations are important determinants of genetic structuring, as was seen by the marked difference in geographic differentiation in schools in Bavaria as opposed to military barracks. Whereas the schools showed local differentiation in the carried meningococci, the military barracks were homogeneous, and showed no correlation between sampling location and genotype. The discrepancy could be explained in this case due to the catchment areas of the two institution types: schools draw pupils from surrounding towns, whereas military barracks in Bavaria draw recruits from the entire region and other areas of Germany too. AMOVA showed that carried meningococci in military recruits are not distinct from those carried by school children, but the large catchment area of the military barracks causes meningococci to be sampled from all over the region. There was no evidence that host age shapes patterns of genetic diversity in meningococcal populations. At the continental scale, there is strong evidence for genetic differentiation between carried meningococci sampled from different countries (section 3.3.2). However, that differentiation did not appear to show a simple relationship with geographic distance. Isolates sampled from Greece and Norway appeared to be less genetically distinct than isolates sampled from either of those countries and the Czech Republic, which lies almost exactly between the two. Transmission routes at the continental level appear to be complex.

Disease-causing meningococci are not a random subset of the carriage population, to the extent that carriage and disease-causing isolates within the same country can exhibit as much genetic differentiation as meningococcal isolates sampled from different countries (section 3.5). Differentiation of disease-causing and carried meningococci was observed at all housekeeping loci. Some 5%-10% of genetic diversity in the total European-wide carriage and disease populations could be

attributed to differences between the carried and disease-causing genotypes, and another 6%-10% could be attributed to differences between genotypes in different countries. The fact that housekeeping loci in disease-causing isolates are not a random subset of housekeeping loci in carried meningococci, the evidence for significant differentiation between the housekeeping loci of disease-causing and carried meningococci, and the comparable levels of genetic diversity between the two suggest that disease-causing meningococci persist alongside carriage populations. To some extent this contradicts the hypothesis that *N. meningitidis* is an accidental pathogen.

Chapter 4

Evolutionary Model of Immune Selection

For a parasite such as *Neisseria meningitidis*, the key to long-term persistence is the successful and ongoing colonisation of a host. Despite its notorious pathogenicity, *N. meningitidis* normally resides as a commensal of the nasopharynx, but that is not to say that *N. meningitidis* is an onlooker in the co-evolutionary arms race between the host immune system and microbial intruders. Antigenic variation is a distinguishing feature of meningococcal populations, indicating that the observed genetic diversity at these loci is caused by strong selective pressures. Indeed, the patterns of genetic variation in samples of antigen gene sequences can be used to locate individual sites that interact directly with the host immune system.

Current methods that attempt to identify sites that interact with the immune system are based on reconstructing the phylogenetic tree of the gene sequences. In a highly recombining organism such as *N. meningitidis*, phylogenetic methods are not appropriate because there may be multiple trees along the sequence. In the presence of high levels of recombination phylogenetic methods that attempt to detect positive selection can have a false positive rate of up to 90% (Anisimova *et al.* 2003; Shriner *et al.* 2003). In this chapter I will begin by discussing the background to the dN/dS ratio (section 4.1.1), and the current phylogenetic methods for detecting immune selection (section 4.1.2). In section 4.2 I present a new population genetics model of immune selection in the presence of recombination, based on an approximation to the coalescent (Li and Stephens 2003). I also describe a model for variation in selection

pressure and the recombination rate within a gene, which is novel in the context of detecting selection. In section 4.3 I describe how to perform Bayesian inference on the selection and recombination parameters under the new model, using reversible-jump Markov chain Monte Carlo (MCMC). Then in section 4.4, I use a simulation study to investigate the properties of the inference method under two scenarios and demonstrate that the new method has the power to detect variability in selection pressure and recombination rate, and does not suffer from a high false positive rate. In Chapter 5 I apply the new method to the *porB* locus of *N. meningitidis* which encodes the PorB outer membrane protein. I use prior sensitivity analysis and model criticism techniques to verify the inferences, and compare the results to those obtained with phylogenetic methods.

4.1 The dN/dS ratio

4.1.1 Models that incorporate the dN/dS ratio

As an indicator of the action of natural selection in gene sequences the ratio of non-synonymous to synonymous substitutions (dN/dS) is a versatile and widely-used method of summarising patterns of genetic diversity. When comparing a pair of nucleotide sequences, synonymous substitutions refer to those codons that differ in their nucleotide sequence but not in the amino acid encoded. Non-synonymous substitutions refer to those codons that differ both in nucleotide sequence and amino acid encoded. In Figure 1 there are nine possible single nucleotide mutations of the triplet `CTT`, two of which are synonymous because leucine is still encoded, the other seven of which are non-synonymous.

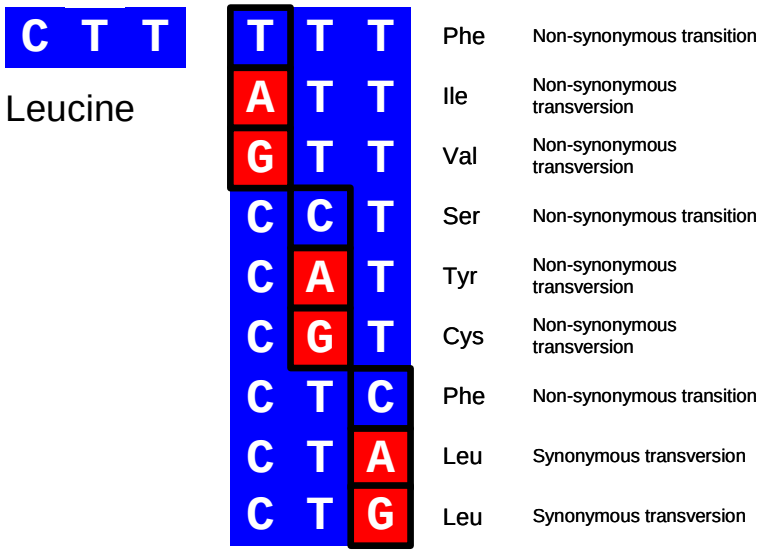


Figure 1 Synonymous and non-synonymous single nucleotide mutations from CTT.

In a strictly neutral model in which single nucleotide mutations occur at a uniform rate, non-synonymous mutations would occur more frequently than synonymous mutations, because there are more potential non-synonymous changes. The dN/dS ratio measures the relative rate at which non-synonymous and synonymous changes occur, adjusting for the fact that there are more potential non-synonymous changes. In a strictly neutral model of evolution, the dN/dS ratio equals one. The dN/dS ratio is an indicator of natural selection, because deviations from a ratio of one suggest that nucleotide changes that alter the amino acid sequence are more or less frequently observed than those that do not.

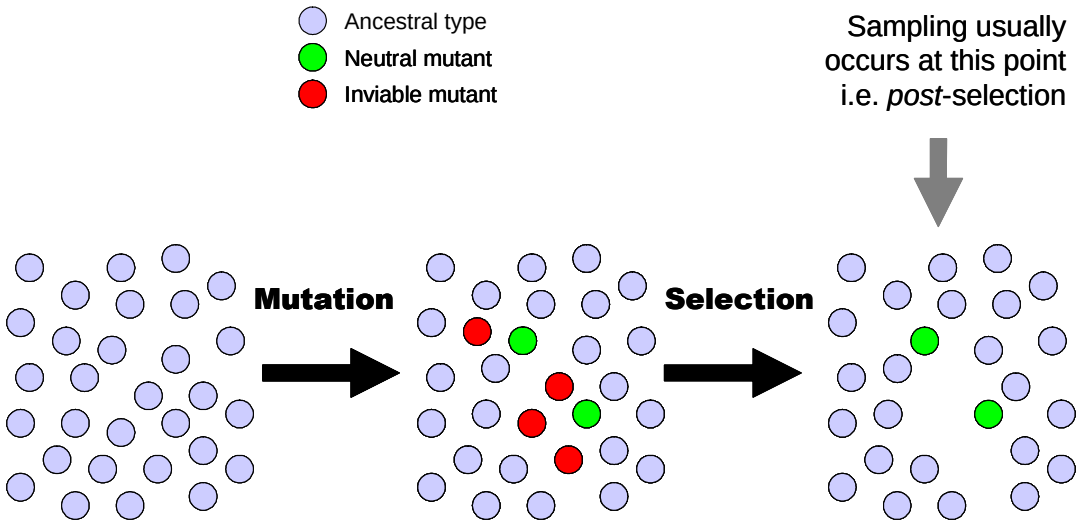


Figure 2 In samples of gene sequences the effects of mutation and selection on patterns of genetic diversity are confounded. For example, non-synonymous polymorphism might be under-represented because of purifying selection.

4.1.1.1 Purifying selection and dN/dS

Figure 2 illustrates how the observed patterns of synonymous and non-synonymous polymorphism represent a confounding between the evolutionary processes of mutation and natural selection. For example, it is generally assumed in studies of adaptation that organisms are optimally adapted to their environment (Dawkins 1982). This is a reasonable assumption because over long periods of time natural selection favours variants that have a selective advantage. If a gene is adapted to its environment, even if it is not optimally adapted, then there will be a great many more worse alternative sequences than better alternative sequences. So, random mutation will tend to produce less-well adapted sequences, not better adapted sequences. As a result of natural selection, those sequences that have reduced survival or reproductive success will be under-represented in a sample taken from the population. None of this applies to synonymous changes, of course, which do not alter the amino acid sequence of the gene product. As a result it is reasonable to expect that purifying, or

negative, selection will cause non-synonymous variants to be under-represented relative to synonymous variants, and the dN/dS ratio will be less than one in a functional gene. This is known as functional constraint.

The fact that mutation and natural selection are confounded in genetic samples serves as the basis for a class of evolutionary models of selection. Models of selection that describe the movement of alleles through the population (e.g. Fisher 1930) are not easily amenable to inference because for each site and each allele the selective advantage conferred by that allele (the selection coefficient), the time since the allele arose, and the way in which selection coefficients interact across sites, all need to be specified, resulting in a great many parameters. Such models exist, usually they make assumptions to reduce the number of parameters, but the inference methods are computationally prohibitive even when recombination is not modelled (e.g. Coop and Griffiths 2004). Evolutionary models that deliberately confound mutation and natural selection (Goldman and Yang 1994; Nielsen and Yang 1998; Sainudiin *et al.* 2005) use a single selection parameter for each site, the dN/dS ratio. In these models natural selection is treated as a form of mutational bias, so that if the dN/dS ratio is less than one then non-synonymous mutations simply occur at a lower rate.

In the codon model of Nielsen and Yang (1998), hereafter NY98, the mutation rate from codon i to j ($i \neq j$), which I will measure in units of PN_e generations (where P is the ploidy and N_e the effective population size) is

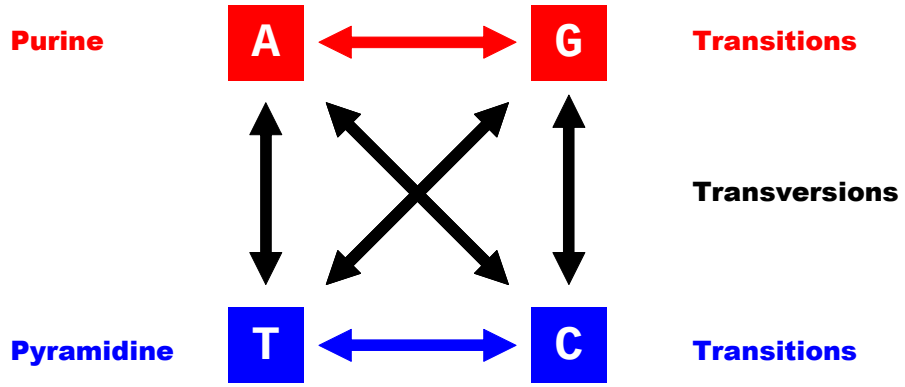


Figure 3 There are two classes of nucleotides, purines (adenosine and guanine) and pyrimidines (thymine, cytosine and uracil). Single nucleotide mutations that do not change the nucleotide class are called transitions, and those that do are called transversions. For any nucleotide there are two possible transversions and one transition. Despite this, transitions are observed more commonly than transversions, so the transition:transversion ratio κ is usually greater than two.

$$q_{ij} = \pi_j \mu \begin{cases} 1 & \text{if i and j differ by a synonymous transversion} \\ \kappa & \text{if i and j differ by a synonymous transition} \\ \omega & \text{if i and j differ by a nonsynonymous transversion,} \\ \kappa\omega & \text{if i and j differ by a nonsynonymous transition} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and $q_{ii} = -\sum_{j \neq i} q_{ij}$, where the frequency of codon j is π_j, κ is the relative rate of transitions to transversions (defined in Figure 3), and ω is the dN/dS ratio. If there were equal codon usage (i.e. $\pi_j = 1/61$ because only the 61 non-stop codons are allowed in NY98) the total rate of synonymous mutation (per PN_e generations) would be approximately,

$$\frac{\theta_s}{2} \approx \frac{(6+5\kappa)\mu}{310}. \quad (2)$$

4.1.1.2 Positive selection and dN/dS

When organisms are already well-adapted to their environment natural selection will purge the population of less-fit variant genes so non-synonymous polymorphism is under-represented relative to synonymous polymorphism and $dN/dS < 1$. The converse scenario, in which non-synonymous polymorphism is over-represented relative to synonymous polymorphism and $dN/dS > 1$ needs careful interpretation. An excess of non-synonymous polymorphism implies that there is a selective advantage to novelty in the amino acid sequence. It might be envisaged that recurrent, adaptive change in a gene will manifest itself as an over-representation of non-synonymous relative to synonymous change because positive selection will drive the adaptive variants to high frequency. Such a model has been used to detect natural selection between species, because it is assumed that multiple adaptive changes are important during speciation (McDonald and Kreitman 1991; Shpaer and Mullins 1993; Long and Langley 1993).

However, some controversy surrounds the generality with which adaptation leads to an excess of non-synonymous polymorphism. When positive dN/dS is observed, it is likely that multiple compensatory, or complementary, changes at several sites in the gene have occurred as a result of adaptation. So an excess of non-synonymous relative to synonymous polymorphism is a clear signal of adaptive change, or positive selection. But a single adaptive substitution at a particular codon is not sufficient to generate a positive dN/dS ratio across a whole gene if much of the gene is functionally constrained. So the dN/dS ratio will under-report the extent of adaptive change for any model in which episodic environmental change causes a transformation from one optimal state to a new optimum.

What an excess of non-synonymous to synonymous polymorphism is truly indicative of is selection for variation in the polypeptide sequence, not change from one conserved state to another. That makes the dN/dS ratio a particularly useful tool for studying the interaction between antigen genes and the immune system. Immunological memory against particular antigens exerts a strong selective pressure for antigenic novelty in the parasite population. This is known as diversifying selection. The antigenic properties of an outer membrane protein such as PorB may be determined by a small number of amino acids that might or might not be contiguous in the codon sequence. The dN/dS ratio can in principle be harnessed to estimate the magnitude of the selection pressure exerted by the immune system on different genes, investigate the evolutionary trade-off between protein functionality and immune evasion in the parasite, and locate the genetic determinants of antigenicity at a locus. The latter might be informative for vaccine development.

4.1.2 Inferring immune selection using dN/dS

Nielsen and Yang (1998) proposed a maximum likelihood phylogenetic approach to estimating the dN/dS ratio that employs a codon-based mutation model (Equation 1), and treats the dN/dS ratio as an unknown parameter ω. This method has subsequently been expanded (Yang *et al*. 2000; Yang and Swanson 2002; Swanson *et al*. 2003), adapted into a Bayesian setting (Huelsenbeck and Dyer 2004), and approximated for the purposes of computational efficiency (Massingham and Goldman 2005). Simulation studies have shown that phylogenetic likelihood-based methods can be substantially more powerful than alternative non-likelihood-based approaches

(Anisimova *et al*. 2001; Anisimova *et al*. 2002; Wong *et al*. 2004; Kosakovsky Pond and Frost 2005).

Estimating the selection parameter ω using these methods has become widespread (e.g. Bishop *et al*. 2000; Ford 2001; Mondragon-Palomino *et al*. 2002; Filip and Mundy 2004) and has been applied to many organisms. Analysis of pathogens such as viruses (Twiddy *et al*. 2002; Moury 2004; de Oliveira *et al*. 2004) and bacteria (Peek *et al*. 2001; Urwin *et al*. 2002) is particularly informative, because they typically have high mutation rates and are consequently genetically diverse, which lends greater statistical power to estimation. As discussed, the diversifying selection imposed by the host immune system may be the most appropriate model for which inference based on the dN/dS ratio can be applied. The ability to observe these populations evolving in real-time makes them especially interesting for the study of evolution (Drummond *et al*. 2003a), and suggests that we may be able to make useful epidemiological inference from molecular sequence data.

4.1.2.1 CODEML

The method of Nielsen and Yang (1998) is the most popular method for estimating the dN/dS ratio for nucleotide sequence data, and has been widely applied to samples within parasite populations. Based on the mutation model specified by Equation 1, in its original incarnation a random effects model is used for variation in ω between sites. To make inference feasible, only three classes of sites, occurring in proportions p_0, p_1 and p_2 are allowed. These have dN/dS ratios ω_0, ω_1 and ω_2 respectively, subject to the constraint that $\omega_0 < \omega_1 < \omega_2$. The method has three stages.

1. A tree topology is supplied or estimated using maximum likelihood (ML) from the data using a simple nucleotide mutation model.
2. Conditional on the topology, the branch lengths, κ, p_0, p_1 and $\boldsymbol{\omega}$ are estimated by maximum likelihood.
3. An empirical Bayes (Robbins 1956) approach is used to obtain the posterior probability that a given site is a member of a particular class.

The posterior probability that site *h* belongs to class *k*, so that the selection parameter at site *h*, w_h say, equals ω_k is taken to be

$$\Pr(w_h = \omega_k \mid \mathbf{X}_h) = \frac{p_k f(\mathbf{X}_h \mid w_h = \omega_k)}{\sum_{l=0}^{2} p_k f(\mathbf{X}_h \mid w_h = \omega_l)}, \quad (2)$$

(Nielsen and Yang 1998) where $\mathbf{X}_h$ is the codon alignment at site *h* and $f(\mathbf{X}_h \mid w_h = \omega_k)$ is the likelihood function. Equation 2 hides some of the conditioning however. The likelihood function in Equation 2 is not marginal to, but conditional upon the ML tree topology, branch lengths and κ, which are estimated using the alignment across all sites, $\mathbf{X}$. The posterior probability of belonging to class *k* is also conditional upon the ML estimates of p_0 and p_1.

The method of Nielsen and Yang (1998) is implemented in the program CODEML, part of the PAML package (Yang 1997). CODEML includes a large number of alternative specifications for the variation in ω over the sequence, including an arbitrary number of classes, gamma, beta and truncated normal distributions for the variation in ω across sites (the distributions have to be discretised for computational feasibility) and combinations thereof (Yang *et al.* 2000). Nielsen and Yang (1998) use a likelihood ratio test to compare nested models of variation in ω. For example, a model with three classes where $\omega_0 = 0$, $\omega_1 = 1$ and $\omega_2 > 1$ can be compared to a model

with only two classes where $\omega_0 = 0$ and $\omega_1 = 1$. This constitutes a test for positive selection. Nielsen and Yang (1998) assume that for nested models, the difference in double the log likelihood (the deviance) follows a χ^2 distribution with degrees of freedom equal to the difference in number of parameters. In practice this asymptotic result might not hold (Anisimova *et al.* 2001).

4.1.2.2 MrBayes

Huelsenbeck and Dyer (2004) implement the model of Nielsen and Yang (1998) in a fully Bayesian setting, available in MrBayes 3 (Ronquist and Huelsenbeck 2003). MrBayes uses MCMC to obtain a posterior distribution for all parameters of the model: codon frequencies, tree topology and branch lengths, κ, **p** and **ω**. Not surprisingly, it is considerably more computationally intensive than CODEML.

Huelsenbeck and Dyer (2004) fit a uniform prior on all unrooted tree topologies, and an exponential prior on branch lengths. Symmetric Dirichlet priors are applied to the frequencies **p** of the ω classes, and the codon frequencies. For the transition:transversion ratio κ, a distribution describing the ratio of two i.i.d. (independently and identically distributed) exponential random variables is used:

$$f(\kappa) = \frac{1}{(1+\kappa)^2}, \kappa > 0 .$$

Under their prior, the selection parameters ω_0, ω_1 and ω_2 are treated as ordered draws ($\omega_0 < \omega_1 < \omega_2$) from a distribution describing the ratio of two i.i.d. exponential random variables:

$$f(\omega_0, \omega_1, \omega_2) = \frac{36}{(1+\omega_0+\omega_1+\omega_2)^4} .$$

Because MrBayes is fully Bayesian, uncertainty in the phylogeny, mutation parameters and ω class frequencies is taken into account in the posterior probability that site *h* belongs to class *k*

$$\Pr(w_h = \omega_k \mid \mathbf{X}) = \int f(w_h = \omega_k, \Theta \mid \mathbf{X}) \mathrm{d}\Theta,$$

where Θ represents $\omega_{[-k]}, \mathbf{p}, \kappa$, the phylogenetic tree topology and branch lengths.

4.1.2.3 SLR

Massingham and Goldman (2005) introduced the sitewise likelihood ratio (SLR) method, which is an approximation to the ML method of Nielsen and Yang (1998). SLR is principally concerned with identifying the mode of selection at each site (i.e. dN/dS < 1 or dN/dS > 1).

The problem with estimating a separate dN/dS ratio for every codon in a sequence is that there are too many parameters. Nielsen and Yang (1998) overcame this problem by using a random effects model for the variation in ω which reduces the number of parameters to only a few. In CODEML, maximum likelihood estimates of the parameters are obtained using a high dimensional optimization procedure which is computationally intensive. In contrast, Massingham and Goldman (2005) use an approximation, described below, that allows a different ω to be estimated for each site. The approximation allows there to be a single multidimensional optimisation for the whole sequence (with fewer parameters than in CODEML) and then a one dimensional optimisation for each site.

The method works as follows

1. Assuming a common selection parameter for the whole sequence ω_0, the maximum likelihood tree, including branch lengths and the parameters κ and ω_0 are jointly estimated.
2. For each site *i* an individual selection parameter ω_i is estimated, assuming that the selection parameter for all other sites is ω_0.
3. For each site *i* a likelihood ratio test is performed for the null hypothesis that $\omega_i = 1$ by assuming that the difference the deviance between $\omega_i = \hat{\omega}_i$ and $\omega_i = 1$ is χ^2 distributed with one degree of freedom.

The method is approximate because for each site ω_i is estimated conditional upon all the other parameters, including ω_0. As a result estimating ω_i is a one dimensional problem for each site. The χ^2 distribution used in the likelihood ratio test is an asymptotic result that may not hold, so a parametric bootstrap procedure (Goldman 1993) can also be used to generate the null distribution of the difference in deviances.

4.1.2.4 Problems with current methods

CODEML, MrBayes and SLR all rely on reconstructions of the phylogenetic tree for the sample of genes. These methods have been applied frequently to within-population samples of micro-organisms (Twiddy *et al*. 2002; Moury 2004; de Oliveira *et al*. 2004; Peek *et al*. 2001; Urwin *et al*. 2002). However, the use of phylogenetic techniques is questionable in organisms that are highly recombining, because recombination leads to not one, but multiple evolutionary trees along the sequence. If the recombination rate is of the same order as the mutation rate, as has been found in some organisms (McVean *et al*. 2002; Stumpf and McVean 2003), then there might be a new evolutionary tree for every polymorphic site along the sequence. In such a scenario, which is plausible for many highly-recombining micro-organisms (Awadalla

2003) and eukaryotic genes containing recombination hotspots (McVean *et al*. 2004, Winckler *et al*. 2005), there is little hope to infer any particular evolutionary tree along the sequence. When a single evolutionary tree is estimated for a sample of gene sequences that have undergone recombination, the resulting tree is likely to have longer terminal branches and total branch length, yet a smaller time to the most recent common ancestor, in a way that superficially resembles the star-shaped topology of an exponentially growing population (Schierup and Hein 2000). The effect on detecting diversifying selection is to produce a high rate of false positives (Anisimova *et al*. 2003), as high as 90% (Shriner *et al*. 2003).

4.2 Modelling selection with recombination

4.2.1 Population genetics inference

When changes in the evolutionary tree are separated by only a few polymorphic sites, there is little hope to infer the tree at any particular site along the sequence. The population genetics approach is to treat the evolutionary trees along the sequence, or genealogy, as missing data. Because the likelihood of a set of molecular sequences needs to be evaluated with reference to a particular genealogy (Felsenstein 1981), it is calculated by averaging over the genealogies, weighted by the probability of that genealogy under the missing data model.

$$P(\mathbf{H} \mid \Theta) = \int P(\mathbf{H} \mid \Theta, G) P(G) \mathrm{d}G , \qquad (3)$$

where $P(\mathbf{H} \mid \Theta)$ is the likelihood of the data **H** given the parameters Θ, $P(G)$ is the missing data model for the genealogy and $P(\mathbf{H} \mid \Theta, G)$ is obtained using the pruning algorithm (Felsenstein 1981). There are various ways to model $P(G)$. In the case of

no recombination Huelsenbeck and Dyer (2004) used a model in which all unrooted tree topologies were uniformly likely, and branch lengths had an exponential distribution. When the sequences are from a single population a natural choice would be the coalescent (Kingman 1982a, 1982b; Hudson 1983; Griffiths and Marjoram 1997) which models a neutrally evolving, randomly mating population of constant size, with or without recombination.

However, $P(\mathbf{H} \mid \Theta, G)$ involves summation over the unknown states of internal nodes in the marginal genealogies (the evolutionary tree at a particular site), so the integration in Equation 3 cannot be solved analytically for any genealogical model, including the coalescent. As a result Equation 3 has to be evaluated numerically, which is not a trivial problem. Naïvely,

$$P(\mathbf{H} \mid \Theta) \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathbf{H} \mid \Theta, G^{(i)}), \tag{4}$$

for large *M*, where $G^{(i)}$ is simulated from $P(G)$. Unfortunately, for all but the simplest problems this method is useless because for most trees drawn from $P(G)$, the conditional likelihood $P(\mathbf{H} \mid \Theta, G)$ is negligibly small. Only once in a million draws would the conditional likelihood contribute significantly to the sum (Stephens 2003).

Importance sampling and Markov Chain Monte Carlo are methods that attempt to calculate Equation 4 more efficiently (see Stephens 2003). Both methods have been applied to a variety of contexts in population genetics (e.g. Kuhner *et al.* 1995, 1998, 2000; Griffiths and Marjoram 1996; Beerli and Felsenstein 1999, 2001; Bahlo and Griffiths 2000; Stephens and Donnelly 2000; Fearnhead and Donnelly 2001; Drummond *et al.* 2002; Wilson *et al.* 2003; Coop and Griffiths 2004; De Iorio *et al.*

2005). The methodology is more tractable in the absence of recombination because the state space of the possible genealogies is much smaller. In the presence of recombination, only the simplest models with two parameters (the mutation rate and recombination rate) have been implemented (Fearnhead and Donnelly 2001; Kuhner *et al.* 2000). Even for a small number of sequences these methods are extremely computationally burdensome. In the context of the NY98 mutation model with variation in the selection parameter and recombination rate amongst sites, such an approach is not feasible.

4.2.2 An approximation to the coalescent

Instead I turn to an approximation to the coalescent likelihood in the presence of recombination (Li and Stephens 2003) called the PAC likelihood ("product of approximate conditionals"). Their approach relies on rewriting the likelihood as

$$P(\mathbf{H} \mid \Theta) = P(H_1 \mid \Theta)P(H_2 \mid H_1, \Theta)\cdots P(H_n \mid H_1, H_2, \ldots, H_{n-1}, \Theta) \qquad (4)$$

where $\mathbf{H} = (H_1, H_2, \ldots, H_n)$ is the sample of n gene sequences (haplotypes). Li and Stephens approximate the $(k+1)$th conditional likelihood

$$P(H_{k+1} \mid H_1, H_2, \ldots, H_k, \Theta) \approx \hat{\pi}(H_{k+1} \mid H_1, H_2, \ldots, H_k, \Theta).$$

The approximate conditional likelihood, $\hat{\pi}$, that they use is a hidden Markov model that is designed to incorporate some key properties of the proper likelihood, notably that (i) the $(k+1)$th haplotype is likely to resemble the first k haplotypes but (ii) recombination means that it may be a mosaic of those haplotypes and (iii) mutation means that it may be an imperfect copy. In terms of averaging over possible evolutionary trees, one can think of the hidden Markov model doing so implicitly, but in an approximate way that is highly computationally efficient.

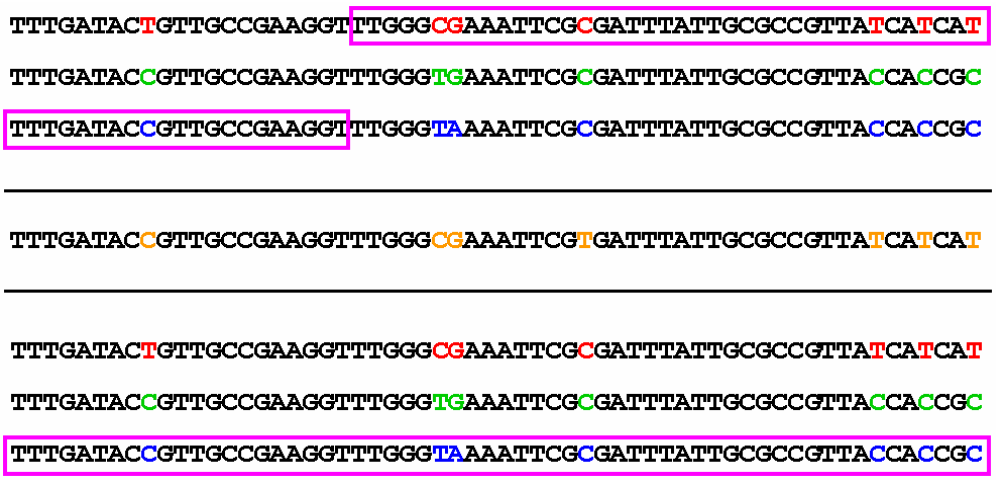


Figure 4 Approximate likelihood of the orange haplotype conditional on the red, green and blue haplotypes. In Li and Stephens' (2003) model, the orange haplotype resembles the others, but recombination means it may be a mosaic and mutation means that it may be an imperfect copy. In the top scenario, the orange haplotype is a mosaic of the red and blue haplotypes, necessitating a C→T mutation. In the bottom scenario, the orange haplotype is a copy of the blue haplotype, necessitating five mutations: T→C, and four C→Ts.

As a result of the approximate nature of the PAC likelihood, the ordering of the n haplotypes can influence the value of the likelihood (were it not for the approximation, the haplotypes would be exchangeable). Therefore, the likelihood is assessed by averaging over multiple orderings of the haplotypes. In the analyses I present throughout this chapter and Chapter 5, I use 10 orderings unless otherwise stated.

4.2.2.1 Sampling formula with recombination

Li and Stephens (2003) use a hidden Markov model (HMM) to approximate the likelihood of the $(k+1)$th haplotype conditional on the first k. Theirs is an approximation to the sampling formula in the sense of Ewens (1972), with the

additional complication of recombination. Li and Stephens think of the $(k+1)$th haplotype as a copy of the first *k* haplotypes. Figure 4 illustrates the idea. At every site, the orange haplotype is a copy of one of the four other haplotypes. This haplotype can be thought of as being closest to the orange haplotype in the evolutionary tree. Parsing the sequence 5′ to 3′, the orange haplotype is a copy of the blue haplotype, so at the first polymorphic site, depending on the mutation rate, it is most likely to share the same nucleotide C. Continuing along the sequence, the orange haplotype can switch between the other four with a given probability. However, if the orange haplotype is a copy of the blue haplotype at site *i*, then it is most likely to continue copying the blue haplotype at site $(i+1)$. This models the way that recombination creates mosaics of contiguous sequences. Between the first and second polymorphic site, the orange haplotype might switch from copying the blue to copying the red haplotype (Figure 4, top). In that case only one mutation need be invoked for the rest of the sequence. However, with some probability the orange continues to copy the blue haplotype (Figure 4, bottom), in which case five more mutation events need to be invoked.

4.2.2.2 Mutation model

In the lexicon of HMMs, the latent variable records which of the first *k* haplotypes the $(k+1)$th is a copy of at a given site. Conditional on the latent variable *x* ($x = 0,1,\ldots,k$), the emission probability models the mutation process, because it specifies the probability of observing state $a = H_{k+1,i}$ in haplotype $(k+1)$ given state $b = H_{x,i}$ in haplotype *x*, at a particular site *i*. Under a coalescent model (Kingman 1981, Hudson 1983), the time (in units of PN_e generations) to the common ancestor of

haplotypes x and $k+1$ is known (R. C. Griffiths, unpublished), and to the order of the approximation is exponentially distributed with rate k. Consider a simple mutation model with two states 0 and 1, and mutation rate $\theta/2$ per PN_e generations. The model is defined by the instantaneous rate matrix

$$\mathbf{Q} = \begin{vmatrix} -\theta/2 & \theta/2 \\ \theta/2 & -\theta/2 \end{vmatrix}. \tag{5}$$

The matrix $\mathbf{P}^{(t)}$ gives the probability $p_{ij}^{(t)}$ of a site being in state j time t after it was in state i.

$$\mathbf{P}^{(t)} = e^{t\mathbf{Q}} \tag{6}$$

(see Grimmett and Stirzaker 2001), which can be solved analytically for this model to give

$$p_{ij}^{(t)} = \begin{cases} \frac{1}{2} + \frac{1}{2}\exp\{-\theta t\} & \text{for } i = j \\ \frac{1}{2} - \frac{1}{2}\exp\{-\theta t\} & \text{for } i \neq j \end{cases}.$$

The probability of observing an (unordered) pair of states (a,b) given the time t to their common ancestor for a reversible mutation rate matrix (such as $\mathbf{Q}$) is

$$P(a,b \mid t) = \delta_{ab}\pi_a p_{ab}^{(2t)}, \tag{7}$$

where $\pi_0 = \pi_1 = 1/2$ are the equilibrium frequencies of states 0 and 1, and

$$\delta_{ab} = \begin{cases} 1 & \text{for } a = b \\ 2 & \text{for } a \neq b \end{cases}.$$

So

$$P(a,b \mid t) = \begin{cases} \frac{1}{4} + \frac{1}{4}\exp\{-\theta t\} & \text{for } a = b \\ \frac{1}{2} - \frac{1}{2}\exp\{-\theta t\} & \text{for } a \neq b \end{cases}.$$

To obtain the probability of observing a pair of states unconditional on the time to their common ancestor involves the integration

$$P(a,b)=\int_0^\infty P(a,b\mid t)P(t)\,\mathrm{d}t\,, \tag{8}$$

where $P(t)=k\exp\{-kt\}$ from before. Therefore the emission probability is defined by

$$P(a,b)=\begin{cases}\dfrac{2k+\theta}{4(k+\theta)} & \text{for } a=b\\ \dfrac{\theta}{2(k+\theta)} & \text{for } a\neq b\end{cases},$$

which is normalised because $P(0,0)+P(0,1)+P(1,1)=1$. Li and Stephens (2003) denote the emission probability

$$\gamma_i(x)=P(H_{k+1,i},H_{x,i}). \tag{9}$$

4.2.2.3 Recombination model

The transmission probability models recombination, because it specifies the probability of a switch from copying one haplotype to copying another between adjacent sites *i* and (*i* + 1). Li and Stephens (2003) model the length of sequence before a switch as exponentially distributed with rate ρ/k. This is based on the informal idea that $E(t)=1/k$, so the average rate of recombination between a pair of sequences is roughly $(\rho/2)\times(2/k)$. Under this crude approximation, the transmission probability is defined by

$$P(X_{i+1}=x'\mid X_i=x)=\begin{cases}\exp\{-\rho_i d_i/k\}+(1-\exp\{-\rho_i d_i/k\})/k & \text{if } x'=x\\ (1-\exp\{-\rho_i d_i/k\})/k & \text{otherwise}\end{cases} \tag{10}$$

where X_i is the copied haplotype at site i, X_{i+1} is the copied haplotype at site $(i + 1)$, and d_i is the distance (in bp) between sites i and $(i + 1)$. In this model there can be a different recombination rate ρ_i between every pair of adjacent sites.

4.2.2.4 Computing the likelihood

To calculate the approximate conditional likelihood requires a summation over all possible combinations of the latent variable at every site; that is to say, all possible mosaics that might constitute the $(k+1)$th haplotype. The advantage of the HMM is that this computation is fast using the forward algorithm (e.g. Rabiner 1989). Suppose that $\alpha_i(x)$ is the joint likelihood of the first i sites and $X_i = x$. Then the approximate conditional likelihood is

$$\hat{\pi}(H_{k+1} \mid H_1, H_2, \ldots, H_k, \Theta) = \sum_{x=1}^{k} \alpha_L(x),$$

when there are L sites. From the forward algorithm,

$$\begin{aligned} \alpha_{i+1}(x) &= \gamma_{i+1}(x) \sum_{x'=1}^{k} \alpha_i(x') P(X_{i+1} = x \mid X_i = x') \\ &= \gamma_{i+1}(x) \left(p_i \alpha_i(x) + (1 - p_i) \frac{1}{k} \sum_{x'=1}^{k} \alpha_i(x') \right), \end{aligned} \tag{11}$$

where $p_i = \exp\{-\rho_i d_i / k\}$. Because the second term in Equation 11 does not depend on x, it only needs to be computed once for each site. As a result, the computational complexity of the approximate conditional likelihood $\hat{\pi}$ is linear in L and linear in the total sample size n. The complexity of the full PAC likelihood is, therefore, linear in L and quadratic in n (Li and Stephens 2003).

4.2.3 NY98 in the coalescent approximation

Incorporating the NY98 mutation model in to the coalescent approximation of Li and Stephens (2003) is straightforward. The instantaneous mutation rate matrix **Q** in Equation 5 is replaced by that defined by Equation 1. However, the exponentiation of the NY98 rate matrix in Equation 6 cannot be solved analytically. Instead, a numerical technique known as diagonalisation is used. Equation 6 can be re-written using the matrix factorisation

$$\mathbf{P}^{(t)} = \mathbf{V}e^{t\mathbf{D}}\mathbf{V}^{-1} \tag{12}$$

(Grimmett and Stirzaker 2001) where **V** is a matrix whose columns are the right eigenvectors of **Q**, $\mathbf{V}^{-1}$ is its inverse and **D** is a diagonal matrix whose diagonal elements are the eigenvalues of **Q**. Exponentiation of a diagonal matrix is trivial, because

$$\exp\{t\mathbf{D}\}_{ij} = \begin{cases} \exp\{d_{ij}t\} & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}. \tag{13}$$

So, breaking down Equation 12 into parts for simplification,

$$\mathbf{P}^{(t)} = \mathbf{M}\mathbf{V}^{-1}$$

where

$$\mathbf{M} = \mathbf{V}e^{t\mathbf{D}}.$$

Now, using Equation 13 and the laws of matrix multiplication,

$$m_{ij} = v_{ij}\exp\{d_{jj}t\}$$

so

$$p_{ij}^{(t)} = \sum_{c \in C} v_{ic}\exp\{d_{cc}t\}v_{cj}^{(-1)}, \tag{14}$$

where C is the state space of **Q**, which consists of the 61 non-stop codons for NY98. Using Equation 7, the probability of observing a pair of states $a = H_{k+1,i}$ and $b = H_{x,i}$ when the $(k+1)$th haplotype is copying from the xth haplotype is,

$$P(a,b \mid t) = \delta_{ab}\pi_a \sum_{c \in C} v_{ac} v_{cb}^{(-1)} \exp\{2d_{cc}t\}.$$

Following Equation 8, one can obtain an expression for the HMM emission probability under any reversible mutation matrix **Q**

$$P(a,b) = \delta_{ab}\pi_a \sum_{c \in C} v_{ac} v_{cb}^{(-1)} \frac{k}{k - 2d_{cc}}. \quad (15)$$

Equation 15 is useful because it means that the PAC likelihood can be adapted to any reversible mutation model, of which NY98 is just an example (e.g. Rodríguez *et al.* 1990; Goldman and Yang 1994; Sainudiin *et al.* 2005). For a particular combination of the mutation rate parameters μ, κ and ω, the rate matrix **Q** must be diagonalised, which is to say its eigenvalues and right eigenvectors must be found (Equation 12). This can be achieved for any general real matrix **Q** using a numerical algorithm, available in libraries such as Numerical Recipes (Press *et al.* 2002), LAPACK (Anderson *et al.* 1999) or NAG. See Wilkinson and Reinsch (1971) for details of the algorithm. One problem with the algorithm for diagonalising a general real matrix is that the eigenvalues and eigenvectors are not guaranteed to be real numbers. In fact the eigenvalues and eigenvectors of a reversible rate matrix are real. I am grateful to Ziheng Yang for showing how further factorisation of Equation 12 leads to diagonalisation of a symmetric real matrix, for which the algorithms are guaranteed to produce real eigenvalues and eigenvectors. The algorithm for diagonalising a symmetric real matrix is also quicker and safer than the algorithm for diagonalising a

general real matrix. The code I used for the implementation of this algorithm was kindly provided by Ziheng Yang.

A reversible, irreducible mutation rate matrix **Q**, which is given by Equation 1 for NY98, can be re-written

$$\mathbf{Q} = \mathbf{S\Pi}$$

where **S** is a symmetric matrix ($s_{ij} = q_{ij} / \pi_j, i \neq j$, cf. Equation 1) and **Π** is a diagonal matrix whose diagonal elements are the stationary frequencies π_j of the rate matrix. The eigenvalues and eigenvectors of **Q** can be obtained by constructing a symmetric matrix

$$\mathbf{A} = \mathbf{\Pi}^{1/2}\mathbf{Q}\mathbf{\Pi}^{-1/2},$$

because the eigenvalues of **A** and **Q** are the same (contained in the diagonal matrix **D**), and the matrix of right eigenvectors **V** for matrix **Q** is related to the matrix of right eigenvectors **R** for matrix **A** by the formulae

$$\mathbf{V} = \mathbf{\Pi}^{-1/2}\mathbf{R},$$
$$\mathbf{V}^{-1} = \mathbf{R}^{-1}\mathbf{\Pi}^{1/2}.$$

Matrices **D** and **R** are obtained by diagonalising **A** using the algorithm for a symmetric real matrix. Because **R** is orthogonal, $\mathbf{R}^{-1} = \mathbf{R}^{T}$, so no matrix inversion is required for obtaining $\mathbf{V}^{-1}$. By matrix multiplication

$$\begin{aligned} v_{ac} &= \pi_a^{-1/2} r_{ac} \\ v_{cb}^{(-1)} &= r_{bc} \pi_b^{1/2} \end{aligned}. \qquad (16)$$

Therefore, Equation 15 can be re-written

$$P(a,b) = \delta_{ab} \pi_a^{1/2} \pi_b^{1/2} \sum_{c \in C} r_{ac} r_{bc} \frac{k}{k - 2d_{cc}}. \qquad (17)$$

This is the actual formula used in the implementation of the model.

4.2.4 An indel model for NY98

Alignments of nucleotide sequences from antigen loci are punctuated by gaps in the alignment caused by insertion or deletion mutations (indels). A sequence alignment is a statement of the homology of particular nucleotides in one sequence to those in the other sequences. Indels cause gaps in the nucleotide sequence alignment in multiples of three when the gene is functional, because otherwise a frameshift will ensue, and the remaining sequence will be nonsense. Indels are an important feature of the evolution of antigen loci, but even simple treatments of indels result in complex models that do not share the nice properties of the reversible nucleotide and codon models in common usage (e.g. Thorne *et al.* 1991, 1992). Here I make a very simple extension of NY98 in order to incorporate an extra indel state. The motivation for using this model is not to provide a realistic model of insertion/deletion, but to capture the information regarding the underlying tree structure and mode of selection at sites segregating for indels in the simplest possible way. The model is only applied to columns in the alignment that are segregating for an indel.

For columns segregating for an indel, codons are assumed to mutate to the indel state at rate $\pi_{\text{indel}}\varphi\omega$ and back at rate $(1-\pi_{\text{indel}})\varphi\omega$. Here π_{indel} is the equilibrium frequency of indels (in sites segregating for indels), φ is the rate of insertion/deletion, and ω is the selection parameter for that site. The model can be thought of in two parts: the NY98 model is nested within a two state codon vs. indel model (0 = codon, 1 = indel) specified by

$$\mathbf{Q}^{*} = \begin{vmatrix} -\pi_{\text{indel}}\varphi\omega & \pi_{\text{indel}}\varphi\omega \\ (1-\pi_{\text{indel}})\varphi\omega & -(1-\pi_{\text{indel}})\varphi\omega \end{vmatrix}. \quad (18)$$

Exponentiating Equation 18 gives the transition probability matrix between codon and indel states. So

$$p_{ij}^{*(t)} = \begin{cases} 1-\pi_{indel}(1-\exp\{-\omega\varphi t\}) & \text{for two (unspecified) codons} \\ \pi_{indel}+(1-\pi_{indel})\exp\{-\omega\varphi t\} & \text{for two indels} \\ \pi_{indel}(1-\exp\{-\omega\varphi t\}) & \text{if i is a codon and j an indel} \\ (1-\pi_{indel})(1-\exp\{-\omega\varphi t\}) & \text{if i is an indel and j a codon} \end{cases}. \quad (19)$$

Denote the full transition probability matrix for the NY98 model with indels $\mathbf{P}^{(t)}$. From Equation 19, part of this matrix is apparent

$$p_{ij}^{(t)} = \begin{cases} \pi_{indel}+(1-\pi_{indel})\exp\{-\omega\varphi t\} & \text{for two indels} \\ \pi_{indel}(1-\exp\{-\omega\varphi t\}) & \text{if i is a codon and j an indel} \\ \pi_{j}(1-\pi_{indel})(1-\exp\{-\omega\varphi t\}) & \text{if i is an indel and j a codon} \end{cases}.$$

When *i* and *j* are both codons, $p_{ij}^{(t)}$ can be found by conditioning on whether there are intermediate indels. Denote $\mathbf{N}^{(t)} = \{\nu_{ij}^{(t)}\}$ for the transition probability matrix of the NY98 model without indels. Conditional on intermediate indels, the transition probability from codon *i* to *j* in time *t* is simply π_j. Conditional on no intermediate indels, the transition probability from codon *i* to *j* in time *t* is $\nu_{ij}^{(t)}$. Since the probability of no intermediate indels is $\exp\{-\pi_{\text{indel}}\varphi\omega\}$, for a pair of codons

$$p_{ij}^{(t)} = \nu_{ij}^{(t)}\exp\{-\pi_{\text{indel}}\varphi\omega\} + \pi_j\left[1-\pi_{indel}(1-\exp\{-\omega\varphi t\}) - \exp\{-\pi_{\text{indel}}\varphi\omega\}\right].$$

Using Equations 8, 14 and 16 the emission probabilities for the PAC likelihood are obtained. For two identical codons

$$P(a,a) = \pi_a^2(1-\pi_{indel})\left[(1-\pi_{indel}) + \frac{k\pi_{indel}}{k+2\omega\varphi} - \frac{k}{k+2\pi_{indel}\omega\varphi} + \frac{1}{\pi_a}\sum_{c\in C}\frac{kr_{ac}r_{bc}}{k+2\pi_{indel}\omega\varphi-2d_{cc}}\right] \quad (20a)$$

where *C* is the state space of the NY98 model. For two non-identical codons

$$P(a,b) = 2\pi_a\pi_b(1-\pi_{indel})\left[(1-\pi_{indel}) + \frac{k\pi_{indel}}{k+2\omega\varphi} - \frac{k}{k+2\pi_{indel}\omega\varphi} + \frac{1}{\sqrt{\pi_a\pi_b}}\sum_{c\in C}\frac{kr_{ac}r_{bc}}{k+2\pi_{indel}\omega\varphi-2d_{cc}}\right] \quad (20b)$$

For two indels

$$P(a,a) = \pi_{indel}^2 + \frac{\pi_{indel}k(1-\pi_{indel})}{k+2\omega\varphi}. \quad (20c)$$

For a codon *a* and an indel *b*

$$P(a,b) = 2\pi_a\pi_b(1-\pi_{indel})\left(1-\frac{k}{k+2\omega\lambda}\right). \quad (20d)$$

4.2.5 Variation in ω and ρ along a gene

The primary aim of the new method is to obtain posterior distributions for ω and ρ, allowing both to vary along the length of the sequence. The information regarding either ω or ρ at a given position along the sequence is limited by the number of mutations in the underlying evolutionary history. This is a potentially serious limitation, particularly for sequences with low diversity. In an attempt to exploit to the full the available information, I use a independent prior distributions on ω and ρ in which adjacent sites may share either parameter in common. I will describe the model of variation in ω for the purposes of information. The model of variation for ρ is of the same form.

For a sequence of length L codons, the prior distribution imposes a 'block-like' structure on the variation in ω with two fixed and B_ω $(0 \le B_\omega \le L-1)$ variable transition points,

$$\mathbf{s}^{(B_\omega)} = (s_0, s_1, \ldots, s_{B_\omega+1}),$$

where $(s_0 = 0) < s_1 < s_2 < \cdots < s_{B\omega} < (s_{B_\omega+1} = L).$

Block j is delimited by transition points (s_j, s_{j+1}) and has a common selection parameter ω_j. I model the number of variable transition points in the region as a binomial distribution with parameters $(L-1, p_\omega)$. Given the number of transition points, the selection parameter for each block is independently and identically distributed. For an exponential prior on ω_j with rate parameter λ, the prior distribution on the transition points and selection parameters can be written

$$P\left(B, \mathbf{s}^{(B_\omega)}, \boldsymbol{\omega}^{(B_\omega)}\right) = p_\omega^{B_\omega} (1-p_\omega)^{L-B_\omega-1} \lambda^{B_\omega+1} \exp\left\{-\lambda(\omega_0 + \omega_1 + \cdots + \omega_{B_\omega})\right\} \quad (21)$$

In this model, the expected length of a block is $L/([L-1]p_\omega + 1) \approx 1/p_\omega$. For $p_\omega = 0$ there is a single block, producing a constant model for ω along the sequence, and for $p_\omega = 1$ every site has its own independent ω.

This prior structure is based on the multiple change-point model of Green (1995) which was adopted by McVean *et al*. (2004) to estimate variable recombination rates along a gene sequence, although the binomial model that I have used here is designed specifically so that transition points must fall between codons at a finite $(L-1)$ number of positions. I implement a block-like prior on ρ of the same form as for ω, but the block structure for ρ is independent of the block structure for ω, and the number of variable transition points is binomially distributed with parameters $(L-2, p_\rho)$. It is assumed that recombination only occurs between codons and not within. To perform inference jointly on variation in ω and ρ along the sequence I will use reversible-jump MCMC.

Table 1 Notation used for Constants

n	Sample size
L	Number of codons
P	Ploidy
N_e	Effective population size

Table 2 Parameters of the Model

μ	Rate of synonymous transversion per PN_e generations
κ	Transition:transversion ratio
B_ω	Number of changes in the dN/dS ratio along the sequence
$s_j^{(B_\omega)}, j = 0 \ldots B_\omega + 1$	Positions at which the dN/dS ratio changes along the sequence
$\omega_j, j = 0 \ldots B_\omega$	dN/dS ratio between sites $s_j^{(\omega)}$ and $s_{j+1}^{(\omega)}$
B_ρ	Number of changes in the recombination rate along the sequence
$s_j^{(B_\rho)}, j = 0 \ldots B_\rho + 1$	Positions at which the recombination rate changes along the sequence
$\rho_j, j = 0 \ldots B_\rho$	Recombination rate between sites $s_j^{(\rho)}$ and $s_{j+1}^{(\rho)}$
φ	Rate of insertion/deletion per PN_e generations

4.3 Bayesian inference

To summarise, Tables 1 and 2 list the constants and parameters of the model. The parameters together in Table 2 are denoted Θ, and the aim of Bayesian inference is to obtain a posterior distribution of Θ given the data **H**. To do so I will use Markov Chain Monte Carlo (MCMC; see for example O'Hagan and Forster [2004] for

Table 3 MCMC Moves

Type	Move	Relative proposal probability
A	Change μ	●●
A	Change κ	●●
A	Change φ	●●
A	Change ω within a block	●●●
A	Change ρ within a block	●●●
B	Extend an ω block 5′ or 3′	●●●
B	Extend an ρ block 5′ or 3′	●●●
C/D	Split or merge ω blocks	●●●●●●
C/D	Split or merge ρ blocks	●●●●●●

details). In brief, the Markov chain is initiated using values taken at random from the priors. Each iteration of the chain one or more parameters are updated according to a proposal distribution, and the proposal is accepted with the acceptance probabilities specified in the next section. There are nine moves that can be proposed, each of which is visited with the relative probability specified in Table 3. This is known as a random sweep. Moves of type A and B (Table 3) are Metropolis-Hastings (Metropolis *et al.* 1953; Hastings 1970) moves that change a single parameter at a time. Moves of type C and D are complementary reversible-jump moves (Green 1995). For the purpose of illustration, I will describe one each of move types A-D, and assume that the prior on the ω_j's specifies i.i.d. exponential distributions with rate λ. The moves below describe in full how variation in ω along the sequence is explored by MCMC.

4.3.1 Type A. Change ω within a block

Metropolis-Hastings move

A block is chosen uniformly at random. A new value ω' is proposed so that $\omega' = \omega \exp(U)$ where U ~ Uniform(-1,1). Thus $\omega e^{-1} < \omega' < \omega e$. The acceptance probability is given by the Metropolis-Hastings ratio

$$\alpha(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{P(\Theta')}{P(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\},$$

where $K(\Theta \to \Theta')$ is the proposal kernel density. To find *K*, note that

$$\Pr(U < u) = \frac{1}{2}(1+u), \ -1 < u < 1.$$

So

$$\begin{aligned} \Pr(\omega' < x) &= \Pr(\omega e^{U} < x) \\ &= \Pr\left(U < \ln\frac{x}{\omega}\right) \\ &= \frac{1}{2}\left(1 + \ln\frac{x}{\omega}\right). \end{aligned}$$

Therefore

$$\begin{aligned} P(\omega' = x) &= \frac{\partial}{\partial x} \Pr(\omega' < x) \\ &= \frac{1}{2x}. \end{aligned}$$

This gives an acceptance probability of

$$\alpha_A(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \exp\{-\lambda(\omega' - \omega)\} \frac{\omega'}{\omega}\right\}. \qquad (22)$$

4.3.2 Type B. Extend an ω block 5′ or 3′

Metropolis-Hastings move

The block to extend is chosen uniformly at random, and for each block the direction is chosen with equal probability. If the 5′-most or 3′-most block is chosen to be extended 5′ or 3′ respectively, the move is rejected. The number of sites to extend the block, $g \in [1, \infty)$ is chosen from a geometric distribution with some parameter. If extending the block *g* sites in the chosen direction would cause it to merge with the adjacent block, the move is rejected.

The proposal distribution is symmetric, so the Hastings ratio is one. The ratio of priors is also one because the prior on the positions of the transition points is uniform. Therefore

$$\alpha_B(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)}\right\}. \tag{23}$$

4.3.3 Types C and D. Split and Merge an ω block

Reversible Jump moves

The acceptance probability for a reversible jump move (Green 1995) is

$$\alpha_m(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{P(\Theta')}{P(\Theta)} \frac{j_m(\Theta')}{j_m(\Theta)} \frac{g'_m(\mathbf{U}')}{g_m(\mathbf{U})} \left|\frac{\partial(\Theta', \mathbf{U}')}{\partial(\Theta, \mathbf{U})}\right|\right\}.$$

Here $j_m(\Theta)$ is the probability of proposing move *m* when at state Θ, and $g_m(\mathbf{U})$ is the joint probability density of the random vector **U** which is generated to facilitate

the transformation from $(\Theta, \mathbf{U})$ to $(\Theta', \mathbf{U}')$. The last term in the acceptance probability is the determinant of the Jacobian of the diffeomorphism (the transformation which must be differentiable in both directions).

4.3.3.1 Ratio of priors

In move C a block that currently has length $(s_{j+1} - s_j)$ is split at position *s**, and its current selection parameter ω_j is transformed, with the aid of a random variable *U*, into two new parameters ω'_j and ω'_{j+1}. The ratio of priors is

$$\frac{p_\omega}{(1-p_\omega)} \lambda \exp\{-\lambda(\omega'_j + \omega'_{j+1} - \omega_j)\}.$$

In move D two adjacent blocks that currently have lengths $(s^* - s_j)$ and $(s_{j+1} - s^*)$ are merged, and their selection parameters ω_j and ω_{j+1} are transformed into a single parameter ω'_j. So the ratio of priors is

$$\frac{(1-p_\omega)}{p_\omega \lambda} \exp\{-\lambda(\omega'_j - \omega_j - \omega_{j+1})\}.$$

4.3.3.2 Ratio of proposal probabilities

Move C splits an existing block. When there are $(B_\omega + 1)$ blocks there are $(L - B_\omega - 1)$ possible positions at which a block could be broken. The position of the split, *s**, is chosen uniformly at random from these. Move type C_i splits the block that spans position *i*; only $(L - B_\omega - 1)$ out of the total possible $L - 1$ type C moves are

available at any one time. So $j_{C_i}(\Theta) = c_B/(L - B_\omega - 1)$, where c_B is the total rate at which type C moves are proposed when there are $(B_\omega + 1)$ blocks.

Move D merges two adjacent blocks. Assuming that the block merges with its 3′ neighbour, there are B_ω possible mergers. The merger is chosen uniformly at random from these B_ω possibilities. So $j_{D_i}(\Theta) = d_B/B_\omega$, where d_B is the total rate at which type D moves are proposed when there are $(B_\omega + 1)$ blocks.

Following Green (1995), when there are B_ω transition points, moves C and D are proposed with relative probabilities c_B and d_B, where

$$\frac{c_B}{d_B} = \frac{\min\{1, P(B_\omega + 1)/P(B_\omega)\}}{\min\{1, P(B_\omega - 1)/P(B_\omega)\}}.$$

Under the prior, the number of transition points B_ω is distributed binomially. This yields

$$\frac{\Pr(B_\omega + 1)}{\Pr(B_\omega)} = \frac{(L - B_\omega - 1)}{(B_\omega + 1)} \frac{p_\omega}{(1 - p_\omega)} \text{ and } \frac{\Pr(B_\omega)}{\Pr(B_\omega - 1)} = \frac{B_\omega}{(L - B_\omega)} \frac{(1 - p_\omega)}{p_\omega}.$$

4.3.3.3 Ratio of density functions

In transforming ω_j to ω'_j and ω'_{j+1}, it is necessary to introduce a random deviate U to match the dimensionality on both sides. So the transformation $(\omega_j, U) \to (\omega'_j, \omega'_{j+1})$ involves the generation of a random deviate U in move C, but not in the inverse move D. This simplifies $g_D(\mathbf{U}')/g_C(\mathbf{U})$ to $1/g_C(U)$. Since U is chosen uniformly on $(0,1)$, this ratio equals one.

4.3.3.4 Jacobian

In Move C the values of the selection parameters for the two resulting blocks, ω'_j and ω'_{j+1} are chosen from the current value of ω_j so that the weighted geometric mean is preserved. The weighting takes into account the relative sizes of the two resulting blocks, which are $(s*-s_j)$ and $(s_{j+1}-s*)$ respectively. Thus

$$\omega_j'^{(s*-s_j)}\omega_{j+1}'^{(s_{j+1}-s*)}=\omega_j^{(s_{j+1}-s_j)}.$$

To introduce a random element,

$$\frac{\omega'_{j+1}}{\omega'_j}=\frac{1-U}{U},$$

where U ~ Uniform(0,1). The determinant of the Jacobian is,

$$J=\begin{vmatrix}\dfrac{\partial\omega'_j}{\partial\omega_j} & \dfrac{\partial\omega'_{j+1}}{\partial\omega_j}\\ \dfrac{\partial\omega'_j}{\partial U} & \dfrac{\partial\omega'_{j+1}}{\partial U}\end{vmatrix},$$

To obtain J, it is necessary to express ω'_j and ω'_{j+1} in terms of ω_j and U, giving

$$\omega'_j=\omega_j\left(\frac{U}{1-U}\right)^{1-a}$$

and

$$\omega'_{j+1}=\omega_j\left(\frac{1-U}{U}\right)^{a},$$

where $a=(s*-s_j)/(s_{j+1}-s_j)$. The determinant of the Jacobian (which is defined to be always positive) comes out as

$$J=\frac{(\omega'_j+\omega'_{j+1})^2}{\omega_j}.$$

4.3.3.5 Acceptance probabilities

For move C,

$$\alpha_C(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{p_\omega \lambda \mathrm{e}^{-\lambda(\omega'_j + \omega'_{j+1})}}{(1-p_\omega)\mathrm{e}^{-\lambda(\omega_j)}} \frac{d_{B_\omega+1}(L - B_\omega - 1)}{c_{B_\omega}(B_\omega + 1)} \frac{(\omega'_j + \omega'_{j+1})^2}{\omega_j}\right\}. \quad (24)$$

For move D,

$$\alpha_D(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{(1-p_\omega)\mathrm{e}^{-\lambda(\omega'_j)}}{p_\omega \lambda \mathrm{e}^{-\lambda(\omega_j + \omega_{j+1})}} \frac{c_{B_\omega - 1} B_\omega}{d_{B_\omega}(L - B_\omega)} \frac{\omega'_j}{(\omega_j + \omega_{j+1})^2}\right\}. \quad (25)$$

Table 4 Structure of the *omegaMap* program

File	Function	# Lines
main.h	Header file for main.cpp	6
omegaMap.h	Header file for omegaMap.cpp	361
main.cpp	Program control	30
omegaMap.cpp	Read in command line and configuration file options. Allocate memory. Initialize the MCMC chain.	1164
likelihood.cpp	Calculate the likelihood. Forward and backward algorithm. Build the mutation rate matrix.	726
mcmc.cpp	Controls the MCMC scheme. Proposes moves. Calculates acceptance probabilities.	1514
io.cpp	Outputs MCMC chain in text format and encoded format. Functions for reading in MCMC chain from encoded format.	504

Table 5 Utilities used by *omegaMap*

File	Function	# Lines
argumentwizard.h	Utility for reading in command line options.	215
controlwizard.h	Utility for reading in configuration files.	659
dna.h	Functions for reading in FASTA files and storing DNA sequences.	486
lotri_matrix.h	Lower triangular matrix class.	144
matrix.h	Matrix class.	226
myerror.h	Error and warning functions.	33
myutils.h	Links these various utility files.	35
random.h	Random number generation.	520
utils.h	Various utilities.	29
vector.h	Vector class.	133

Table 6 *PAML* package, linked to by *omegaMap*

File	Function	# Lines
paml.h	Header file for tools.c	335
tools.c	PAML functions	4369

PAML was written by Ziheng Yang and is available from

http://abacus.gene.ucl.ac.uk/software/paml.html

4.3.4 Implementation

I implemented the likelihood calculation and inference scheme in C++. The program, called *omegaMap*, was built up progressively, from testing the likelihood function on simple examples that could be verified using a calculator, to a Metropolis-Hastings MCMC scheme without variation in ω and ρ, to the full reversible-jump MCMC scheme. The code was developed using *Microsoft Visual C++* and then switched to Linux *gcc* for testing on datasets of realistic size. The MCMC scheme was debugged principally by using a flat likelihood, in which case one expects to recover the prior from the posterior. This proved important when, having moved from a dual-node 64-bit AMD machine (mcv1@stats.ox.ac.uk) I recompiled the program on a multi-node 64-bit AMD machine (genecluster@stats.ox.ac.uk), the posterior began to produce a systematic bias in the recombination rate estimates, so that rates declined 5′-3′, even when the same sequence was reversed. Using a flat likelihood revealed that there was a numerical inconsistency, probably caused by a difference in compilers on the two machines. The problem was solved in a makeshift fashion by running the executable compiled on mcv1 on genecluster. This was a compromise because the executable compiled on mcv1 ran somewhat slower on genecluster than the executable compiled on genecluster. This is a cause for concern because the expectation is that C++ code is portable between machines and compilers. As a result when the code is distributed I will stress the need to test the program by compiling it first with flat likelihoods (which can be done using the flag `-D _TESTPRIOR`) and ensuring the prior is recovered from the posterior.

Table 7 Structure of the *analyse* program

File	Function	# Lines
analyse.h	Header file for analyse.cpp	45
main.cpp	Program control	73
analyse.cpp	Functions for reconstructing the MCMC chain based on an encoded file.	411

Tables 4-6 show the structure of the *omegaMap* program. In total there are 6,785 lines of novel code (Tables 4 and 5). *omegaMap* uses some functions in the *PAML* package (Table 6), written by Ziheng Yang. *PAML* (Phylogenetic Analysis by Maximum Likelihood) is freely available from http://abacus.gene.ucl.ac.uk/software/paml.html. In addition, many functions in the C++ standard template library are used, so the total size of the code is unknown. *omegaMap* can output the results in two formats. The first is a tab-delimited text file with a column for each parameter in the model and a number of other diagnostics such as the acceptance probability and computational time. The thinning interval dictates the number of iterations before the parameter state is output. This text file can be read by software such as *R* or *Excel*. However, outputting the entire MCMC chain using a thinning interval of one creates an enormous text file with a great deal of redundancy because only a subset of the parameters are changed in any iteration. Therefore *omegaMap* can output in a second format, an encoded version of the MCMC chain. The program *analyse* (Table 7) can read this file, reconstruct the MCMC chain internally (orders of magnitude faster than the original MCMC chain was generated) and output a text file for use with *R* or *Excel*.

4.4 Simulation study

To investigate the performance of the method, I undertook two simulation studies. In one data was simulated with variation in the selection parameter along the sequence, and a constant recombination rate. In the other, data was simulated with variation in the recombination rate along the sequence, and a constant selection parameter. Each study consisted of simulating 100 datasets of $n = 20$ sequences each of length $L = 200$ codons using the coalescent with recombination (Hudson 1983, Griffiths and Marjoram 1997) and the NY98 mutation model. Every simulated dataset was analysed twice, using 250,000 iterations of the MCMC and a burn-in of 20,000 iterations. Initial values were chosen randomly from the priors independently for the two runs. The runs were compared for convergence and merged to obtain the posterior distributions.

4.4.1 Permutation test for recombination

Before the datasets were analysed, each was subjected to a permutation test for recombination (McVean *et al.* 2001; Meunier and Eyre-Walker 2001). Phylogenetic analysis is inappropriate for gene sequences taken from populations that are demonstrably recombinogenic. The aim of the permutation tests was to demonstrate the recombinogenic nature of the data.

The permutation test is a goodness-of-fit test for the model of no recombination. When there is no recombination, there ought to be no correlation between physical distance and LD, so sites are exchangeable. It should be noted that sites are also exchangeable in the case of complete linkage equilibrium. If LD tails off with

physical distance then recombination must have occurred in the ancestral history of the sequences. The test proceeds as follows

1. The observed correlation between a measure of LD and physical distance is recorded as c_{obs}.
2. The nucleotide positions are reordered at random and the correlation between LD and physical distance is calculated.
3. Step 2 is repeated 999 times.

Three measures of LD can be used: r^2 (Hill and Robertson 1968), D' (Lewontin 1964) and the four-gamete test (*G*4; Hudson and Kaplan 1985). In section 2.3.2 cor(r^2,*d*), where *d* is physical distance, was used for testing the goodness-of-fit of the standard neutral coalescent. If c_{obs} lies in the tail of the reference distribution then the model of exchangeability of sites is not a good fit to the data, and we can conclude that there is good evidence for recombination in the data. The probability of obtaining a result as extreme as observed under the model can be expressed as a *p* value, where *p* is estimated to be

$$p = \frac{n+1}{N+1}$$

(Sokal and Rohlf 1995). Here *n* is the number times a value more extreme than c_{obs} was observed out of a total of *N* simulations.

Using *p* values to reject a "null" model might seem to be a particularly frequentist thing to do. In fact a frequentist *p* value and a Bayesian posterior predictive *p* value (Rubin 1984) are equivalent in the model of exchangeability described here, because the model has no parameters. I will discuss the use of posterior predictive *p* values for goodness-of-fit testing more in chapter 5.

4.4.2 Simulation study A

This study was designed to simulate data with variation in ω but not in ρ. I varied ω between 0.1 and 10, as shown by the red line in Figure 5a. I created more fine detail in variation in ω for $\omega > 1$ because, biologically, a scenario in which there is an excess of non-synonymous relative to synonymous polymorphism is of greater interest. For the same reason ω is plotted on a natural, rather than a logarithmic scale. The mutation parameters were set at $\mu = 0.7$ and $\kappa = 3.0$, which gives $\theta_S = 0.1$. The recombination rate was set constant at $\rho = 0.1$, giving a total recombination distance for the region of $R = \sum \rho = 19.9$. The mutation and recombination parameters were chosen to mimic those estimated for the housekeeping genes of *Neisseria meningitidis* (see Chapter 1). Exponential distributions were used for the priors on μ, κ, ω and ρ, with means 0.7, 3.0, 1.0 and 0.1.

Permutation tests showed that phylogenetic analysis of these datasets was inappropriate because of the presence of recombination. The number of datasets for which the *p*-value was less than 0.05 was 99, 93 and 93 for the three measures of LD (r^2, D' and *G*4) respectively.

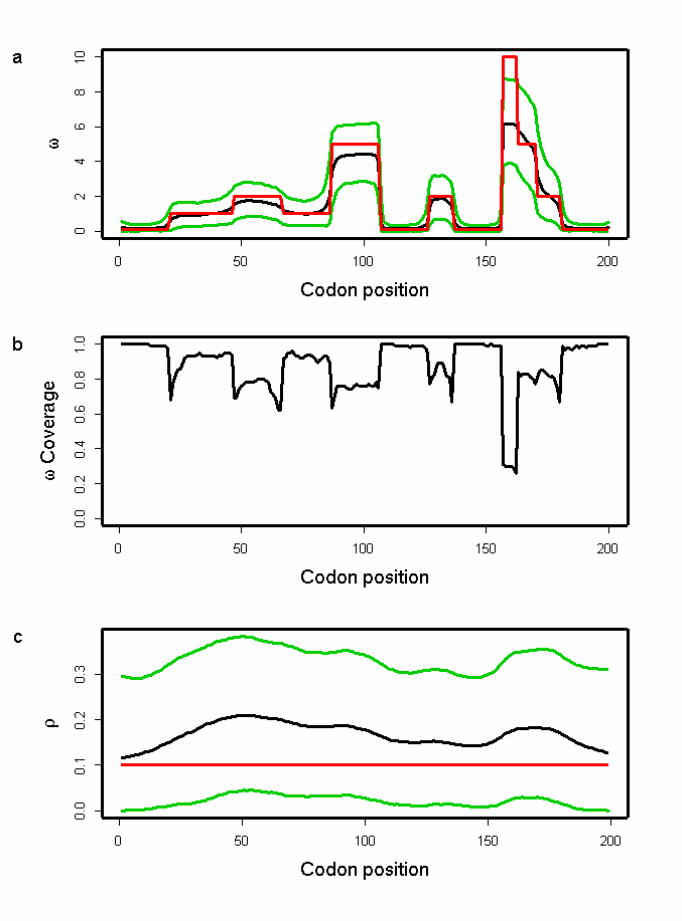


Figure 5 Results of simulation study A. (a) Average posterior of ω, (b) coverage of ω and (c) average posterior of ρ. In (a) and (c) the red line indicates the truth, the black line indicates the average mean of the posterior and the green lines indicate the average 95% HPD interval of the posterior. The averages are taken over 100 simulated datasets. In (b) coverage is defined as the proportion of the 100 datasets for which the 95% HPD interval encloses the truth.

Figure 5a shows the average over the 100 simulated datasets of the mean and 95% highest posterior density (HPD) interval for the posterior distribution of ω at each site. The average mean posterior density follows the truth closely. Likewise the average 95% HPD interval generally encloses the true value of ω. As expected, the effect of fitting a prior with mean 1 was to cause the posterior to underestimate ω when $\omega > 1$

Table 8 Summary of posteriors for simulation study A

Parameter	Truth	Prior	Average posterior			Coverage
		Mean	Lower 95% HPD	Mean	Upper 95% HPD	
μ	0.7	0.7	0.7	0.9	1.1	0.63
κ	3.0	3.0	2.3	3.1	3.9	0.91
R	19.9	19.9	22.4	33.3	44.7	0.43

and overestimate ω when $\omega < 1$. The effect is not great except for the most extreme values where $\omega = 10$.

However, even where the average 95% HPD interval encloses the truth, that does not mean the 95% HPD interval encloses the truth for all simulated datasets. Figure 5b shows the relevant quantity, the coverage of ω, for each site. Coverage is defined here as the proportion of datasets for which the 95% HPD interval encloses the truth. Half of sites have coverage better than 93%, and 95% of sites have coverage better than 66%. If a false positive is defined as the lower bound of the 95% HPD interval exceeding 1 when in truth $\omega \leq 1$, then the false positive rate was 0.5%. The estimate of the synonymous transversion rate μ exhibits upward bias (average 0.90), with 63% coverage (Table 8), and the transition-transversion ratio κ is estimated to be 3.1 on average, with 91% coverage.

Consistent with the findings of Li and Stephens (2003), I observed that the recombination rate estimator has a small upward bias (Figure 5c). The average mean posterior is almost flat, and the average 95% confidence intervals enclose the truth completely, suggesting that the estimator is good notwithstanding its bias. The

Table 9 MCMC Moves Acceptance Probabilities

Type	Move	Mean acceptance probability α
A	Change μ	0.139
A	Change κ	0.157
A	Change ω within a block	0.573
A	Change ρ within a block	0.727
B	Extend an ω block 5′ or 3′	0.403
B	Extend an ρ block 5′ or 3′	0.825
C	Split an ω block	0.381
D	Merge ω blocks	0.242
C	Split a ρ block	0.635
D	Merge ρ blocks	0.660

coverage is almost constant across sites at 95%. Table 8 shows that the estimate of the total recombination distance, *R*, is also upwardly biased. Coverage of *R*, however, was only 43%, suggesting that the good coverage for ρ at individual sites may be in part because of poor information. Importantly, Figures 5a and 5b show that the effect of the selection parameter on the estimate of ρ is negligible, indicating that inference on ρ is not confounded by ω.

4.4.3 Mixing properties of reversible jump moves

Achieving satisfactory acceptance probabilities can be an issue in reversible-jump MCMC (Green 1995). This was not found to be a problem in the MCMC scheme

presented here. For illustrative purposes, Table 9 shows the acceptance probabilities for the MCMC moves, averaged over a pair of independent analyses of the same dataset from simulation study A. The reversible-jump moves (those of type C or D) had high acceptance probabilities (for example, $\alpha = 0.381$ when splitting an ω block and $\alpha = 0.242$ when merging ω blocks). Of the other moves, acceptance probabilities ranged from 0.139 to 0.825. The lowest acceptance probabilities were for moves changing μ and κ ($\alpha = 0.139$ and 0.157 respectively), perhaps because these changes affect all sites in the sequence unlike any other move. Changes to moves involving ρ had high acceptance probabilities ($\alpha = 0.635$ to 0.825), which may be indicative of the low information regarding variation in recombination rate within the region.

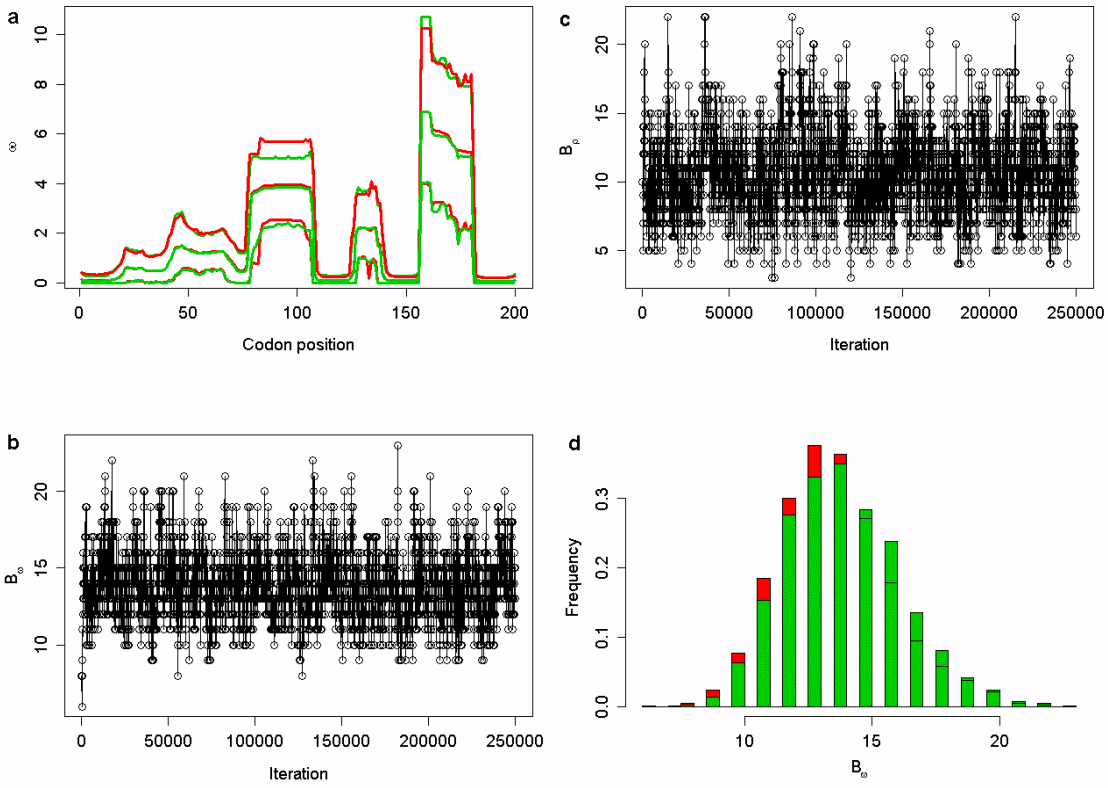


Figure 6 a Convergence of the mean and upper and lower 95% HPD bounds of the posterior on ω for two analyses (red and green lines) of the same dataset from simulation study A. **b** Trace of B_ω for one of the two analyses. **c** Trace of B_ρ for one of the analyses. **d** Convergence of the posterior distribution of B_ω for the two analyses (red and green histograms).

In Figure 6 the mixing properties of the two chains for the same dataset are shown. Figure 6a shows the convergence of the two chains for the posterior distribution on ω across sites. The mean and upper and lower 95% HPD bounds are indicated. One chain is plotted in red, the other in green. The agreement is good; more so for the mean than the 95% HPD bounds. One would expect estimates of the latter to have greater variance. Figure 6b is a trace of B_ω through iterations of one of the Markov chains, and 6c is the corresponding trace of B_ρ. B_ω and B_ρ can only be changed by reversible-jump moves. There is no evidence of poor mixing in either of the traces. Figure 6d shows a histogram of the posterior distribution of B_ω for both the chains

(one in red, the other green). The two appear to converge well throughout the distribution. When the chains are merged the variance in the estimate of the posterior will be reduced. However, if this were an analysis of a real dataset of special interest, rather than one of a hundred simulated datasets, then there is some argument for running the two chains longer to further improve convergence.

4.4.4 Simulation study B

This study was designed to simulate data with variation in ρ but not in ω. Along the sequence ρ was allowed to vary at 0.005, 0.1, 0.5 and 1, for which one would expect 0.018, 0.35, 1.8 and 3.5 recombination events respectively per site in the ancestral history under a coalescent model (Griffiths and Marjoram 1997). The total recombination distance was $R = 37.5$. I let $\mu = 3.6$ and $\kappa = 3.0$ giving $\theta_S = 0.5$, and a constant selection parameter of $\omega = 0.2$. Exponential distributions were used for the priors on μ, κ, ω and ρ, with means 3.6, 3.0, 1.0 and 0.2.

Permutation tests showed that these datasets were not amenable to phylogenetic analysis because of the presence of recombination. All 100 datasets yielded *p*-values less than 0.05 for all three measures of LD.

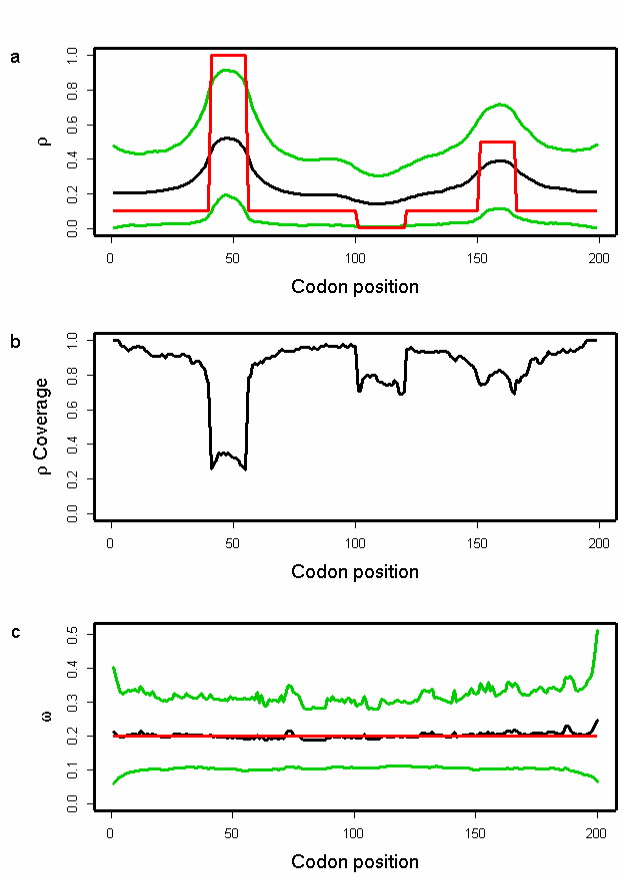


Figure 7 Results of simulation study B. (a) Average posterior of ρ, (b) coverage of ρ and (c) average posterior of ω. In (a) and (c) the red line indicates the truth, the black line indicates the average mean of the posterior and the green lines indicate the average 95% HPD interval of the posterior. The averages are taken over 100 simulated datasets. In (b) coverage is defined as the proportion of the 100 datasets for which the 95% HPD interval encloses the truth.

Variation in the recombination rate was detected by the new method, as seen in Figure 7a. The average over the 100 datasets shows that the mean and 95% HPD interval for the posterior distribution of ρ at each site pick up the rate variation, but not to the full extent. As a result, the coverage shown in Figure 7b is generally good, on average 85%, but performs worst for the most extreme peak in rate between sites 41 and 55, where it consistently underestimates the height. The properties of the estimate of the

Table 10 Summary of posteriors for simulation study B

Parameter	Truth	Prior	Average posterior			Coverage
		Mean	Lower 95% HPD	Mean	Upper 95% HPD	
μ	3.6	3.6	3.4	4.2	5.1	0.53
κ	3.0	3.0	2.5	3.1	3.8	0.95
R	37.5	39.8	37.4	50.9	65.0	0.49

total recombination distance R (Table 10) are similar to those in simulation study A. There is a tendency to overestimate (average 50.9) and as a result coverage is 49%. This bias could be corrected empirically, as in Li and Stephens (2003). Nevertheless, there is power to detect rate variation on such fine scales. The extent to which the posteriors underestimate the deviations from the mean recombination rate reflects the constraining effect of the prior when the signal in the data is weak.

Figure 7c shows that on average the estimates of ω are very close to the truth, with the average 95% HPD intervals completely enclosing the true value. Along the sequence, the estimates are flat, with mean 0.21 and coverage 90%. The false positive rate was zero. Reflecting simulation study A, there was no evidence that variation in the recombination rate confounded inference on the selection parameter. Table 10 shows that there was some upward bias in the mean estimate of $\mu = 4.1$, with 58% coverage, and the transition-transversion ratio was estimated to be 3.2 on average, with 89% coverage. Most importantly, both simulation studies show that when there is variation in ω or ρ it can be detected, when there is no variation none is detected, and there is little or no confounding between ω and ρ.

4.5 Summary

In this chapter I have described a new model for detecting immune selection in nucleotide sequences, based on an approximation to the coalescent. The model uses the NY98 codon model of molecular evolution which incorporates the ratio of non-synonymous to synonymous substitution, dN/dS. Values of dN/dS less than one are interpreted as purifying selection imposed by functional constraint and values greater than one are interpreted as diversifying selection imposed by interaction with the host immune system. Those sites under strong diversifying selection are predicted to be the major determinants of immunogenicity for the gene product. In order to exploit information about the underlying tree structure and mode of selection at sites segregating for insertions/deletions, I have described a simple extension to the NY98 mutation model. I have proposed a model for the variation in the dN/dS ratio and recombination rate along a sequence and a reversible-jump MCMC scheme for exploring that variation. The primary aim of the Bayesian inference framework described is to obtain a posterior distribution for the dN/dS ratio and recombination rate for every site along the sequence, but the underlying mutation rate, transition:transversion ratio and rate of insertion/deletion are also estimated. Finally I performed simulation studies to assess the performance of the inference method for two caricatures of variation in the dN/dS ratio and recombination rate. The method was found to have good coverage for the dN/dS ratio, but some upward bias in estimates of the recombination rate, in agreement with previous work. Most importantly, the simulation studies showed that when there is variation in the dN/dS ratio or recombination rate it can be detected, when there is no variation none is

detected, and there is little or no confounding between dN/dS and the recombination rate.

In the next chapter I will apply the new method to the *porB* locus of *N. meningitidis*, which encodes the antigenic PorB outer membrane protein. I will give a brief background to *porB* and the results of previous phylogenetic estimates of variation in the dN/dS ratio at the locus. In order to verify the conclusions of the *porB* analysis with the new method, I will apply a variety of model criticism techniques including prior sensitivity analysis and goodness-of-fit testing. Goodness-of-fit testing requires datasets to be simulated under the new model, so I will describe how to do that. I will briefly investigate the effect of violating the coalescent assumption of random sampling by comparing datasets of *porB* that represent a random and non-random sample. Finally, I will compare the results of the new method to previous phylogenetic methods to look for evidence of false positives caused by the assumption of no recombination.

Chapter 5

Evidence for Immune Selection in an Antigen of *Neisseria meningitidis*

Using the *porB* locus of *Neisseria meningitidis*, this chapter demonstrates the application of the Bayesian approach to inferring selection and recombination introduced in Chapter 4. PorB is ideal for exploring the new method, because it is a strongly immunogenic, constitutively expressed outer membrane protein whose molecular structure has been elucidated. Previous studies have analysed *porB* and I will compare the results of my method to those published before. In particular, I will compare the results to those of phylogenetic-based analyses for evidence of false positives introduced by the assumption of no recombination. Using *porB* as an example, I will outline a coherent approach to model-based analysis, from rejection of a model with no recombination through to prior sensitivity analysis and model criticism. Using different datasets there is the opportunity to informally study the effect of violating the coalescent assumption of random sampling. I will also contrast the patterns of variation in selection pressure in *porB* to those in the seven MLST housekeeping loci.

5.1 Analysis of the *porB* locus

PorB is a porin expressed on the surface of the meningococcus, and thought to be important both for proper cell growth and pathogenesis. There exist two classes of

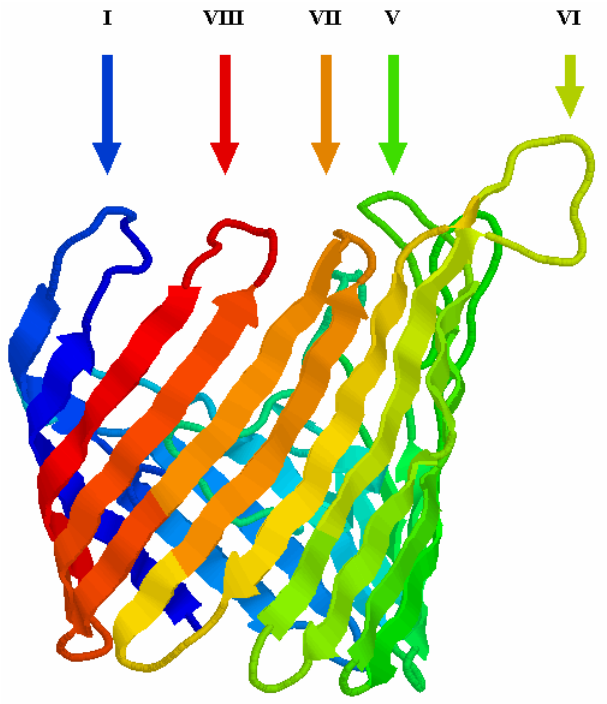


Figure 1 Molecular structure of the *Neisseria meningitidis* class 3 outer membrane protein, PorB3. The molecule is colour-coded from the N (blue) to the C (red) terminus. The molecule spans the outer membrane, with eight exposed variable loop regions, of which five are marked (I, V, VI, VII and VIII). This image was generated using Protein Explorer (Martz 2002). The molecular structure was determined by Derrick *et al.* (1999). Jeremy Derrick kindly provided the molecular structure file.

PorB protein with somewhat different molecular structure and evolutionary ancestry (Smith *et al.* 1995; Derrick *et al.* 1999), called PorB class 2 and PorB class 3, or PorB2 and PorB3. These classes are defined on the basis of sequence homology and immunological properties (Smith *et al.* 1995). PorB is an important outer membrane protein (OMP) which is expressed constitutively at high levels, and is strongly immunogenic; epitopes of PorB define the serotypes of *N. meningitidis*. The molecular structure of meningococcal PorB comprises eight highly variable surface-exposed loop regions (I–VIII), consisting mainly of hydrophilic amino acid residues, between nine highly conserved membrane-spanning β-sheets. Figure 1 shows the

molecular structure of the PorB3 molecule (Derrick *et al.* 1995). The molecule is oriented with the surface-exposed regions at the top, and is colour-coded from the N (blue) to the C (red) terminus. Five of the eight loop regions are indicated by arrows.

5.1.1 Previous analyses

PorB is encoded by the *porB* locus. There have been several studies into the influence of natural selection on the genetic diversity of *porB2* and *porB3* alleles (Smith *et al.* 1995; Urwin *et al.* 2002). In an analysis based on 5 *porB2* sequences and 4 *porB3* sequences, Smith *et al.* (1995) counted the number of synonymous and non-synonymous differences to estimate that the relative rate of synonymous and non-synonymous change was dN/dS = 1.2 for *porB2* and dN/dS = 0.62 for *porB3* across the sequences as a whole. These values represent an average over the conserved and variable regions. Loop regions exhibited an elevated number of non-synonymous relative to synonymous substitutions; estimated to be 2.3 compared with 0.28 for non-loop regions, averaged over the *porB* classes. As discussed in Chapter 4, an elevated rate of non-synonymous change is indicative of relaxed functional constraint, or positive selection for variation in the amino acid sequence (diversifying selection). Evidence for diversifying selection in the surface-exposed loop regions is consistent with the hypothesis that the immune system exerts a selection pressure for antigenic novelty in PorB. However, on the basis of these estimates, that selection pressure is not enormous, which would suggest that PorB may be a less important vaccine constituent than other OMPs, for example PorA (Urwin *et al.* 2002).

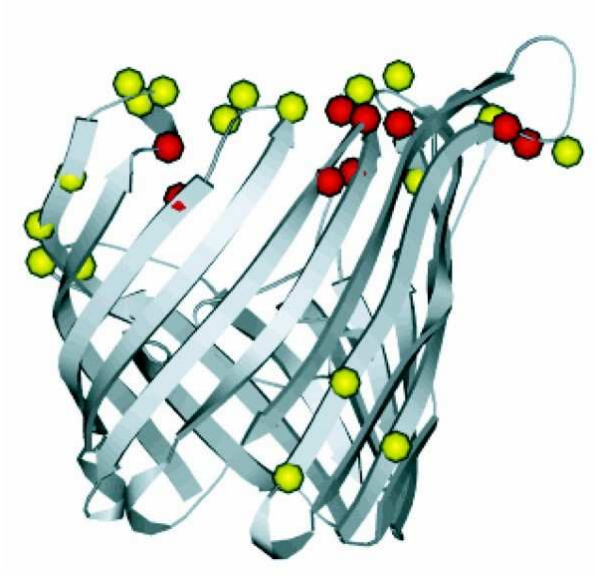


Figure 2 Location of sites under weak (yellow circles) and strong (red circles) positive selection in the PorB3 molecule, inferred using CODEML. All but 3 positively selected sites lie in the surface exposed loop regions, and seven out of the eight loops contain some sites under positive selection. Cf. Figure 1. Source: Urwin *et al.* (2002).

Urwin *et al.* (2002) used a maximum-likelihood method (Yang *et al.* 2000; described in section 4.1.2.1) implemented in the CODEML program of the PAML package (Yang 1997) to infer selection in the *porB* locus, taking the *porB2* and *porB3* allelic classes separately. The aim of the analysis was to improve the estimates of the dN/dS ratio by using many more sequences, applying a likelihood-based model for the ancestry of the sequences, and obtaining an estimate for each site along the amino acid sequence. They found evidence for extremely high selection pressures in *porB* surface exposed loops, as great as in HIV-1 surface glycoproteins, whilst the membrane spanning regions were under strong purifying selection. Based on the model of variation in selection pressure described in section 4.1.2.1, Urwin *et al.* (2002) estimated that in *porB2*, 94% of sites were under purifying selection (dN/dS = 0.067), 4.5% of sites were under weak positive selection (dN/dS = 4.2) and 1.1% of sites were under strong positive selection (dN/dS = 18.6). In *porB3*, 95% of

sites were under purifying selection (dN/dS = 0.033), 4.1% of sites were under weak positive selection (dN/dS = 3.2) and 0.7% of sites were under strong positive selection (dN/dS = 13.9). The likelihood ratio tests for positive selection were very highly significant for both *porB* classes. Figure 2 shows the location of sites identified as experiencing weak (yellow circles) or strong (red circles) positive selection in *porB3*. Almost all positively selected sites lie in the surface exposed loop regions, except for one weakly selected site between loops IV and V, and two between loops V and VI. There is evidence for some positive selection in all loops except III.

Urwin *et al.* (2002) also conducted an analysis of the frequency of recombination in *porB*. The method is similar to that of Holmes *et al.* (1999), described in Chapter 1. A maximum likelihood tree was estimated for each half of the *porB2* sequence. Phylogenetic incongruence was quantified as the difference in log likelihood, δ, between the maximum likelihood (ML) tree for the first half and the ML tree for the second half fit to the first (tree topology estimated from second half of the sequence and branch lengths estimated from the first half of the sequence conditional on the topology). Two hypotheses were tested: Firstly, the hypothesis that there is no recombination so that the topology is the same in both halves of the sequence. Secondly, the hypothesis that there is so much recombination that the topology estimated for the second half of the sequence is no better a fit to the first half of the sequence than a random topology. Both hypotheses were rejected for *porB2*, and for *porB3*, indicating that there is frequent recombination but not to the extent that the phylogenetic signal is utterly obliterated. Recombination in *porB* creates multiple, correlated, evolutionary histories for different parts of the sequence. The problem of this for CODEML is that it can inflate the false positive rate for detecting positive

selection (Anisimova *et al.* 2003; Shriner *et al.* 2003). The aim of the new method presented in Chapter 4, and implemented in the program omegaMap, is to co-estimate the dN/dS ratio and recombination rate along a sequence, with a more flexible model of variation in these parameters. In the sections that follow I apply the new method to the *porB3* sequences in order to visualise the variation in dN/dS and recombination rates, and demonstrate a coherent approach to testing the fit of the model so that it might be falsified. In the process I will compare the results to those of the CODEML analysis to look for potential false positives.

5.1.2 Isolates

For the analysis I used the 79 *porB3* alleles sequenced by Urwin *et al.* (2002). The 79 alleles were sequenced from an assorted collection of isolates including carriage and disease from around the world. As a result, the 79 alleles do not constitute a random sample of any population in a meaningful sense, thus violating one of the assumptions of the coalescent model. The effect of violating the coalescent assumption of random sampling is unknown. Therefore only a subset of the 79 alleles was used: 37 alleles sequenced from isolates obtained during a swabbing programme at a military recruit training camp. Nasopharyngeal swabs were taken from healthy recruits several weeks after arriving at the camp. The catchment area of the training camp was England and Wales.

Of the 37 isolates, 19 were obtained by repeatedly swabbing 5 of the carriers; the remaining 18 were sampled from one carrier each. In Chapter 1 a metapopulation model was described in which each host corresponded to a single deme. It was shown that when no more than one isolate is sampled from each host, the ancestry of the

Table 1 Permutation test for recombination

	Carriage study		Global study	
	Correlation	*p*	Correlation	*p*
r^2	-0.18	0.001	-0.15	0.001
D'	-0.24	0.001	-0.16	0.001
$G4$	-0.23	0.001	-0.15	0.001

isolates can be modelled as a coalescent process in which the effective population size is a function of several epidemiological parameters including the duration of infection and the primary and secondary infection rates. Therefore, the collection of 37 isolates was thinned to 23 so that each host was represented by only one isolate each. In this chapter I will refer to this sample of 23 alleles as the *carriage study*, and to the full collection of 79 alleles as the *global study*. Whereas the global study consisted of 77 unique haplotypes, the carriage study consisted of 12 unique haplotypes. Rachel Urwin kindly provided her sequence alignments for the analyses presented in this section.

5.1.3 Test for recombination

Urwin *et al.* (2002) have already rejected the hypothesis that there is no recombination in the *porB3* alleles by quantifying phylogenetic incongruence within the sequences. The permutation test described in section 4.4.1 agrees with this conclusion for both the carriage and global studies. Table 1 shows the results. For the carriage study there was a 0.1% probability of observing as extreme a correlation

between physical distance and LD under the model of no recombination, regardless of choice of LD statistic, and the result was the same for the global study. Although the two methods reach the same conclusion, the permutation test for exchangeability of sites is much faster than the phylogenetic incongruence test because it does not involve estimating tree topologies.

5.1.4 Codon frequencies

In the inference scheme presented in Chapter 4, the codon equilibrium frequencies are not treated as parameters but rather as known constants. Specifying the equilibrium frequencies of the 61 non-stop codons can be done in several ways. One can make the simplifying assumption that all codons have equal equilibrium frequency. This is not supported empirically, but the assumption had little effect on the analyses presented in this chapter (data not shown). One can take the observed codon usage in the data. However, for a small dataset there will be considerable sampling error. This is not recommended because in the worst case, a codon may not be observed in the sample. Then adopting this approach would be to exclude the possibility of the codon appearing anywhere in the ancestral history of the sample, which is very undesirable. A possible solution is to take the observed nucleotide frequencies, and assume that codon frequencies are proportional to the product of their constituent nucleotide frequencies, but this is not supported empirically. In the analyses that follow the observed codon frequencies (Nakamura *et al.* 2000) in the complete genomic sequence of *N. meningitidis* Z2491 serogroup A (Parkhill *et al.* 2000) were used, after removing the stop codons. There are 730,000 codons in the meningococcal genome, which is more than adequate to overcome sampling problems.

Table 2 Prior distributions

	Prior A	Prior B
μ	Exponential	Uniform
	mean 0.07	0 to 10
κ	Exponential	Exponential ratio
	mean 3.0	
φ	Exponential	Exponential
	mean 0.1	mean 1.0
ω	Exponential	Gamma
	mean 1.0	shape 2, scale 0.5
ρ	Exponential	Uniform
	mean 0.1	0 to 10

5.1.5 Priors

I chose to use exponential distributions for the priors on μ, κ, φ, ω and ρ. Because the robustness of the results to the prior will later be tested by proposing an alternative prior, the prior specified here will be referred to as Prior A (Table 2). All the model parameters can take on any positive value. However, it might be more natural to consider them on a logarithmic scale. For example, the strength of purifying selection when $\omega = 0.5$ might be interpreted as of an equal magnitude to the strength of positive selection when $\omega = 2$. This is because ω is the ratio of the rate of non-synonymous mutation to synonymous mutation. Under the null model of selective neutrality, the rate of non-synonymous and synonymous mutation might be thought of as i.i.d. random variables, and as a result ω or ω^{-1} are equally valid parameterizations. I chose

a mean of 1 for the prior on ω to represent the null model of selective neutrality. On a log scale, using the exponential distribution gives an approximately symmetric distribution. In fact the right tail decreases more rapidly than the left, so the prior favours less extreme values of ω in the direction of diversifying selection than in the direction of purifying selection.

Because of the natural multiplicative interpretation of the other model parameters, and because for practical reasons it was originally simpler to program only a single distributional form for the priors, exponential distributions were fit to the other parameters. Other distributions were implemented for prior sensitivity analysis. It should be noted that for ω and ρ, the reversible-jump MCMC scheme means that improper priors cannot be used because doing so forces the number of blocks to be 1 or L, ($L - 1$ in the case of ρ). The means of the priors were chosen based on analyses of *N. meningitidis* housekeeping loci (see Chapter 2). The mean of the prior on μ was 0.07, and the mean for κ was put at 3.0. The rate of insertion/deletion was given a mean of $\varphi = 0.1$. For ρ, the mean was set at 0.1.

The prior on the number of blocks for ω and ρ has already been described in section 4.2.5. The model for the variation in ω and ρ can be considered a prior if the block structure is considered to be a parameter. That is to say that B_ω, B_ρ, $\mathbf{s}^{(B_\omega)}$ and $\mathbf{s}^{(B_\rho)}$ are parameters and p_ω and p_ρ would be hyperparameters of the priors on B_ω and B_ρ. (See Chapter 4 Tables 1 and 2 for definitions of the notation.) If the block structure is considered to be missing data, then it is not a prior but a random effects model. The two are essentially equivalent in a Bayesian analysis. In order to specify that the average length of a block would be 10% of the sequence length ($L = 298$ codons), the

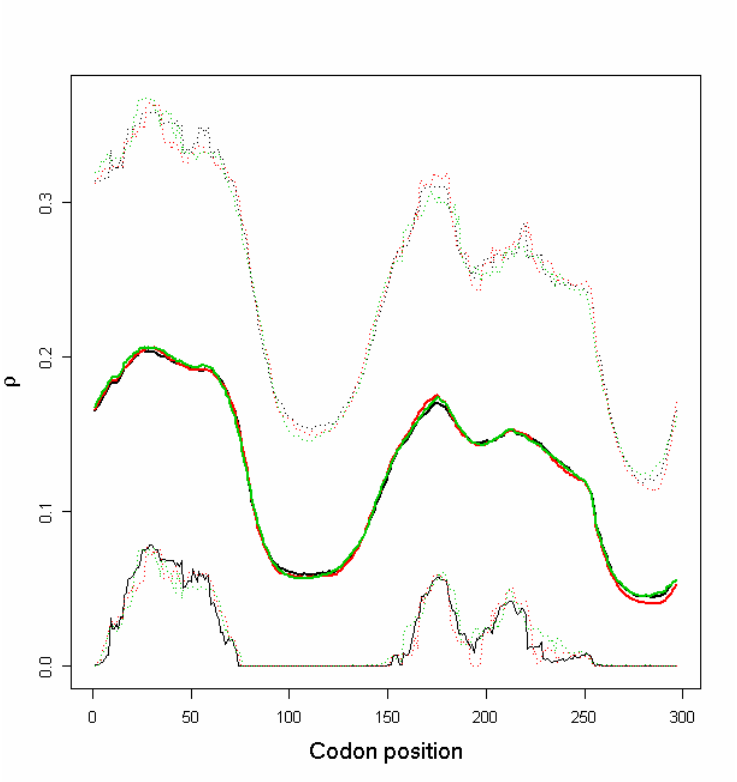


Figure 3 Convergence of the marginal posterior for ρ along the sequence for the three MCMC chains (red, green and black) after 500,000 iterations. The burn-in was 20,000 iterations. For each codon the marginal mean and upper and lower 95% HPD bounds are shown.

prior on the number of ω blocks was binomial with $p_\omega = 1/30$. Similarly, the prior on the number of ρ blocks was binomial with $p_\rho = 1/30$.

5.1.6 Results

Three MCMC chains were run each for 500,000 iterations. Convergence was judged informally by comparing the posteriors obtained from the three chains. Figure 3 illustrates the posterior distribution of ρ along the sequence for the three chains, which are colour-coded red, green and black (the mean and 95% HPD interval is shown for each codon). A burn-in of 20,000 iterations has been removed from the start of each

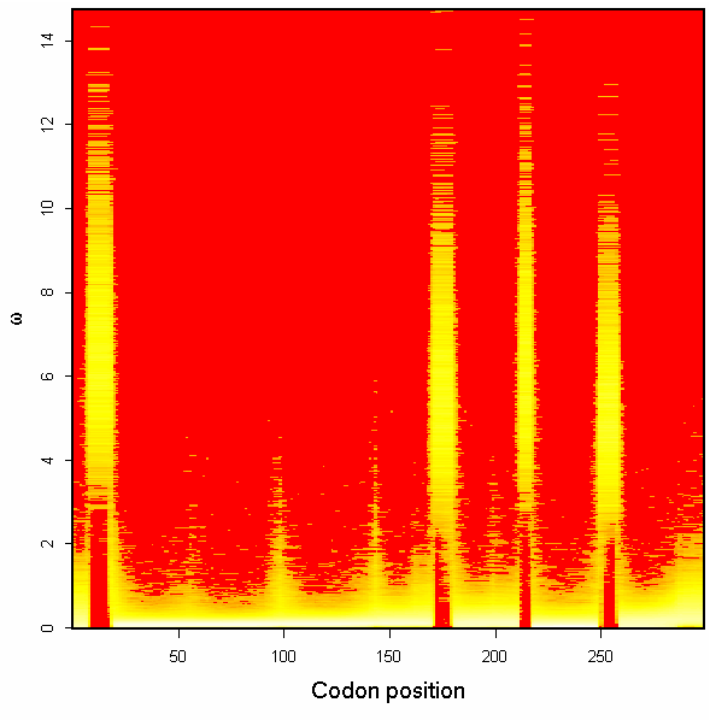


Figure 4 Fire-plot showing the site-wise posterior of ω in the *Neisseria meningitidis* carriage study. More intense colours (closer to white) represent high posterior density, and less intense (closer to red) low posterior density.

chain. Figure 3 shows that the posteriors are in close agreement. Convergence is always best for the mean compared to the HPD bounds. Once convergence has satisfactorily been established for all the parameters, the chains are merged (minus the burn-in), and the results of the merged chains are presented.

Figure 4 shows a fire-plot for the posterior distribution of ω at each site. More intense colours (closer to white) represent high posterior density, and less intense (closer to red) low posterior density. The structure of PorB3 consists of eight loop regions that extend out of the cell. Of these, there is clear and strong evidence for diversifying selection at four of the eight loops. In these loop regions the 95% HPD interval for the peak ω is (3.58, 9.76), (3.01, 8.92), (3.26, 9.68) and (2.58, 7.57) for loops 1, 5, 6 and 7 respectively. Taking the point estimate of ω at a site, $\hat{\omega}$, as the mean of the

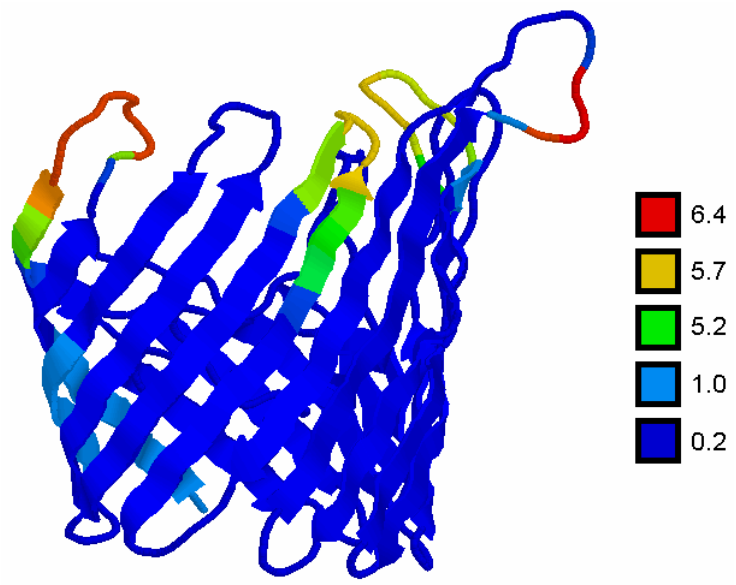


Figure 5 Molecular structure of PorB3 colour-coded according to $\hat{\omega}$, the mean of the posterior of ω at each site. Dark blue indicates strong functional constraint and red indicates strong diversifying selection. This image was produced using Protein Explorer (Martz 2002).

posterior distribution, then the average $\hat{\omega}$ for the sequence is 0.90. Excluding sites for which $\hat{\omega} > 1$, this drops to 0.16. So the majority of the sequence is under strong functional constraint, but four of the eight loop regions are under strong diversifying selection.

Superimposing $\hat{\omega}$ onto the three-dimensional structure of the PorB3 protein (Figure 5) illustrates the external position of loops 1, 5, 6 and 7. Because PorB3 is an outer membrane protein, these loops are especially exposed to the immune system, and are prime sites for recognition by antibody. It is striking that there is no evidence for diversifying selection outside the loops. Loops 2, 3 and 4 do not appear to be under diversifying selection; the three-dimensional structure suggests that they may be less exposed than the other loops. However, loop 8 is surprising because despite its prominent position (the dark blue loop second from left in Figure 5, cf. Figure 1),

Table 3 Posterior distributions

		Carriage study			Global study
		Prior A	Prior B	Prior A $\rho = 0$	Prior A
μ	mean	0.27	0.35	0.45	0.31
	95% HPD	(0.18, 0.36)	(0.23, 0.48)	(0.33, 0.58)	(0.22, 0.40)
κ	mean	3.61	3.09	3.69	3.34
	95% HPD	(2.38, 5.00)	(1.94, 4.24)	(2.69, 4.83)	(2.41, 4.33)
φ	mean	0.09	0.17	0.29	0.08
	95% HPD	(0.02, 0.19)	(0.03, 0.37)	(0.08, 0.56)	(0.02, 0.16)
R	mean	37.7	46.8	-	78.0
	95% HPD	(27.2, 49.0)	(26.2, 75.0)	-	(61.6, 94.5)

there is very little support for diversifying selection between codons 280-295 (Figure 4). The light blue shading in Figure 5 occurs at the N and C termini, outside the nucleotide alignment I analysed. Therefore they have been assigned the mean of the prior, $\hat{\omega} = 1$.

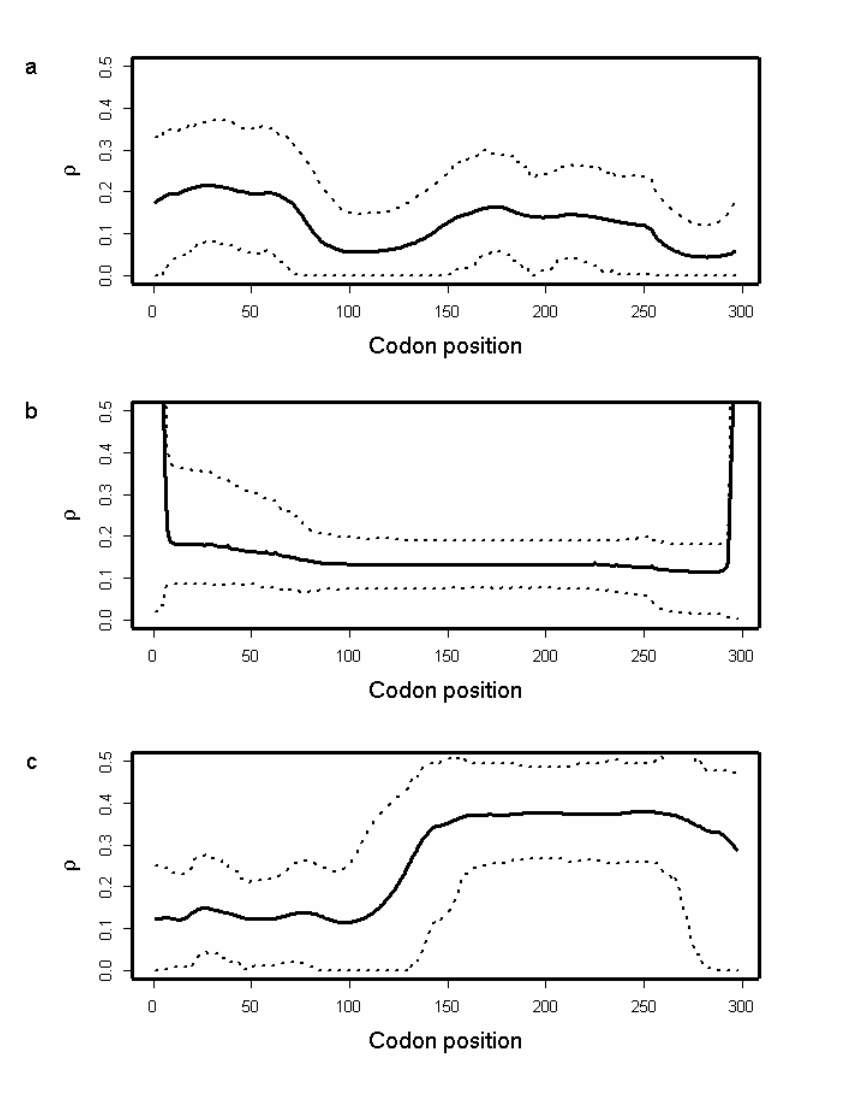


Figure 6 Posterior distribution of ρ in the carriage and global studies. The mean (solid line) and 95% HPD interval (dotted lines) are shown for (a) the carriage study under prior A, (b) the carriage study under prior B and (c) the global study under prior A.

There was some evidence for variation in the recombination rate (Figure 6a). The posterior mean for the total recombination distance, $\hat{R} = 37.7$ (Table 3), was twice the prior mean of 19.9. The posterior on μ was very different to the prior ($\hat{\mu} = 0.27$), while there was little discrepancy for κ and φ ($\hat{\kappa} = 3.61$, $\hat{\varphi} = 0.09$).

5.2 Model criticism

The application of phylogenetics-based techniques to detecting selection in genes sampled from within populations of microparasites such as viruses (Twiddy *et al.* 2002; Moury *et al.* 2004; de Oliveira *et al.* 2004) and bacteria (Peek *et al.* 2001; Urwin *et al.* 2002) has become widespread. However, many of these organisms are highly recombinogenic (McVean *et al.* 2002; Awadalla 2003), so the use of phylogenetic methods is inappropriate because a bifurcating tree is not an adequate model of the evolutionary history of these organisms (Holmes *et al.* 1999). This model misspecification can elevate the rate of false positives when detecting sites under positive selection (Anisimova *et al.* 2003; Shriner *et al.* 2003), which casts doubt over the validity of the inferences made. Model criticism is an important part of formulating and fitting an evolutionary model, because it allows that model to be falsified if it is a poor description of the data. In this section I discuss how model criticism can be integrated into the Bayesian analysis itself, using *porB3* as an example. The aspects of model criticism that I will focus on are the sensitivity of inference to the priors, goodness-of-fit testing using posterior predictive distributions of summary statistics, and the effect of violating the assumption of random sampling in the coalescent model.

5.2.1 Prior sensitivity analysis

To determine the sensitivity of inference to the choice of priors, the analysis was repeated with alternative priors (Prior B in Table 2). The choice of distributional forms and parameters for prior B is largely arbitrary, and necessarily so, because prior A was supposed to represent earnest prior beliefs. For practical purposes, a variety of

distributional forms was used to test the computer program. For μ and ρ a uniform prior between 0 and 10 was fit (10 being the highest value considered plausible for either parameter). Following Huelsenbeck and Dyer (2004) a prior distribution on κ was fit that describes the ratio of two independent and identically distributed exponential random variables (see section 4.1.2.2). The moments, including the mean, for this distribution are undefined, but the median equals 1. For φ the mean of the exponential prior was changed from 0.1 to 1. Finally, for ω a gamma distribution was used, still with a mean of 1, but with shape parameter 2, giving the distribution a mode at 0.5. This distribution retains the case of selective neutrality for its mean, but it tails off towards zero rather than increasing. Three MCMC chains were run, each 250,000 iterations in length, with a burn-in of 20,000 iterations. Having checked for convergence, the chains were merged to obtain the posteriors.

95% HPD intervals for the peak ω in loops 1, 5, 6 and 7 show that the magnitude of the estimates has been reduced by the gamma prior to (2.76, 6.80), (2.16, 5.79), (2.31, 6.70) and (2.16, 5.66) respectively. Despite this, the relative height of the peaks is conserved. The average $\hat{\omega}$ for the sequence is 0.68, reflecting the more conservative effect of the gamma prior. Excluding sites for which $\hat{\omega} > 1$, this drops to 0.17, which is almost identical to the inference based on prior A. This suggests that information about the absolute magnitude of sites under functional constraint is less influenced by the prior. Despite differences concerning the magnitude of ω, the priors strongly agree on which sites are under diversifying selection (Figure 7).

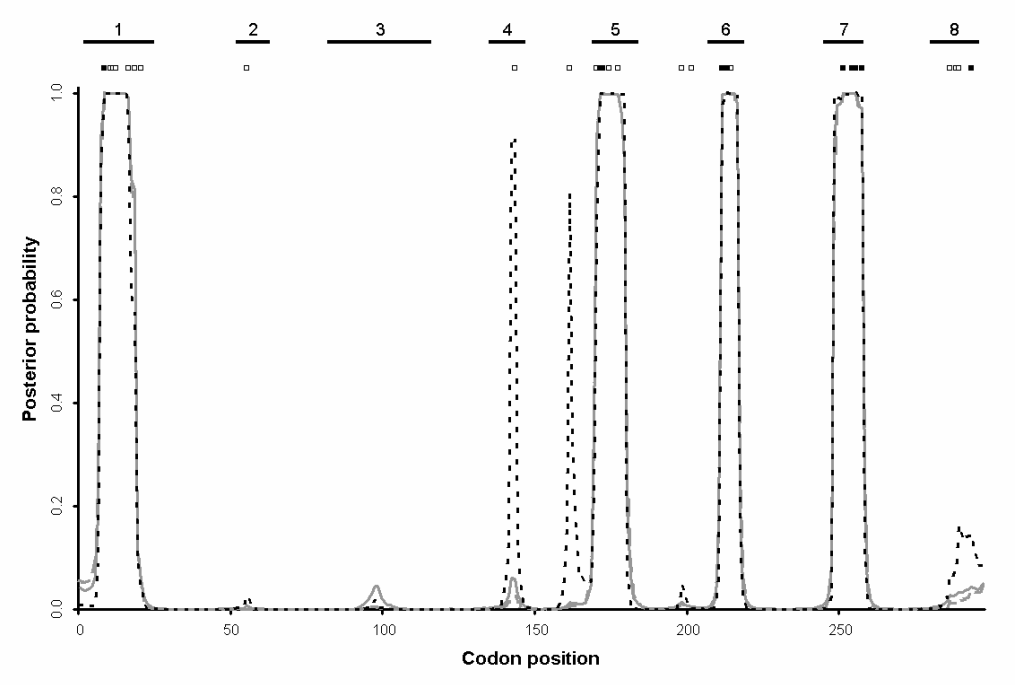


Figure 7 Site-wise posterior probability of diversifying selection ($\omega > 1$) for the *porB3* carriage study, under prior A (grey solid line), prior B (grey dashed line), and prior A with the recombination rate forced to equal zero (black dotted line). The loop regions are numbered above. Those sites identified as under weak (empty squares) or strong (filled squares) positive selection by Urwin *et al.* (2002) are shown.

The posterior probability of diversifying selection at a given site is

$$\Pr(\omega > 1) = \int_{1}^{\infty} \Pr(\omega \mid \mathbf{H}) \mathrm{d}\,\omega .$$

Prior A is represented by the solid grey line, and prior B by the dashed grey line. The two lines are virtually indistinguishable from one another at every site, indicating that identification of sites under diversifying selection is robust to the choice of prior.

Figures 6a and 6b compare the posterior probability of ρ given priors A and B. Under prior B, the posterior on ρ is somewhat flatter, with tighter credible intervals. The

average $\hat{\rho}$ is largely the same for most of the sequence, except at the far ends, where $\hat{\rho}$ increases sharply. This is an edge effect where, in the lack of information about the recombination rate, the posterior has been overwhelmed by the prior. The uniform prior on ρ has mean 5, explaining the rapid increase. The effect is reflected in the posterior on *R* (Table 3), which has a similar lower bound, but much increased upper bound. This striking sensitivity to the prior at the edges suggests that one should be cautious in interpreting the recombination rates at the edges of the sequence.

The posterior on μ is influenced by the high mean of the uniform prior (Table 3), to the extent that $\hat{\mu} = 0.35$ under prior B, which is only just inside the upper bound of the credible interval under prior A. In contrast, κ is not particularly sensitive to the prior, with largely overlapping credible intervals. φ shows a similar sensitivity to μ in responding to a considerable increase in the prior mean. The lower bound is almost unaffected, but the mean and upper bound show a marked increase.

5.2.2 Posterior predictive *p*-values

An essential property of any statistical model is that it should be falsifiable. A useful approach in Bayesian inference, and the one used here, is that of posterior predictive *p*-values (Rubin 1984). Here the model is taken to mean the probability model together with the posterior distribution of the model parameters. In essence, if the model is a good description of the data, then further datasets simulated under that model ought to resemble the real data. If they do not, then the model is failing in some important way. By *resemble* what is meant is that with respect to some statistic *D*, the

observed value of that statistic, $D_{\mathbf{H}}$ should fall well within the range of values for the simulated datasets, $D_{\mathbf{H}'}$, where $\mathbf{H}'$ is used to denote a simulated dataset.

The posterior predictive *p*-value is defined as the probability under the model of observing a discrepancy statistic *D* as large as that observed.

$$p = \int P(D_{\mathbf{H}'} \geq D_{\mathbf{H}} \mid \Theta) P(\Theta \mid \mathbf{H}) \mathrm{d}\Theta ,$$

where the integration is approximated by

$$p \approx \frac{1}{M} \sum_{i=1}^{M} I(D_{\mathbf{H}'_i} \geq D_{\mathbf{H}}). \quad (1)$$

In Equation 1, *M* is a large number (I used $M \approx 15{,}000$), $\mathbf{H}'_i$ is simulated from the posterior distribution $P(\Theta \mid \mathbf{H})$, and *I* is the indicator function. It is important to note that $\mathbf{H}'_i$ is simulated under the exact probability model specified by the PAC likelihood and used in inference, which is not the coalescent but an approximation to it.

5.2.3 Simulating under a PAC model

The algorithm here for simulating under the PAC model follows directly from the description of the model (see section 4.2). Here the total number of sequences is denoted *n*.

1. Generate the first haplotype by drawing the codon independently at each site from the equilibrium frequency of codons. Let the number of haplotypes, $k = 1$.

2. Generate the $(k + 1)$th haplotype conditional on the first k by parsing the sequence 5′ to 3′ in the following way. For the 5′-most site, choose one of the k haplotypes to copy from uniformly at random. Call this haplotype x.
3. Mutate the current site j by drawing a time t, independently of all other sites, from an exponential distribution with rate k. Then use the transition probability matrix $\mathbf{P}^{(t)}$ for the NY98 mutation model with parameters μ, κ and ω_j to draw a codon conditional on codon j in haplotype x and time $2t$.
4. Move to the next site, and continue to copy haplotype x with probability $\left(1-\exp\left\{-\rho_j d_j / k\right\}\right)$, where $\rho_j/2$ is the recombination rate between the sites (per bp per PN_e generations) and d_j is the distance (in bp). Otherwise choose one of the k haplotypes to copy from uniformly at random (including x). Call this haplotype x.
5. Return to step 3 until at the 3′-most codon.
6. Let $k = k + 1$. If $k < n$ return to step 2.

5.2.4 Combining *p*-values

A large number of datasets, M, are simulated as described in section 5.2.3. For each dataset the parameters are drawn from one of the iterations of the combined MCMC chain. Then for any particular discrepancy statistic, a marginal posterior predictive p-value can be calculated using Equation 1. The p-value is made two-tailed in the usual way. Combining p-values might be done using the Bonferroni correction, but this would be conservative even if the discrepancy statistics were independent. Since they are unlikely to be independent, Bonferroni would be too conservative for the goodness-of-fit test. That is to say that the p-value would under-estimate the extremity

of the observed data under the model, so the model would be less likely to be falsified. I am grateful to Jonathan Marchini for explaining the following way to combine *p*-values, which transforms the *p*-values into standard normal variates. By assuming that the transformed *p*-values can be made independent by removing the linear correlation structure (by further transformation of the multivariate normal distribution), a combined *p*-value can be obtained.

To combine two-tailed *p*-values for *N* different discrepancy statistics, denote the vector of discrepancy statistics for dataset *j*

$$\mathbf{D}_j = \left(D_{1j}, D_{2j}, ..., D_{Nj}\right).$$

Transform the marginal distribution of each discrepancy statistic i ($D_{i1}, D_{i2}, ...D_{iM}$) into a standard normal distribution, so that

$$Z_{ij} = \Phi^{-1}\left(\frac{W_{ij}+1}{M+1}\right),$$

where W_{ij} is the marginal rank (with respect to *j*) of discrepancy statistic D_{ij}, and Φ^{-1} is the quantile function (inverse cdf) for the standard normal distribution. Next assume that the joint distribution of $\mathbf{Z}_j = \left(Z_{1j}, Z_{2j}, ..., Z_{Nj}\right)$ is multivariate normal with zero mean and variance-covariance matrix $\mathbf{\Sigma}$, where

$$\Sigma_{kl} = \begin{cases} r_{kl} & \text{if } k \neq l \\ 1 & \text{if } k = l \end{cases},$$

where r_{kl} is the correlation coefficient between the transformed discrepancy statistics *k* and *l* (Z_{kj} and Z_{lj}) over datasets *j*. Next transform $\mathbf{Z}_j$ to remove the correlation structure

$$\mathbf{Y}_j = \mathbf{\Lambda}^{-1}\mathbf{Z}_j,$$

where $\mathbf{\Lambda}$ is obtained from the matrix factorization

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T .$$

$\mathbf{\Lambda}$ is computed by singular value decomposition (see, for example, Press *et al.* 2002).

Include the observed values of the discrepancy statistics $\mathbf{D_H}$ in the above procedure to obtain $\mathbf{Y_H}$. Assuming that the uncorrelated transformed discrepancy statistics are independent, then

$$X_j = \sum_{i=1}^{N} Y_{ij}^2$$

has a chi-squared distribution with *N* degrees of freedom. This can be verified by a histogram of the X_j's. A one-tailed chi-square test of $X_{\mathbf{H}}$ combines the two-tailed posterior predictive *p*-values.

5.2.5 Choice of statistics and results

Discrepancy statistics have to be chosen that describe some aspect of the data that should be fit well by the model. This is important because it is unlikely that a model will fit all aspects of the data well. Statistics that are sensitive to mutation are *S*, the number of segregating sites and $\bar{\pi}$, the average number of pairwise differences. For recombination, the variance in the number of pairwise differences $V(\pi)$ and the minimum number of recombination events R_m (Hudson and Kaplan 1985) are useful statistics (see section 2.1.3). The correlation between physical distance *d* and LD (r^2, D' and *G*4) that was used previously in the permutation test is also sensitive to recombination. For selection the statistic *U* is sensitive to any tendency for the

Table 4 Posterior predictive *p*-values

	Carriage study				Global study	
	Observed	Prior A	Prior B	Prior A $\rho = 0$	Observed	Prior A
S	67	0.236	0.039	0.008	92	0.391
$\bar{\pi}$	25.3	0.340	0.179	0.003	26.9	0.068
$V(\pi)$	94.0	0.268	0.391	0.000	98.2	0.118
R_m	15	0.293	0.658	0.070	12	0.036
$\mathrm{cor}(r^2, d)$	-0.13	0.247	0.265	0.002	-0.07	0.002
$\mathrm{cor}(D', d)$	-0.24	0.440	0.353	0.000	-0.10	0.059
$\mathrm{cor}(G4, d)$	0.22	0.443	0.332	0.000	0.09	0.144
U	0.5	0.543	0.878	0.711	0.5	0.621
D	1.05	0.121	0.058	0.567	0.97	0.398
Combined		0.268	0.103	0.001		0.013

simulated data to have too much or too little non-synonymous polymorphism on average.

$$U = \frac{\sum_{i=1}^{L} I\left(u_{\mathbf{H}'}^{(i)} \geq u_{\mathbf{H}}^{(i)}\right)}{\sum_{i=1}^{L} I\left(u_{\mathbf{H}'}^{(i)} \neq u_{\mathbf{H}}^{(i)}\right)},$$

where $u^{(i)}$ is the number of non-synonymous pairwise differences minus the number of synonymous pairwise differences at site *i*. *U* should be centred around 0.5. $U > 0.5$ indicates a bias towards diversifying selection and $U < 0.5$ a bias towards functional constraint. Finally Tajima's *D* (Tajima 1989) is used, which is sensitive to directional

selection, balancing selection and demography; not forces that were modelled explicitly.

As with a classical *p*-value, if *p* is very small then the model does not fit the data well. Table 4 shows the observed values of all the discrepancy statistics and the two-tailed posterior predictive *p*-values for the carriage study under priors A and B. Of all the discrepancy statistics, the only posterior predictive *p*-value in the first two columns less than 0.05 is *S* for prior B. To obtain a single posterior predictive *p*-value for each model, the marginal *p*-values from one each of the mutation-sensitive, recombination-sensitive and selection-sensitive statistics (*S*, $\mathrm{cor}\left(r^2, d\right)$ and *U*) were combined following section 5.2.4. Table 4 shows that the combined posterior predictive *p*-values for the carriage study under priors A and B are $p = 0.268$ and $p = 0.103$ respectively. Neither is in the 5% tail of the distribution, suggesting the model fit is adequate with respect to mutation, recombination, and selection insofar as the dN/dS ratio is concerned. Tajima's *D* was positive ($D = 1.05$), which may indicate balancing selection or population structure. The *p*-value for neither prior was in the 5% tail, so while these forces have not been modelled explicitly, the fit appears to be adequate. In fact for finite-sites mutation models, Tajima's *D* can have an expectation greater than zero under the standard neutral model

5.2.6 Analysis of the global study

As an informal test of how violating the coalescent assumption of random sampling would affect inference, the 79-sequence PorB3 data (the global study) of Urwin *et al.* (2002) were analysed using prior A. For computational tractability one randomly

chosen ordering of the haplotypes was used. Three MCMC chains were run, each 500,000 iterations in length, with a burn-in of 20,000 iterations. Having checked for convergence, the chains were merged to obtain the posteriors. Table 4 shows that $\hat{\mu} = 0.31$ was barely larger than for the carriage study, and the credible intervals overlapped almost entirely. The rate of insertion/deletion, φ was not greatly affected ($\hat{\varphi} = 0.08$), nor was the transition-transversion ratio ($\hat{\kappa} = 3.34$). But the total recombination rate doubled to $\hat{R} = 78.0$ with no overlap in the credible intervals. Across the sites, the recombination map (Figure 6c) does not differ greatly in the left half of the sequence (c.f. Figure 6a), but thereafter rises rapidly to about $\rho = 0.38$. The low posterior predictive *p*-values for the recombination-sensitive discrepancy statistics (Table 4) advises caution on the interpretation of $\hat{\rho}$.

However, inference on ω was hardly affected. Loops 1, 5, 6 and 7 still have very high posterior probabilities of diversifying selection. The magnitude of ω inferred for each loop is comparable, with the 95% HPD intervals for the four loops (2.89, 7.28), (3.47, 8.17), (3.22, 8.79) and (3.10, 7.60). The only substantive difference is in loop 8, which now also has high posterior probability of $\omega > 1$. The 95% HPD interval for the peak ω in loop 8 is (0.66, 2.87) and $\Pr(\omega > 1) = 0.92$. This difference can be explained by sites in loop 8 that exhibit amino acid variation in the global study but not the carriage study. The average $\hat{\omega}$ for the whole sequence is 0.91, and excluding sites for which $\hat{\omega} > 1$, it drops to 0.22, both comparable to the carriage study.

5.3 Evidence for false positives

Ancestral recombination can cause false positives in phylogenetic methods (Shriner *et al.* 2003, Anisimova *et al.* 2003). If this has had an important effect on the analysis of meningococcal PorB3 then one should expect to see those false positives when the results of the CODEML analysis (Urwin *et al.* 2002) are compared to those presented here. Those sites identified as under weak (empty squares) and strong (filled squares) diversifying selection by CODEML are illustrated in Figure 7. All of the strongly selected sites and all but five of the weakly selected sites fall within loops 1, and 5-8. With the exception of loop 8 all these sites had high posterior probability of diversifying selection for the carriage study (Figure 7). When the global study is analysed, loop 8 also has high posterior probability of diversifying selection. Therefore there are just five sites where CODEML inferred diversifying selection but omegaMap did not. These are candidates for false positives.

There are a number of possible explanations for discrepancies of this kind, including

1. The approximation in omegaMap has given rise to false negatives. The PAC likelihood does not explicitly model the genealogy and this might have unexpected effects.
2. The block-like prior in omegaMap caused false negatives. Imposing a model in which adjacent sites share a common selection parameter might disfavour isolated sites under diversifying selection.
3. Recombination has caused CODEML to give false positives.

In an attempt to distinguish between the explanations, an analysis of the carriage study was performed in which the recombination rate was forced to equal zero. Using

prior A, three chains were run for 500,000 iterations each. After a burn-in of 20,000 iterations the chains were compared for convergence and merged to give the posterior.

In Figure 7 the site-wise posterior probability of diversifying selection is plotted (black dotted line) for comparison with the other analyses. The false-positive candidates are located at sites 55, 143, 161, 198 and 201. Of these, the first two are located in loops 2 and 4 respectively. The remaining three are not in loops. Comparison of Figure 4 to Figure 2 shows that these latter three disputed sites are located in a cytoplasmic region of the protein. The site-wise posterior probability of diversifying selection is very similar to the other analyses (Figure 7), except at two positions. These two positions correspond to two of the five false-positive candidates: sites 143 and 161. Although one cannot be certain that these sites are false positives, the results are suggestive.

The posterior predictive *p*-values (Table 4) show that the deleterious effect of assuming no recombination is not confined to recombination-sensitive discrepancy statistics. The mutation-sensitive parameters also have extremely low *p*-values (0.008 and 0.003 for S and $\overline{\pi}$ respectively). The combined test shows that the model as a whole is a very poor description of the data ($p = 0.001$). Although the selection-sensitive parameters do not have significant *p*-values, the consequence of the model inadequacy is to cast doubt on all inferences made from it.

The PAC model in the absence of recombination does not default to the coalescent with no recombination because the tree is still not modelled explicitly. Therefore it is unlikely that the assumption of no recombination will affect a PAC model and a

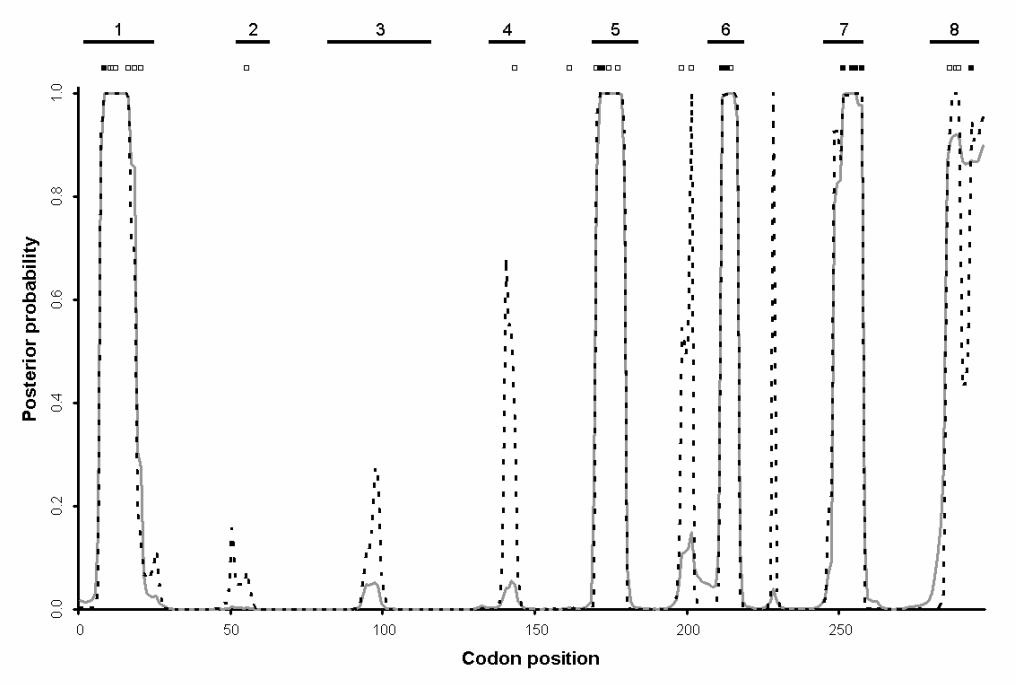


Figure 8 Site-wise posterior probability of diversifying selection ($\omega > 1$) for the *porB3* global study, under prior A (grey solid line), and prior A with the recombination rate forced to equal zero (black dotted line). The loop regions are numbered above. Those sites identified as under weak (empty squares) or strong (filled squares) positive selection by Urwin *et al.* (2002) are shown. Cf. Figure 7.

phylogenetic model in an exactly equivalent fashion. Nevertheless, when it is assumed that there is no recombination, sites that otherwise had low posterior probability of diversifying selection attained high posterior probabilities. This outcome is exactly what is predicted by the work of Shriner *et al.* (2003) and Anisimova *et al.* (2003).

When the global study is analysed under the constraint that the recombination rate equals zero, the same effect is seen: sites that have low posterior probability of diversifying selection when the recombination rate is unconstrained attain high posterior probabilities. But the actual sites affected differ when the global study is

analysed compared to the carriage study (compare dotted line in Figure 8 to Figure 7). Of the false positive candidates, site 55 remains at low posterior probability. Sites 143 and 161 have reduced posterior probabilities, 0.35 compared to 0.91, and 0.01 compared to 0.59, respectively. Sites 198 and 201 have increased posterior probabilities, 0.54 compared to 0.02, and 0.99 compared to 0.00, respectively. In addition, sites 140 and 228 reach high posterior probability. These sites exhibit amino acid polymorphism in the global study but not the carriage study. Loop 8 sites also show high posterior probabilities, compared to the carriage study with zero recombination. Like sites 140 and 228, these sites exhibit amino acid polymorphism in the global but not the carriage study, but are not candidates for false positives because they are identified as under positive selection in the global study with unconstrained recombination.

The effect on inference of assuming no recombination is complex, and as a result it is not possible to say with any confidence that particular sites are false positives. In general, however, by constraining the recombination rate to equal zero in the PAC models, the number of sites identified as under positive selection is inflated, which is predicted by the work of Anisimova *et al.* (2003) and Shriner *et al.* (2003). Only sites that exhibit amino acid polymorphism or immediately adjacent sites are identified as under positive selection. Why might imposing a phylogenetic tree on recombining genes lead to some sites that exhibit amino acid variation to have an inflated estimate of the dN/dS ratio? Recombination can cause homoplasies in a phylogenetic tree, i.e. a pair of sequences that appear to be distantly related overall share a rare allele at a particular site which the rest of the sample suggests was not shared by the common ancestor of those sequences. Therefore, recurrent mutation must be invoked to explain

the homoplasy. (See for example Figure 4 in Chapter 4; homoplasies can be explained by recombination or by extra mutation). As a result, some sites will appear to be hypermutable. If ω is allowed to vary but the synonymous rate of mutation is not, then sites with non-synonymous homoplasies will be best fit by an elevated ω at that site. These sites will appear to be positively selected. Yet sites with synonymous homoplasies, which will also appear hypermutable, cannot be fit by simply lowering ω because the synonymous mutation rate is constrained. Therefore there is no symmetric effect in which some sites have very low ω.

5.4 Analysis of housekeeping loci

Evolutionarily, antigen genes such as *porB* and housekeeping genes responsible for essential metabolic processes are under entirely different selection regimes. Whilst an antigen such as *porB* is exposed to immense selection pressure for antigenic novelty imposed by the host immune system, a housekeeping gene is, perhaps as a result, shielded from such co-evolutionary conflict and has the opportunity to adapt to an optimal functional form provided the necessary mutations arise. Therefore one would expect that housekeeping loci exhibit strong functional constraint. The loci used in MLST were chosen to be conserved, with no unusual signatures of selection or recombination, but sufficiently polymorphic to provide resolution for typing (Urwin and Maiden 2003). Contrasting the variation in the dN/dS ratio and recombination rate (selection and recombination 'maps', say) for the *porB* locus with those estimated from the MLST loci should make for interesting comparisons.

Table 5 Genetic diversity in Czech carriage study

Locus	Length (bp)	No. alleles	No. poly-morphic sites	dN/dS
abcZ	432	21	75	0.074
adk	465	19	25	0.011
aroE	489	21	135	0.295
fumC	465	29	48	0.010
gdh	501	19	26	0.049
pdhC	480	25	83	0.068
pgm	450	24	80	0.112

Table 5 summarises the genetic diversity in the seven MLST loci sequenced from a population of carried meningococci in the Czech Republic in 1993 (Jolley *et al.* 2000). There were 217 isolates in total. Polymorphism at the nucleotide level ranges from 5-20% across the loci. The dN/dS ratios, estimated using the number of non-synonymous and synonymous polymorphisms observed in the data, range from 0.010-0.295. *fumC* is the least polymorphic locus and has the lowest dN/dS ratio. *aroE* is the most polymorphic locus and has the highest dN/dS ratio. All dN/dS ratios are less than one, suggesting that differences may be due to the relative functional constraint rather than there being any evidence for positive selection. However, because these ratios are averaged across the sequence any positive selection occurring might be diluted by surrounding highly constrained regions.

For each locus, a random subset of fifty out of the 217 Czech sequences was chosen for analysis. This mainly reflects a computational constraint in that the complexity of

the PAC likelihood is quadratic in the number of sequences (Chapter 4). For the same reason, only one ordering of the PAC likelihood was used in these analyses. For direct comparison with the *porB3* results, the same priors on μ, κ, φ, ω, ρ (Prior A, Table 2) and the number of blocks were used. This is not entirely justifiable because the Czech carriage data were used originally to inform the choice of priors on *porB3*. For each locus three MCMC chains were run, each of length 500,000 iterations, with a burn-in of 20,000 iterations. Having established convergence, the three chains for each locus were merged to obtain the posteriors.

Figure 9 shows the estimated selection and recombination maps for the seven loci. From the selection maps it is clear that the sequence-averaged dN/dS ratios presented in Table 5 belie the true extent of variation in selection pressure in some of the housekeeping genes. Recombination rates are more conserved, with little compelling evidence for major peaks or troughs in any of the loci. There is some evidence for limited recombination rate variation in *adk* and *pdhC*. Interestingly, *fumC* is particularly functionally constrained, with almost no variation in ω along the sequence (average $\hat{\omega} = 0.016$) and *aroE* exhibits the greatest variation in and highest values of ω (average $\hat{\omega} = 0.409$). The *abcZ*, *gdh* and *pdhC* loci all show some spikes in the selection map.

Table 6 Point estimates for MLST loci

	abcZ	*adk*	*aroE*	*fumC*	*gdh*	*pdhC*	*pgm*
$\bar{\omega}$	0.124	0.076	0.409	0.016	0.101	0.136	0.184
μ	0.298	0.106	0.696	0.195	0.113	0.452	0.601
κ	10.2	7.63	3.38	9.51	6.88	5.86	3.77
φ	0.056	0.050	0.257	0.071	0.097	0.051	0.159
R	10.7	12.6	10.0	12.3	7.86	29.9	16.7

Table 6 shows the point estimates (mean posterior) for μ, κ, φ, R and the average ω ($\bar{\omega}$) along the sequence for each locus. The estimates of $\bar{\omega}$ agree well with the relative estimates of dN/dS in Table 5, but not the absolute values. Likewise, the estimate of μ (Table 6) agrees well with the relative polymorphism in the seven loci (Table 5). The recombination rates are of the same order of magnitude as those estimated in Chapter 2, but appear to be slightly elevated, possibly a manifestation of the upwards bias noted in Chapter 4. The relative magnitude of recombination rates across loci agree well.

There is little evidence for diversifying selection in any of the MLST loci in the Czech carriage study (Figure 9), although sections of *aroE* appear to be close to selective neutrality. Some loci appear to show an increase in ω at the 5′ and 3′ extremes of the sequence. This is most likely an edge effect, caused by reduced sequence information at the extremes so the prior has more influence over the inferred dN/dS ratio. In no locus does the mean of the posterior appear to exceed 1 at the extremes of the sequence.

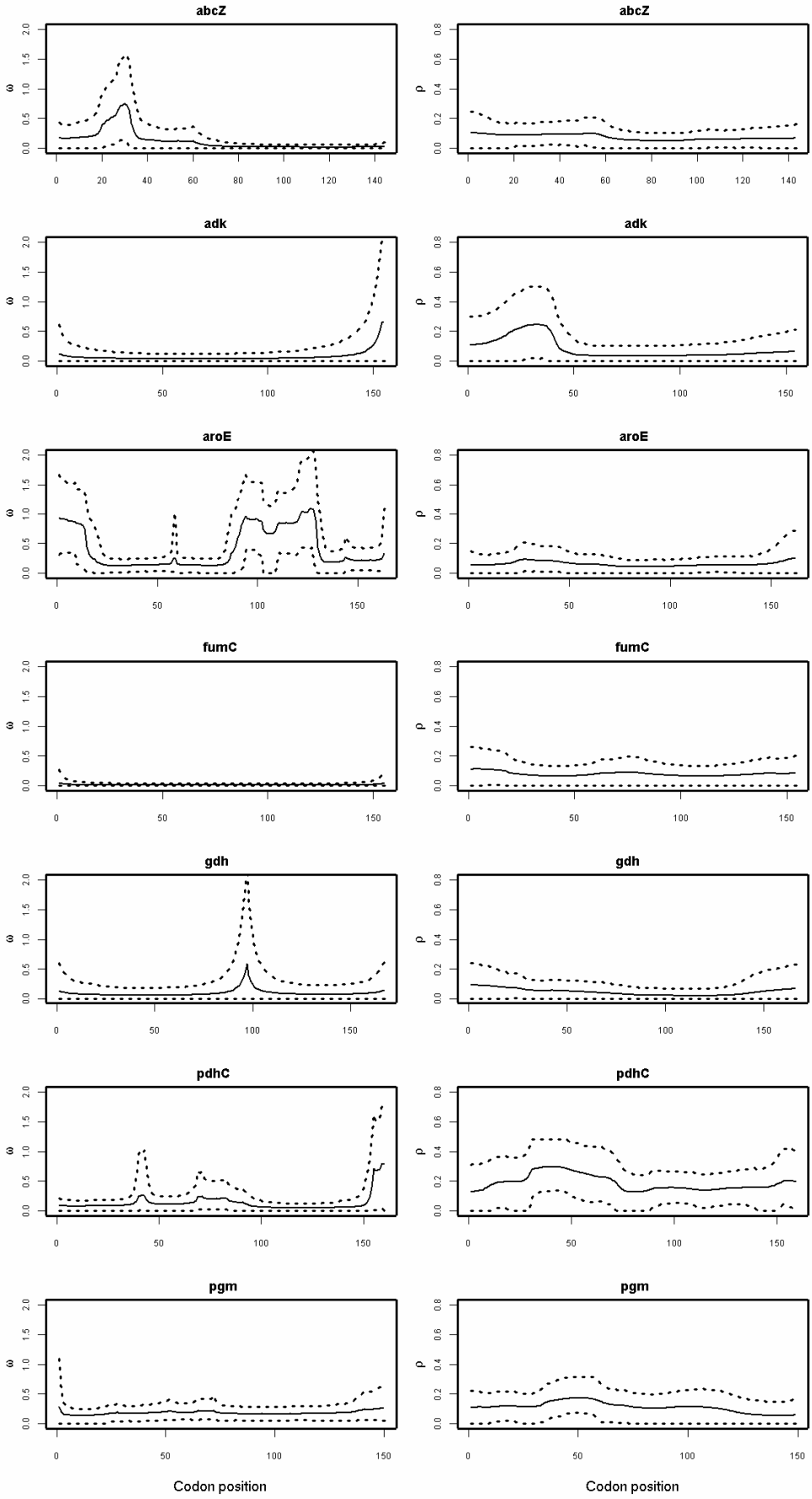


Figure 9 Variation in ω and ρ in MLST genes in the Czech carriage study.

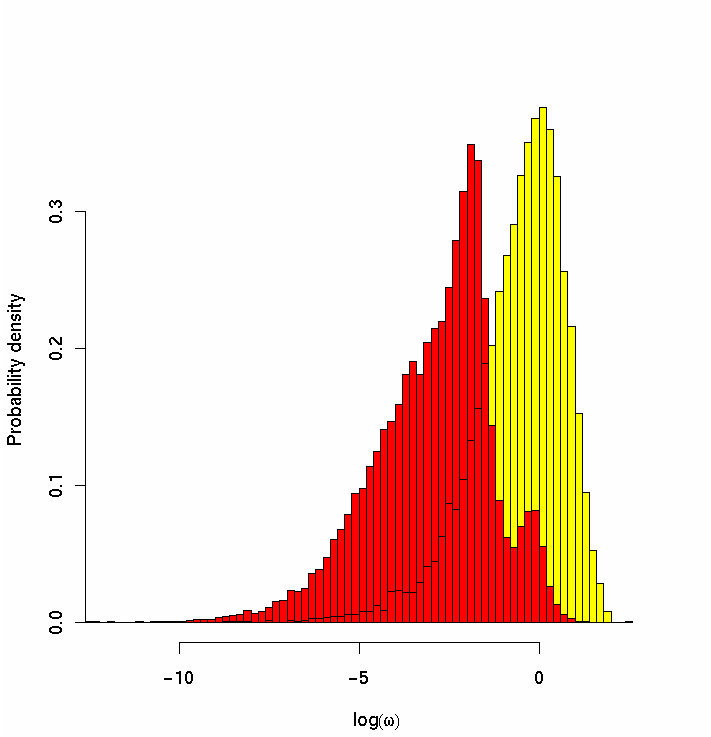


Figure 10 Red histogram: posterior distribution of $\log(\omega)$ amalgamated across MLST loci in the Czech carriage study. Yellow histogram: prior distribution of $\log(\omega)$ for comparison. The prior on ω is an exponential distribution with mean 1.

On the whole, the MLST loci are functionally constrained, which is not unexpected. Whilst the prior favoured selective neutrality, the posterior distributions clearly favour lower values of ω. This is illustrated on a logarithmic scale in Figure 10. The posterior distribution of $\log(\omega)$ has been amalgamated across sites and across loci (red histogram). For comparison, the exponential prior with mean 1 is plotted (yellow histogram). There are two patterns of note. The whole posterior distribution is left-shifted compared to the prior, indicating good evidence of purifying selection in these housekeeping genes. However, the amalgamated posterior is bimodal, with a small peak at $\log(\omega) = 0$, suggesting that a minority of sites are actually selectively neutral, presumably due to relaxed functional constraint. There is no evidence for diversifying

selection in the amalgamated posterior. Figure 10 might serve as a useful prior in future studies of variation in ω in housekeeping loci.

5.5 Summary

In Chapter 4 I presented a new method for estimating the selection parameter ω and the recombination rate ρ from a sample of gene sequences. Uncertainty in the evolutionary history was taken into account using a coalescent-based approximate (PAC) likelihood. Variation in ω and ρ was modelled as a block-like structure with a variable number of blocks. A Bayesian inference scheme was used to average over the number and position of the blocks using reversible-jump MCMC and obtain the posterior distribution of the other parameters. Using simulations, the new method showed good power to detect variation in ω and ρ, and did not appear to confound the two. The method has a low false positive rate for detecting sites under diversifying selection. In this chapter I applied the method to the *porB* locus of *Neisseria meningitidis* and performed prior sensitivity analysis and model criticism to verify the results. I constrained the recombination rate to equal zero to discover the effect on inferring positive selection, and compared the results to those of phylogenetic methods that assume no recombination. The comparison suggests that some sites identified as under positive selection by CODEML may have been false positives. Seven housekeeping loci were analysed from a carriage study in the Czech Republic, and the inferred levels of selection and recombination were compared to those observed in the antigen gene *porB*. There was no evidence for diversifying selection in any of the housekeeping genes.

Chapter 6

Further Developments

Two distinct roles for mathematical modelling of meningococcal biology have been pursued in this thesis. On the understanding that any model is a caricature of reality, the pertinent question is not whether the model is right or wrong, but which aspects of reality does it fail to adequately capture? Answering this question is the first of those roles, and the one that I have utilised for learning about the population structure of *Neisseria meningitidis*. Beginning with a standard neutral coalescent model, which, I argued in Chapter 1, is the appropriate starting point for modelling infectious microorganisms, I used cross-validation to demonstrate that carried populations of meningococci have greater than expected population structuring. As seen in Chapter 2, that structure manifests itself as a dearth of unique sequence types (STs), positive values of Tajima's *D*, and stronger than expected correlation between linkage disequilibrium (LD) and physical distance. In Chapter 3 I showed that the excess structuring cannot be explained simply by over-sampling of closely related meningococci, as proposed by the neutral microepidemic model. Description of meningococcal populations using AMOVA and the Mantel test revealed that within Bavaria, there is some genetic differentiation caused by isolation by distance. By contrast, in the Czech Republic there was no evidence for genetic differentiation, whereas between the European countries of the Czech Republic, Greece and Norway, there is significant differentiation, the nature of which is not simple isolation by distance but reflects more complex sub-continental transmission routes. Disease-

causing meningococci do not appear to be a random subset of carried meningococci; observed genetic diversity suggest that disease-causing meningococci may persist alongside asymptomatic meningococci. In the discussion below I concentrate on how the results of these descriptive analyses can be used to refine a coalescent-based model of meningococcal populations.

Second of the two roles is estimating parameters of evolutionary relevance. When the model is a good fit, which was established in Chapter 5 using posterior predictive *p*-values, the parameter estimates can be biologically interpreted. Fitting the model of immune selection introduced in Chapter 4 revealed variation in selection pressure but not recombination rate within meningococcal antigen genes. Amino acid residues under strong diversifying selection corresponded to extracellular loop regions in the tertiary structure of the outer membrane protein PorB. This is consistent with selection for antigenic novelty in regions of the protein exposed to the immune system. In meningococcal housekeeping genes the dN/dS ratio was estimated to be consistently below the neutral expectation of one, concordant with the view that their gene products are not under diversifying selection. Quantitative inference allows evolutionary parameters from different genes to be compared; in addition to inferring that housekeeping genes are under strong functional constraint, it was shown that there is much less variation in the dN/dS ratio compared to the *porB* locus. I based the model of immune selection on the NY98 codon model of molecular evolution (Nielsen and Yang 1998), together with the PAC likelihood model of Li and Stephens (2003) and a piecewise constant model of variation in the dN/dS ratio and recombination rate (Green 1995; McVean *et al.* 2004). Below I discuss the advantages

and limitations of the new model, including the implications for vaccine research, and possible future extensions to it.

6.1 Meningococcal population structure

To begin with I will discuss the advantages of explicit evolutionary models over descriptive analyses such as AMOVA and the Mantel test. Descriptive analyses can help to direct the refinement of coalescent-based evolutionary models, and I will go on to describe how the results of the descriptive analyses performed in Chapter 3 might be used to inform development of a coalescent model with population structure. Then I will pursue the way in which inference might be performed using approximate Bayesian computation with conditional density estimation (ABC-CDE).

6.1.1 Advantages of explicit evolutionary models

Methods of analysis such as AMOVA, Mantel tests and logistic regression are useful ways to explore patterns of genetic diversity in meningococcal populations, and allow various informal scenarios to be explored. Such exploratory analyses can be used to help inform a more coherent approach to understanding the evolution of meningococci. None of the methods listed above make use of explicit evolutionary models. The advantage of an explicit evolutionary model is that parameters of interest can be estimated and the deficiencies of the model can be explored in a way that is readily interpretable. For example, in the analyses presented in Chapter 3 it was claimed that there is no evidence for geographic differentiation within the Czech Republic but there is in Bavaria, on the basis of whether the permutation test yielded a significant *p*-value or not. If an explicit evolutionary model with migration were fitted

to the data, estimates of transmission rates between geographic localities could be directly estimated and the results compared in a quantitative fashion. Such an approach might be more informative than the polarised result obtained from the permutation test that there either is or is not structure.

Population structure can be incorporated into the coalescent in several ways. When migration rates are high, so that the per generation probability of migration *m* is larger than $O(1/N)$ where *N* is the population size within a geographic deme, but the number of geographic demes *D* is finite,

$$M = \lim_{N\to\infty} PNm = \infty .$$

In this scenario, known as the strong migration limit, the genealogy of the population is that of a standard coalescent model with an altered timescale (see for example Nordborg 2003). Unlike the separation of timescales result obtained for the large *D* approximation (Wakeley 1998, 2001; Wakeley and Aliacar 20001), there is no scattering phase adjustment for the configuration of the sample amongst demes. In Chapter 2 a model of meningococcal evolution with a standard coalescent genealogy was rejected. Thus, the strong migration model of population structure can be rejected for meningococcal evolution, as can any other model that results in a standard coalescent.

When the migration rate is $O(1/N)$, so that

$$M = \lim_{N\to\infty} PNm$$

is finite, the resulting genealogy is known as the structured coalescent (Wilkinson-Herbots 1998). In the structured coalescent the number of geographic demes *D* is

finite, and the relative population size of all demes as well as all pairwise migration rates between demes, need to be specified. Therefore the structured coalescent has a large number of parameters, although various simplifications can be made, such as symmetric migration (Wright 1931, Maruyama 1970), one or two-dimensional stepping stone models (Kimura and Weiss 1964), or the large *D* approximation (Wakeley 2001). Another way of overcoming the problem of many parameters in the context of meningococcal evolution is discussed in the next section.

Using the structured coalescent would allow relevant evolutionary and epidemiological parameters to be estimated. For example, if population structure within Bavaria were to be modelled by treating each sampling location as a separate deme, then in principal the rates of migration between all pairs of towns could be estimated. These rates of migration correspond to transmission rates between demes, which is of epidemiological relevance for understanding how meningococci spread through human populations. If disease-causing and carried meningococci were to be treated as separate demes, then the migration rate could be interpreted as the rate of genetic exchange between the two groups, or the degree of genetic isolation. This again is of epidemiological relevance, for example estimating the rate of acquisition of antibiotic resistance in bacteria, or the speed at which capsule switching may occur in response to a specific vaccine (see section 1.1.3.2).

Using explicit evolutionary models allows for more meaningful hypothesis testing. For example, in Chapter 1 I discussed the problems with using classical methods such as the χ^2 goodness-of-fit test for rejecting a model of linkage equilibrium. The χ^2 test is inappropriate essentially because the wrong null distribution is obtained; genetic

drift can cause deviations from linkage equilibrium even for unlinked loci. Therefore the null hypothesis of linkage equilibrium might be rejected when it is true; the test is anti-conservation. By contrast, the permutation test used by AMOVA to detect population differentiation in the Czech Republic is conservative. This might explain why genetic differentiation was detected between sampling locations in Bavaria but not the Czech Republic even though the two regions are roughly the same size. In the context of using an explicit evolutionary model, more powerful analyses might be obtained by constructing a likelihood ratio test in a classical setting, or using Bayes factors or posterior predictive *p*-values in a Bayesian setting, (although it might be necessary to utilize summaries of the data for computational reasons). In this example the evolutionary model would be the structured coalescent.

6.1.2 Bayesian inference in the structured coalescent

The parameters of the structured coalescent could be estimated in the ABC-CDE (approximate Bayesian computation with conditional density estimation) setting described in Chapter 2. In the most general formulation of the structured coalescent, there is a migration rate for each pair of populations, or demes. Currently there are no 'full-likelihood' methods that estimate migration rates in the coalescent in the presence of recombination using the full sequence data. The problem obviously gets more difficult with increasing number of demes, but one approach that could be harnessed in the ABC-CDE framework would be to restrict the number of free migration parameters to one average rate.

The approach that I propose is to pre-determine the relative pairwise migration rates according to the population sizes of those demes (which could be measured from

census data for human towns), the geographic distance of those demes (which is straightforward to measure for towns), or some other geographic measure of connectivity, which might be as simple as the shortest road distance between towns. Several summary statistics exist that would be sensitive to the migration rate, including F_{ST} and the correlation between genetic and geographic distance. The various ways of defining the relative pairwise migration rates could quite easily be compared in the Bayesian framework using Bayes factors, cross-validation (discussed in section 2.3.4) or posterior predictive *p*-values (discussed in section 5.5.2). In my opinion, coalescent-based models, which can be fitted using methods such as ABC-CDE, offer a common thread which, in an iterative framework of model criticism and refinement, will be the best way to improve understanding of the evolution and epidemiology of meningococci.

6.2 Detecting selection in *Neisseria meningitidis*

In Chapter 4 I introduced a new model for detecting selection in genes of interest in *N. meningitidis*. The model was based on the NY98 codon model of molecular evolution (Nielsen and Yang 1998), together with the PAC likelihood model of Li and Stephens (2003) and a piecewise constant model of variation in the dN/dS ratio and recombination rate (Green 1995; McVean *et al.* 2004). In this section I will focus on the differences in inference based on the new model and existing inference methods, including the advantages of the Bayesian approach. I will discuss the important assumptions of the model, its limitations for inferring natural selection and future extensions to the model. The implications of such methods for vaccine research are

discussed briefly, and finally I will discuss a separation of timescales approach for allowing intra-host genetic diversity without invoking coinfection as an explanation.

6.2.1 Comparison of PorB3 analyses

On the supposition that the analysis using the new PAC method is the most faithful account of the evolutionary history of selection and recombination in the *porB3* locus, then analyses based on comparison of the observed number of pairwise synonymous and non-synonymous differences (Smith *et al.* 1995) seriously underestimate the true extent of diversifying selection in the antigenic locus. Smith *et al.* (1995) estimated that the average dN/dS ratio was 0.62 for *porB3*, whereas in section 5.1.6 it was estimated to be 0.90 on average (under Prior A for the carriage study). Smith *et al.* (1995) estimated that the number of non-synonymous relative to synonymous substitutions in loop regions was 2.3, and 0.28 in non-loop regions. In the analysis presented in section 5.1.6 the average dN/dS ratio was estimated to be around 6 in loops I, V, VI and VII, and 0.16 elsewhere. So not only was the average dN/dS ratio underestimated, but the extent of the differences between positively and negatively selected sites was also greatly under-estimated. This is partly because some loop regions do not experience positive selection. As a result the analysis undervalues the importance of PorB as a potential vaccine target.

Using the phylogenetic CODEML method, Urwin *et al.* (2002) estimated an average dN/dS ratio of 0.26 for *porB3*, which is considerably smaller than the 0.90 estimated using the new method. Most of the difference probably lies in the way that the new method allows adjacent sites to share a common selection parameter, and that the new method also models insertions/deletions; sites segregating for an indel are deemed to

exhibit non-synonymous polymorphism, whereas in CODEML these sites are excluded from the analysis. For sites identified as experiencing strong positive selection, CODEML estimated a dN/dS ratio of 13.9, which is considerably higher than estimates of around 6 using the new method. This might in part reflect the constraining effect of the exponential prior on ω, which disfavours values of ω far from 1. However, it may also reflect the model misspecification that CODEML suffers from when the genes have undergone recombination. The tendency for phylogenetic methods to infer hypermutability at sites with homoplasies caused by recombination, discussed in section 5.3, may artificially elevate the inferred value of dN/dS.

CODEML and the new method differ substantially in the inferred patterns of variation in ω spatially along the gene. The distribution of positively selected sites inferred by CODEML is distinctly sporadic (see Chapter 5 Figure 7), whereas the new method generally infers much smoother variation in the mode of selection. This is a direct result of the prior on variation in ω, which models blocks of contiguous codons which share a common selection parameter, and the fact that the new method is able to infer selection at sites segregating for indels, which CODEML does not. The prior model of variation in ω is used deliberately to create a smoother posterior and share information between adjacent sites. The smoothness can be controlled by the prior value of the parameter p_ω. However, the length of a block follows a truncated geometric distributed, so very short blocks consisting of only a single codon have the highest probability mass under the prior. Therefore I would argue that if the data did not support smooth variation in the selection parameter along the sequence then the posterior would not be smooth. CODEML also infers positive selection at single sites

in loop II, loop IV, between loops IV and V, and two sites between loops V and VI, which the new method does not agree with. For the same reason that I do not believe the smoothness of the variation in ω within loops I, V, VI and VII is an artefact of the prior, I do not believe that the new method has failed to pick these sites out because of over-smoothing. Instead some at least may be false positives, caused by the assumption of no recombination in CODEML (see section 5.3).

6.2.2 Aspects of the Bayesian approach

Besides the ability to co-estimate ω and ρ, there are several advantages to the new method. Some of these are a consequence of the Bayesian approach, and all of them rely on the computational tractability of the PAC model. First among these is that the posterior probabilities of diversifying selection are fully Bayesian, so they incorporate uncertainty about the evolutionary history, as well as uncertainty in the other parameters. In the presence of recombination, there is likely to be a great deal of uncertainty in the evolutionary history. The computationally efficient PAC likelihood means that in the posterior, ω can take on any positive value, rather than having to constrain it to a discrete number of points or approximate a continuous distribution in a similar manner.

The main objection to a Bayesian approach is the requirement to specify a prior distribution on all parameters. In a scientific context it may seem absurd to prejudice the outcome of statistical inference with the researcher's prior subjective beliefs. In practice it is possible to represent a lack of prior knowledge with relatively flat priors, such as the proper and improper uniform priors used in ABC-CDE in Chapters 2 and 3, although it should be noted that in reversible-jump MCMC it is not possible to use

improper priors (Green 1995). In Chapter 5 I took a different approach, that of prior sensitivity analysis. Prior sensitivity analysis reveals which aspects of the posterior distribution, if any, are unduly influenced by the choice of prior. This in turn reveals which aspects of the model the data are uninformative about. For example, Chapter 5 Figure 6b shows that the data contained very little information about recombination rates at the extremes of the sequence. In contrast, inference about diversifying selection (Chapter 5 Figure 7) was robust to the prior.

In a Bayesian setting it is entirely natural to impose a block-like structure on the joint distribution of ω across sites. At sites where the data is compatible with a block-like structure this allows information about ω to be combined across sites, but when the signal in the data is strong enough it will overwhelm the block-like model. The sensitivity to the signal is controlled by p_ω. This is a biologically realistic model insofar as adjacent sites in the primary sequence will be closely juxtaposed in the tertiary structure, and, as such, are more likely to perform similar functional duties. If anything, the model is overly simplistic because the tertiary structure could in principle be used to impose longer-range dependencies on the prior. In a maximum likelihood setting, implementing the block-structure described here would be computationally unfeasible.

On the basis of previous work (Schierup and Hein 2000; Shriner *et al.* 2003; Anisimova *et al.* 2003) and because of clear model misspecification I have claimed that it is inappropriate to analyse data that shows evidence for recombination using phylogenetic methods. Yet neither the coalescent, nor the approximation to the coalescent used in Chapters 4 and 5, inevitably fit data from a recombining

population. That is why the importance of goodness-of-fit testing has been emphasised. Posterior predictive *p*-values allow for goodness-of-fit testing in a Bayesian setting when there is no explicit alternative model specified. The posterior predictive *p*-values in Chapter 5 Table 4 showed that the model with no recombination is a very poor fit to the data, and Chapter 5 Figure 7 showed that in the PAC model the assumption of no recombination leads to an increase in the number of sites experiencing diversifying selection, which would be expected if this assumption increases the false positive rate.

Posterior predictive *p*-values (Chapter 5 Table 4) suggested that the coalescent approximation was not a good fit to the *N. meningitidis* global study. This was not unexpected because the global study did not represent a random sample from any population in a meaningful sense. In constructing the carriage study care was taken not to include more than one haplotype from any one host. The idea was to envisage the bacterial population as a metapopulation, as described in Chapter 1. Consistent with this model, the posterior predictive *p*-values showed that the coalescent approximation did provide an adequate fit to the carriage study. There is more work to be done on formalizing the relationship between genetic models, such as the coalescent, and epidemiological models, but it may be possible in future to use models such as the one presented here to estimate parameters of epidemiological relevance. This is discussed further in section 6.2.6 below.

6.2.3 Limitations of the method

Fundamentally, the likelihood model used for inference is an approximation to the coalescent, and in that sense it is a compromise. In Chapter 1 I discussed why the

coalescent model is an appropriate null model for the evolution of microparasites such as *N. meningitidis*. The coalescent is a model of the genealogy of a random sample of genes in a selectively neutral population. However, integrating over the many possible, unknown, genealogies is computationally unfeasible for all but the simplest problems. The PAC model (Li and Stephens 2003) attempts to perform this integration implicitly, in a computationally convenient, but approximate, fashion. There are several trade-offs in this approximation. Broadly speaking, the biggest disadvantage is that there is no formal relationship between the coalescent and the PAC model, therefore it is very difficult to predict how the PAC model will behave differently to the coalescent. More specifically, one major problem is that the ordering in which the conditional likelihoods are calculated influences the likelihood, so that haplotypes are no longer exchangeable. For any reasonable number of sequences n, the number of possible orderings $n!$ is too large to fully explore, so the likelihood must be calculated using a finite number of orderings. Following Li and Stephens (2003) the new method averages over a fixed number of random orderings to calculate the likelihood. In contrast, Stephens and Scheet (2005) treat the ordering as a model selection problem, and integrate over the orderings numerically using MCMC. In hindsight, this might be a more elegant solution to the problem.

The PAC model is an approximation to a sampling formula (in the sense of Ewens [1972]) for a finite-sites mutation model in the presence of recombination. The $(k + 1)$th haplotype is a copy of the first k, but recombination means that it may be a mosaic, and mutation means it may be an imperfect copy. Recombination causes mosaicism because adjacent sites are more likely to share the same evolutionary history than a random pair of sites. In section 4.2 I argued that the haplotype from

which the (k + 1)th copies can be thought of the nearest neighbour in the genealogy at that site. In calculating the emission probabilities for the HMM, the time to the mrca of the (k + 1)th haplotype and its nearest neighbour is integrated out marginally for each site. The probability distribution is exponential with rate k to the order of the approximation. However, when adjacent sites share the same nearest neighbour, they are more than likely to share the same time to the mrca with that neighbour. Therefore integrating out the time marginally for each site is inconsistent with the coalescent model. In fact adjacent sites could share the same time to the mrca if time is discretized to retain the structure of the HMM (Stephens and Scheet 2005), but this increases the time to calculate the PAC likelihood. Discretizing time using Gaussian quadrature was the basis of the original importance sampler of Fearnhead and Donnelly (2001) that motivated the model of Li and Stephens (2003). Further investigation is needed to discern whether the additional complexity of discretizing time in the PAC likelihood would improve inference in the context of inferring selection and recombination rates.

Even if it were possible to use the actual coalescent model for the genealogical history of the gene sequences, the utility of the method presented here would remain limited by the biological realism of the mutation model. The NY98 mutation model is a useful way to treat selection that is confounded with mutation in samples of gene sequences. However, in Chapter 4 I argued that the ability of the model to detect positive directional selection in which one functionally constrained form is replaced by another is seriously questionable. The NY98 model may be best put to use in the context of inferring positive diversifying selection in genes that interact with the host immune system, because positive selection in the NY98 model is really selection for

genetic novelty. Detecting single adaptive events is best left to other methods that exploit other signals of selection in the data, such as strong haplotype structure and lowered diversity surrounding a site that has undergone a selective sweep (e.g. Przeworski *et al.* 2003; Coop and Griffiths 2004). PAC models may still be a useful approximation to the coalescent in this context.

6.2.4 Extensions to the method

There are several obvious extensions to the method, some of which would be relatively easy to implement. Among these is the replacement of the mutation model with any other reversible nucleotide, codon or amino acid mutation model. One interesting avenue that might be especially useful in the study of microparasites is to adapt the PAC likelihood to model genes sampled at different points in time. Serially-sampled genetic data is potentially a very powerful resource, because (i) genes sampled further back in time are informative about the genealogy at that time, (ii) serially-sampled data allows the effective population size to be deconfounded from the estimates of the mutation and recombination rates (Drummond *et al.* 2003a; Drummond *et al.* 2003b) and (iii) changes in selection pressures over time might be modelled. In this sense, microparasites are measurably evolving populations, because the mutation and recombination rates are sufficiently high and generation length sufficiently short for their evolution to be observed in real-time (Drummond *et al.* 2003b). This might be interesting not only from an evolutionary perspective, but could be instructive for control and prevention of emerging infections or epidemics.

In contrast to the problem of incorporating serially-sampled data into a PAC likelihood, a straightforward extension to the method would be to allow more than

one isolate per host to be included in the sample. Provided that the origin of each of the isolates is known, Wakeley and Aliacar (2001) provide a genealogical model for repeated sampling of some hosts that could be easily incorporated into the model presented here. In essence, the PAC model used for inference in this chapter models the collecting phase of Wakeley and Aliacar's metapopulation genealogy. This extension would allow their scattering phase to be modelled as well. Incorporating the scattering phase into the model would involve a variable number of haplotypes at the beginning of the collecting phase, which could be integrated out as part of the MCMC scheme. In Wakeley and Aliacar's model (see section 1.4), sequences sampled from the same host either coalesce with others sampled from the same host or migrate (backwards-in-time, as a result of a transmission event) to a new host, before commencing the collecting phase. From the perspective of inference, only sequences that are identical can coalesce during the scattering phase if the mutation rate is finite on the timescale of the collecting phase, which implies that sequence variation within a host can only be explained by multiple infection under the model. In section 6.2.6 I discuss why this might not need to be the case.

6.2.5 Implications for vaccine research

Identifying sites in an antigen locus that are under positive selection can help to locate the determinants of antigenicity, because interaction with the host immune system causes selection for antigenic novelty, which is brought about by variation in the amino acid sequence. It has been claimed that identifying the determinants of antigenicity might inform vaccine research (de Oliveira *et al.* 2004). Currently meningococcal vaccines for non-serogroup-B meningococcal disease use the capsular polysaccharide or a conjugate polysaccharide-protein complex (Stuart 2001). Genetic

analysis such as that presented here cannot inform such studies because there is not a one-to-one correspondence between the nucleotide sequence and the antigen (the protein in the conjugate vaccine is only a carrier). However, much current research in vaccine development is focussing on serogroup B (Snape and Pollard 2005), for which no efficacious vaccine currently exists, owing to the poor immunogenicity of the serogroup B polysaccharide (see section 1.1.3). An alternative to conjugate polysaccharide vaccines are outer membrane vesicle (OMV) vaccines which can be readily obtained from the blebs constantly secreted by the outer membrane. The outer membrane proteins (OMPs) that are the immunodominant components of the OMV vaccines can also be synthesised in the laboratory, but it is difficult to simulate the conditions required for the natural conformation of the proteins (Frasch 1995).

In the context of OMV vaccines, genetic analyses are potentially of use. On the premise that positive selection in known antigens reflects interaction with the immune system, then the strength of interaction can be quantified using the selection parameter, either at a particular site or summed across sites. This could direct research into the protein components of the OMV and find the genetic determinants of immunodominance. When combined with knowledge of the tertiary structure of the proteins in their natural conformation, this might prove to be a useful tool. Of particular interest might be, not sites that are demonstrably under strong diversifying selection, which in this study were found to be those that were surface-exposed, but rather those that maintain strong functional constraint despite a prominent surface-exposed position. Such a result might suggest that the conservation of the residue or oligopeptide is so essential to the correct functioning of the protein that it is constrained despite strong selection for antigenic variation. Generally speaking, the

technology of OMV vaccines is currently too crude for fine mapping of immune selection on a gene sequence to be of great use. However, genetic analyses such as that presented here are inexpensive in comparison to laboratory experiments or field trials. This alone is a persuasive argument for pursuing population genetic analysis in the context of vaccine research even if the rewards are slight.

6.2.6 Separation of timescales in microparasite evolution

In Chapter 1 a metapopulation was used to model a population of hosts infected with a microparasite, where the epidemiological dynamics are described by a simple SIRS-style differential equation model. Using the coalescent model of a metapopulation (Wakeley and Aliacar 2001) provides a starting point for modelling the genealogy of gene sequences sampled from a microparasite population. When all isolates are sampled from different hosts, the genealogy is simply the coalescent, where the effective population size is a function of the epidemiological parameters. When some isolates are sampled from the same host, the genealogical history has two phases, the scattering phase and the collecting phase (Wakeley and Aliacar 2001). In the scattering phase the lineages ancestral to the sample coalesce within each host, or migrate backwards-in-time to other hosts. This occurs rapidly relative to the subsequent collecting phase, which is a coalescent process with altered time scale. In Wakeley and Aliacar's model, if the mutation rate in the collecting phase is finite, then no mutation events occur in the scattering phase because it occurs so rapidly relative to the collecting phase. The same is true of recombination. This implies that genetic variation in the parasite population within a single host must be caused by coinfection.

In reality there may be appreciable genetic variation in a host even if the infection had a single founder. This poses a problem for the genealogical model because mutation would then occur infinitely quickly during the collecting phase. It is trivial to show that this is untrue simply by examining a sample in which each host is represented by only a single isolate. One explanation is that the genealogical model is wrong. The key assumption is that there are a large number of demes, which allows the separation of timescales into a scattering and collecting phase. This assumption, however, seems reasonable. The separation of timescales is a very convenient result, and it might be premature to abandon it at this stage. In addition to the assumption that the within-host population size is large $(N_P \to \infty)$ and the number of hosts is large $(D \to \infty)$, Wakeley and Aliacar assume that $D >> N_P$. For viral microparasites such as HIV, clearly this assumption may not hold. However, it may be possible to observe intra-host variation without invoking multiple infection even when $D >> N_P$ is a reasonable assumption, using the same idea as the NY98 mutation model that mutation and selection are confounded.

During transmission, there is a bottleneck in genetic diversity. The bottleneck in diversity is caused by selection for genotypes that are adapted to transmission. Put another way, upon colonisation of a host, there is a relaxing of the selection pressure allowing the population to diversify. In any parasite population, only a fraction of genotypes will be competent for transmission. In a model such as NY98, selection is confounded with the mutation process. This is useful in the context of resolving the conflict between intra-host and inter-host genetic diversity. To illustrate the point, assume that the mutation rate μ within a gene is finite on the timescale of the scattering phase (explaining intra-host diversity) and that within the host there is no

selection acting on the gene; it is neutral. If that gene is important for transmission, so that only a fraction f of all forms are competent for transmission, then the effective mutation rate in an inter-host sample will be $f\mu$. If f is sufficiently small so that $O(1/f) > O(\mu)$, then the effective mutation rate might be finite on the timescale of the collecting phase, so that the gene would exhibit intra-host diversity yet also inter-host genetic structuring.

The idea of an effective mutation rate, where selection is modelled as a form of mutational bias, is the essence of the NY98 model. This suggests that intra-host and inter-host selection pressures should be separately parameterised, which is biologically reasonable because the adaptations required for surviving in an infected host may be quite different to those required for successful transmission to a new host, and there may be limited overlap between the two. Inference of intra-host and inter-host selection pressures could be performed in the context of the model presented here, when extended to incorporate the scattering phase of the genealogical process. In the terminology of Wakeley and Aliacar (2001), it would require integration (by MCMC) over the sample configuration at the end of the scattering phase. In principle, intra-host and inter-host selection maps could be co-estimated using the method presented here as the foundation.

6.3 Summary

Patterns of genetic diversity in parasite populations contain an, albeit corrupted, account of the evolutionary history of the population. Understanding that evolutionary history can help inform control and prevention strategies for pathologically

importance parasites such as *N. meningitidis*. The large genetic diversity, short generation times and intimate co-evolutionary interaction between host and parasite also make pathogens interesting case studies in the study of evolution. The right way to model genetic data is to take account of the strong inter-dependency of gene sequences imposed by the evolutionary tree. For gene sequences sampled at random from a population, the coalescent provides the appropriate null model (or prior distribution) for the underlying evolutionary tree, or trees in the presence of recombination. Evolutionary models for genetic data are the only way to obtain biologically relevant and interpretable parameter estimates. Improving understanding of the biology of pathogens by iteratively refining and criticizing the model of the evolutionary ancestry, arguably the most important role for mathematical models in evolution, can only be achieved by using biologically interpretable explicit models of evolution. Even so, descriptive methods such as AMOVA and Mantel tests have a role for exploring genetic datasets and generating hypotheses. Because of the inherent computational difficulties in performing inference on evolutionary models, further research is required into approximate techniques such as the PAC model and techniques based on summaries of the full data such as ABC-CDE. Used in combination, there is the potential to learn a great deal about the evolution of the microorganisms responsible for important infectious diseases of humans.

Glossary of Acronyms

ABC	approximate Bayesian computation
AMOVA	analysis of molecular variance
ANOVA	analysis of variance
ARG	ancestral recombination graph
bp	base pair
CDE	conditional density estimation
DLV	double locus variant
DNA	deoxyribonucleic acid
ET	electrophoretic type
HMM	hidden Markov model
i.i.d.	independently and identically distributed
LD	linkage disequilibrium
LPS	lipopolysaccharide
Mb	megabase
MCMC	Markov Chain Monte Carlo
ML	maximum likelihood
MLE	maximum likelihood estimate
MLEE	multilocus enzyme electrophoresis
MLST	multilocus sequence typing
mrca	most recent common ancestor
MVP	meningitis vaccine project
OMP	outer membrane protein
OMV	outer membrane protein vesicle
SLV	single locus variant
ST	sequence type

Literature Cited

ACHTMAN, M., 1995 Global epidemiology of meningococcal disease in *Meningococcal Disease*, edited by K. CARTWRIGHT. John Wiley & Sons Ltd, Chichester.

ANDERSON, E., Z. BAI, C. BISCHOF, S. BLACKFORD, J. DEMMEL *et al.*, 1999 *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

ANDERSON, R. M., and R. M. MAY, 1991 *Infectious diseases of humans: dynamics and control*. Oxford University Press, Oxford.

ANDERSON, T. J., 2004 Mapping drug resistance genes in *Plasmodium falciparum* by genome-wide association. *Curr. Drug Targets. Infect. Disord.* **4:** 65-78.

ANDERSON, T. J., B. HAUBOLD, J. T. WILLIAMS, J. G. ESTRADA-FRANCO, L. RICHARDSON *et al.*, 2000 Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol.* **17:** 1467-1482.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* **18:** 1585-1592.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2002 Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* **19:** 950-958.

ANISIMOVA, M., R. NIELSEN and Z. YANG, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164:** 1229-1236.

AWADALLA, P., 2003 The evolutionary genomics of pathogen recombination. *Nat. Rev. Genet.* **4:** 50-60.

AYCOCK, W. L., and J. H. MUELLER, 1950 Meningococcus carrier rates and meningitis incidence. *Bacteriol. Rev.* **14:** 115-160.

BAHLO, M., and R. C. GRIFFITHS, 2000 Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* **57:** 79-95.

BANDELT, H. J., and A. W. DRESS, 1992 Split decomposition: a new and useful

approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* **1:** 242-252.

BAUM, J., A. W. THOMAS and D. J. CONWAY, 2003 Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. *Genetics* **163:** 1327-1336.

BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population genetics. *Genetics* **162:** 2025-2035.

BEERLI, P., and J. FELSENSTEIN, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152:** 763-773.

BEERLI, P., and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. U.S.A.* **98:** 4563-4568.

BENNETT, J. S., D. T. GRIFFITHS, N. D. MCCARTHY, K. L. SLEEMAN, K. A. JOLLEY *et al.*, 2005 Genetic diversity and carriage dynamics of *Neisseria lactamica* in infants. *Infect. Immun.* **73:** 2424-2432.

BIELAWSKI, J. P., and Z. YANG, 2001 Positive and negative selection in the DAZ gene family. *Mol. Biol. Evol.* **18:** 523-529.

BISHOP, J. G., A. M. DEAN and T. MITCHELL-OLDS, 2000 Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci. U.S.A.* **97:** 5322-5327.

BJUNE, G., E. A. HOIBY, J. K. GRONNESBY, O. ARNESEN, J. H. FREDRIKSEN *et al.*, 1991 Effect of outer membrane vesicle vaccine against group B meningococcal disease in Norway. *Lancet* **338:** 1093-1096.

BLAKEBROUGH, I. S., B. M. GREENWOOD, H. C. WHITTLE, A. K. BRADLEY and H. M. GILLES, 1982 The epidemiology of infections due to *Neisseria meningitidis* and *Neisseria lactamica* in a northern Nigerian community. *J Infect Dis* **146:** 626-637.

BOWDEN, R., H. SAKAOKA, P. DONNELLY and R. WARD, 2004 High recombination rate in herpes simplex virus type 1 natural populations suggests significant co-infection. *Infect. Genet. Evol.* **4:** 115-123.

BROOME, C. V., 1986 The carrier state: *Neisseria meningitidis*. *J. Antimicrob. Chemother.* **18 Suppl A:** 25-34.

BROWN, A. H. D., M. W. FELDMAN and E. NEVO, 1980 Multilocus structure of natural

populations of *Hordeum spontaneum*. *Genetics* **96:** 523-536.

BUCKEE, C. O., K. KOELLE, M. J. MUSTARD and S. GUPTA, 2004 The effects of host contact network structure on pathogen diversity and strain structure. *Pro.c Natl. Acad. Sci. U.S.A.* **101:** 10839-10844.

BUSH, R. M., W. M. FITCH, C. A. BENDER and N. J. COX, 1999 Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16:** 1457-1465.

CARBONE, I., Y. C. LIU, B. I. HILLMAN and M. G. MILGROOM, 2004 Recombination and migration of Cryphonectria hypovirus 1 as inferred from gene genealogies and the coalescent. *Genetics* **166:** 1611-1629.

CARTWRIGHT, K., 1995 Meningococcal carriage and disease in *Meningococcal Disease*, edited by K. CARTWRIGHT. John Wiley & Sons Ltd, Chichester.

CARTWRIGHT, K. A., J. M. STUART, D. M. JONES and N. D. NOAH, 1987 The Stonehouse survey: nasopharyngeal carriage of meningococci and *Neisseria lactamica*. *Epidemiol. Infect.* **99:** 591-601.

CAUGANT, D. A., 2001 Global trends in meningococcal disease in *Meningococcal Disease. Methods and Protocols*, edited by A. J. POLLARD and M. C. J. MAIDEN. Humana Press Inc., Totowa, New Jersey.

CAUGANT, D. A., L. O. FROHOLM, K. BOVRE, E. HOLTEN, C. E. FRASCH *et al.*, 1986 Intercontinental spread of a genetically distinctive complex of clones of *Neisseria meningitidis* causing epidemic disease. *Proc. Natl. Acad. Sci. U.S.A.* **83:** 4927-4931.

CAUGANT, D. A., E. A. HOIBY, P. MAGNUS, O. SCHEEL, T. HOEL *et al.*, 1994 Asymptomatic carriage of *Neisseria meningitidis* in a randomly sampled population. *J. Clin. Microbiol.* **32:** 323-330.

CAUGANT, D. A., B. E. KRISTIANSEN, L. O. FROHOLM, K. BOVRE and R. K. SELANDER, 1988 Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers. *Infect. Immun.* **56:** 2060-2068.

CAUGANT, D. A., L. F. MOCCA, C. E. FRASCH, L. O. FROHOLM, W. D. ZOLLINGER *et al.*, 1987 Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern. *J. Bacteriol.* **169:** 2781-2792.

CENTERS FOR DISEASE CONTROL AND PREVENTION, U.S.A., 2000 Active Bacterial Core Surveillance (ABCs) Report Emerging Infections Program Network.

http://www.cdc.gov/abcs/.

CHOISY, M., C. H. WOELK, J. F. GUEGAN and D. L. ROBERTSON, 2004 Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J. Virol.* **78:** 1962-1970.

CLAUS, H., M. C. MAIDEN, R. MAAG, M. FROSCH and U. VOGEL, 2002 Many carried meningococci lack the genes required for capsule synthesis and transport. *Microbiology* **148:** 1813-1819.

CLAUS, H., M. C. MAIDEN, D. J. WILSON, N. D. MCCARTHY, K. A. JOLLEY *et al.*, 2005 Genetic analysis of meningococci carried by children and young adults. *J. Infect. Dis.* **191:** 1263-1271.

COOP, G., and R. C. GRIFFITHS, 2004 Ancestral inference on gene trees under selection. *Theor. Popul. Biol.* **66:** 219-232.

DAWKINS, R., 1982 *The Extended Phenotype*. WH Freeman, San Francisco.

DE IORIO, M., R. C. GRIFFITHS, R. LEBLOIS and F. ROUSSET, 2005 Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popul. Biol.* **68:** 41-53.

DE MORAES, J. C., B. A. PERKINS, M. C. CAMARGO, N. T. HIDALGO, H. A. BARBOSA *et al.*, 1992 Protective efficacy of a serogroup B meningococcal vaccine in São Paulo, Brazil. *Lancet* **340:** 1074-1078.

DE OLIVEIRA, T., M. SALEMI, M. GORDON, A. M. VANDAMME, E. J. VAN RENSBURG *et al.*, 2004 Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? *Genetics* **167:** 1047-1058.

DE WALS, P., C. GILQUIN, S. DE MAEYER, A. BOUCKAERT, A. NOEL *et al.*, 1983 Longitudinal study of asymptomatic meningococcal carriage in two Belgian populations of schoolchildren. *J. Infect.* **6:** 147-156.

DERRICK, J. P., R. URWIN, J. SUKER, I. M. FEAVERS and M. C. J. MAIDEN, 1999 Structural and evolutionary inference from molecular variation in Neisseria porins. *Infect. Immun. Infect. Immun.* **67:** 2406-2413.

DOPAZO, J., A. DRESS and A. VON HAESELER, 1993 Split decomposition: a technique to analyze viral evolution. *Proc. Natl. Acad. Sci. U.S.A.* **90:** 10320-10324.

DRUMMOND, A. J., G. K. NICHOLLS, A. G. RODRIGO and W. SOLOMON, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161:** 1307-

1320.

DRUMMOND, A., O. G. PYBUS and A. RAMBAUT, 2003a Inference of viral evolutionary rates from molecular sequences. *Adv. Parasitol.* **54:** 331-358.

DRUMMOND, A. J., O. G. PYBUS, A. RAMBAUT, R. FORSBERG and A. G. RODRIGO, 2003b Measurably evolving populations. *Trends Ecol. Evol.* **18:** 481-488.

EWENS, W., 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3:** 87-112.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131:** 479-491.

FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164:** 1567-1587.

FALUSH, D., T. WIRTH, B. LINZ, J. K. PRITCHARD, M. STEPHENS *et al.*, 2003 Traces of human migrations in *Helicobacter pylori* populations. *Science* **299:** 1582-1585.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405-1413.

FEARNHEAD, P., and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. *Genetics* **159:** 1299-1318.

FEAVERS, I. M., A. J. FOX, S. GRAY, D. M. JONES and M. C. MAIDEN, 1996 Antigenic diversity of meningococcal outer membrane protein PorA has implications for epidemiological analysis and vaccine design. *Clin. Diagn. Lab. Immunol.* **3:** 444-450.

FEIL, E., G. CARPENTER and B. G. SPRATT, 1995 Electrophoretic variation in adenylate kinase of *Neisseria meningitidis* is due to inter- and intraspecies recombination. *Proc. Natl. Acad. Sci. U.S.A.* **92:** 10535-10539.

FEIL, E., J. ZHOU, J. MAYNARD SMITH and B. G. SPRATT, 1996 A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal Neisseria species: evidence for extensive interspecies recombination within *adk*. *J. Mol. Evol.* **43:** 631-640.

FEIL, E. J., M. C. MAIDEN, M. ACHTMAN and B. G. SPRATT, 1999 The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol. Biol. Evol.* **16:** 1496-1502.

FEIL, E. J., and B. G. SPRATT, 2001 Recombination and the population structure of bacterial pathogens. *Annu. Rev. Microbiol.* **55:** 561-590.

FEIL, E. J., E. C. HOLMES, M. C. ENRIGHT, D. E. BESSEN, N. P. J. DAY, M.-S. CHAN, D. W. HOOD, J. ZHOU, and B. G. SPRATT, 2001 Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U.S.A.* **98:** 182-187.

FEIL, E. J., B. C. LI, D. M. AANENSEN, W. P. HANAGE and B. G. SPRATT, 2004 eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186:** 1518-1530.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17:** 368-376.

FERGUSON, N., R. ANDERSON and S. GUPTA, 1999 The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens. *Proc. Natl. Acad. Sci. U.S.A* **96:** 790-794.

FERGUSON, N. M., A. P. GALVANI and R. M. BUSH, 2003 Ecological and immunological determinants of influenza evolution. *Nature* **422:** 428-433.

FILIP, L. C., and N. I. MUNDY, 2004 Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates. *Mol. Biol. Evol.* **21:** 1504-1511.

FISHER, R. A., 1925 *Statistical Methods for Research Workers*. Oliver and Boyd, London.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

FITCH, W. M., R. M. BUSH, C. A. BENDER and N. J. COX, 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. U.S.A.* **94:** 7712-7718.

FORD, M. J., 2001 Molecular evolution of transferrin: evidence for positive selection in salmonids. *Mol. Biol. Evol.* **18:** 639-647.

FRASCH, C. E., 1995 Meningococcal vaccines: past, present and future in *Meningococcal Disease*, edited by K. CARTWRIGHT. John Wiley & Sons Ltd, Chichester.

FRASER, C., W. P. HANAGE and B. G. SPRATT, 2005 Neutral microepidemic evolution of bacterial pathogens. *Proc. Natl. Acad. Sci. U.S.A.* **102:** 1968-1973.

FU, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. *Genetics* **143:** 557-570.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133:** 693-709.

FU, Y. X., and W. H. LI, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14:** 195-199.

GOJOBORI, T., and M. NEI, 1986 Relative contributions of germline gene variation and somatic mutation to immunoglobulin diversity in the mouse. *Mol. Biol. Evol.* **3:** 156-167.

GOLDACRE, M. J., S. E. ROBERTS and D. YEATES, 2003 Case fatality rates for meningococcal disease in an English population, 1963-98: database study. *BMJ* **327:** 596-597.

GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725-736.

GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82:** 711-732.

GRENFELL, B. T., O. G. PYBUS, J. R. GOG, J. L. WOOD, J. M. DALY *et al.*, 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303:** 327-332.

GRIFFITHS, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3:** 479-502.

GRIFFITHS, R. C., and P. MARJORAM, 1997 An ancestral recombination graph, pp. 257-270 in *Progress in Population Genetics and Human Evolution*, edited by P. DONNELLY and S. TAVARE. Springer, Verlag.

GRIMMETT, G., and D. STIRZAKER, 2001 *Probability and Random Processes* 3rd edition. OUP, Oxford.

GUPTA, S., M. C. MAIDEN, I. M. FEAVERS, S. NEE, R. M. MAY *et al.*, 1996 The maintenance of strain structure in populations of recombining infectious agents. *Nat. Med.* **2:** 437-442.

HARAGUCHI, Y., and A. SASAKI, 1997 Evolutionary pattern of intra-host pathogen antigenic drift: effect of cross-reactivity in immune response. *Philos. Trans. R. Soc. Lond. Ser. B* **352:** 11-20.

HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57:** 97-109.

HEATH, P. T., 1998 Haemophilus influenzae type b conjugate vaccines: a review of efficacy data. *Pediatr. Infect. Dis. J* **17:** S117-122.

HEIN, J., M. H. SCHIERUP and C. WIUF, 2005 *Gene Genealogies, Varation and Evolution*. Oxford University Press, Oxford.

HILL, W. G., and A. ROBERTSON, 1968 The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60:** 615-628.

HOLMES, E. C., R. URWIN and M. C. MAIDEN, 1999 The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.* **16:** 741-749.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23:** 183-201.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*. Oxford University Press, Oxford.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. *Genetics* **159:** 1805-1817.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. *Genetics* **111:** 147-164.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153-159.

HUELSENBECK, J. P., and K. A. DYER, 2004 Bayesian estimation of positively selected sites. *J. Mol. Evol.* **58:** 661-672.

HUSMEIER, D., and G. MCGUIRE, 2003 Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol. Biol. Evol.* **20:** 315-337.

ICHHPUJANI, R. L., R. MOHAN, S. S. GROVER, P. R. JOSHI and S. KUMARI, 1990 Nasopharyngeal carriage of *Neisseria meningitidis* in general population and meningococcal disease. *J. Commun. Dis.* **22:** 264-268.

JOLLEY, K. A., M. S. CHAN and M. C. J. MAIDEN, 2004 mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* **5:** 86.

JOLLEY, K. A., J. KALMUSOVA, E. J. FEIL, S. GUPTA, M. MUSILEK *et al.*, 2000 Carried meningococci in the Czech Republic: A diverse recombining population. *J. Clin. Microbiol.* **38:** 4492-4498.

———, 2002 Author's correction. *J. Clin. Microbiol.* **40:** 3549-3550.

JOLLEY, K. A., D. J. WILSON, P. KRIZ, G. MCVEAN and M. C. J. MAIDEN, 2005 The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Mol. Biol. Evol.* **22:** 562-569.

KIMURA, M., 1968 Evolutionary rate at the molecular level. *Nature* **217:** 624-626.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111-120.

KIMURA, M., and G. H. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49:** 561-576.

KINGMAN, J. F. C., 1982a On the genealogy of large populations. *Journal of Applied Probability* **19A:** 27-43.

KINGMAN, J. F. C., 1982b The coalescent. *Stochastic Process. Appl.* **13:** 235-248.

KORBER, B., M. MULDOON, J. THEILER, F. GAO, R. GUPTA *et al.*, 2000 Timing the ancestor of the HIV-1 pandemic strains. *Science* **288:** 1789-1796.

KOSAKOVSKY POND, S. L., and S. D. FROST, 2005 Not so different after all: a comparison of methods for detecting amino-acid sites under selection. *Mol. Biol. Evol.* **22:** 1208-1222.

KREITMAN, M., 2000 Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1:** 539-559.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140:** 1421-1430.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149:** 429-434.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. *Genetics* **156:** 1393-1401.

LANE, R. P., J. YOUNG, T. NEWMAN and B. J. TRASK, 2004 Species specificity in rodent pheromone receptor repertoires. *Genome Res* **14:** 603-608.

LAPEYSSONIE, L., 1963 La méningite cérébrospinale en Afrique. *Bulletin of the World Health Organization* **28 (Suppl):** 3-114.

LAPORTE, V., and B. CHARLESWORTH, 2002 Effective population size and population subdivision in demographically structured populations. *Genetics* **162:** 501-519.

LEMEY, P., O. G. PYBUS, A. RAMBAUT, A. J. DRUMMOND, D. L. ROBERTSON *et al.*, 2004 The molecular population genetics of HIV-1 group O. *Genetics* **167:** 1059-1068.

LEMEY, P., O. G. PYBUS, B. WANG, N. K. SAKSENA, M. SALEMI *et al.*, 2003 Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl. Acad. Sci. U.S.A.* **100:** 6588-6592.

LESLIE, A. J., K. J. PFAFFEROTT, P. CHETTY, R. DRAENERT, M. M. ADDO *et al.*, 2004 HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.* **10:** 282-289.

LESSARD, S., and J. WAKELEY, 2004 The two-locus ancestral graph in a subdivided population: convergence as the number of demes grows in the island model. *J. Math. Biol.* **48:** 275-292.

LEVIN, B. R., and J. J. BULL, 1994 Short-sighted evolution and the virulence of pathogenic microorganisms. *Trends Microbiol.* **2:** 76-81.

LEVINS, R., 1968 *Evolution in Changing Environments*. Princeton University Press, Princeton, N.J.

LEVINS, R., 1969 Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.* **15**.

LEVY, D. N., G. M. ALDROVANDI, O. KUTSCH and G. M. SHAW, 2004 Dynamics of HIV-1 recombination in its natural target cells. *Proc. Natl. Acad. Sci. U.S.A* **101:** 4204-4209.

LI, K. S., Y. GUAN, J. WANG, G. J. SMITH, K. M. XU *et al.*, 2004 Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430:** 209-213.

LI, N., and M. STEPHENS, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165:** 2213-2233.

LI, W. H., C. I. WU and C. C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2:** 150-174.

LINZ, B., M. SCHENKER, P. ZHU and M. ACHTMAN, 2000 Frequent interspecific genetic exchange between commensal Neisseriae and *Neisseria meningitidis*. *Mol. Microbiol.* **36:** 1049-1058.

LOADER, C. R., 1996 Local likelihood density estimation. *Annals of Statistics* **24:**

1602-1618.

LONG, M., and C. H. LANGLEY, 1993 Natural selection and the origin of *jingwei*, a chimeric processed functional gene in Drosophila. *Science* **260:** 91-95.

LORENZ, M. G., and W. WACKERNAGEL, 1994 Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* **58:** 563-602.

LYNCH, M., and J. S. CONERY, 2003 The origins of genome complexity. *Science* **302:** 1401-1404.

LYTHGOE, K. A., 2002 Effects of acquired immunity and mating strategy on the genetic structure of parasite populations. *Am. Nat.* **159:** 519-529.

MAIDEN, M. C., J. A. BYGRAVES, E. FEIL, G. MORELLI, J. E. RUSSELL *et al.*, 1998 Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 3140-3145.

MAIDEN, M. C. J., 2002 Population structure of *Neisseria meningitidis* in *Emerging strategies in the fight against meningitis: molecular and cellular aspects*, edited by C. FERREIRÓS, M. T. CRIADO and J. VÁZQUEZ. Horizon Scientific Press, Wymondham, Norfolk.

MANTEL, N., 1967 The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27:** 209-220.

MARJORAM, P., J. MOLITOR, V. PLAGNOL and S. TAVARE, 2003 Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U.S.A.* **100:** 15324-15328.

MARTZ, E., 2002 Protein Explorer: easy yet powerful macromolecular visualization. *Trends Biochem. Sci.* **27:** 107-109.

MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theor. Popul. Biol.* **1:** 273-306.

MASSINGHAM, T., and N. GOLDMAN, 2005 Detecting amino Acid sites under positive selection and purifying selection. *Genetics* **169:** 1753-1762.

MAYNARD SMITH, J., E. J. FEIL and N. H. SMITH, 2000 Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays* **22:** 1115-1122.

MAYNARD SMITH, J., N. H. SMITH, M. O'ROURKE and B. G. SPRATT, 1993 How clonal are bacteria? *Proc. Natl. Acad. Sci. U.S.A.* **90:** 4384-4388.

MCCULLAGH, P., and J. A. NELDER, 1989 *Generalized Linear Models* 2nd edition. Chapman and Hall, London.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh

locus in Drosophila. *Nature* **351:** 652-654.

MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160:** 1231-1241.

MCVEAN, G. A. T., S. R. MYERS, S. HUNT, P. DELOUKAS, D. R. BENTLEY *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581-584.

MENINGITIS RESEARCH FOUNDATION, 2005 http://www.meningitis.org/.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER, 1953 Equations of state calculations by fast computing machine. *J. Chem. Phys.* **21:** 1087-1091.

MEUNIER, J., and A. EYRE-WALKER, 2001 The correlation between linkage disequilibrium and distance: Implications for recombination in hominid mitochondria. *Mol. Biol. Evol.* **18:** 2132-2135.

MIMS, C., J. PLAYFAIR, I. ROITT, D. WAKELIN and R. WILLIAMS, 1998 *Medical Microbiology*. Mosby, London.

MONDRAGON-PALOMINO, M., B. C. MEYERS, R. W. MICHELMORE and B. S. GAUT, 2002 Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* **12:** 1305-1315.

MOURY, B., 2004 Differential selection of genes of cucumber mosaic virus subgroups. *Mol. Biol. Evol.* **21:** 1602-1611.

MYERS, S. R., and R. C. GRIFFITHS, 2003 Bounds on the minimum number of recombination events in a sample history. *Genetics* **163:** 375-394.

NAKAMURA, Y., T. GOJOBORI and T. IKEMURA, 2000 Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28:** 292.

NATIONAL OFFICE OF STATISTICS, U. K., 2002 Mortality statistics general. http://www.statistics.gov.uk/.

NEI, M., and T. GOJOBORI, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418-426.

NIELSEN, R., 2001 Statistical tests of selective neutrality in the age of genomics. *Heredity* **86:** 641-647.

NIELSEN, R., 2002 Mapping mutations on phylogenies. *Syst. Biol.* **51:** 729-739.

NIELSEN, R., and J. P. HUELSENBECK, 2002 Detecting positively selected amino acid sites using posterior predictive P-values. *Pac. Symp. Biocomput.* pp. 576-588.

NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929-936.

NORDBORG, M., 2003 Coalescent theory, pp. 602-635 in *Handbook of Statistical Genetics*, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. Wiley, Chichester.

O'HAGAN, A., and J. FORSTER, 2004 *Kendall's Advanced Theory of Statistics*. Volume 2B *Bayesian Inference*. Arnold, London.

PANNELL, J. R., 2003 Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* **57:** 949-961.

PARKHILL, J., M. ACHTMAN, K. D. JAMES, S. D. BENTLEY, C. CHURCHER *et al.*, 2000 Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404:** 502-506.

PEEK, A. S., V. SOUZA, L. E. EGUIARTE and B. S. GAUT, 2001 The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (fimA) from Escherichia coli. *J. Mol. Evol.* **52:** 193-204.

PERKINS, B. A., K. JONSDOTTIR, H. BRIEM, E. GRIFFITHS, B. D. PLIKAYTIS *et al.*, 1998 Immunogenicity of two efficacious outer membrane protein-based serogroup B meningococcal vaccines among young adults in Iceland. *J. Infect. Dis.* **177:** 683-691.

POLLARD, A. J., D. SCHEIFELE and N. ROSENSTEIN, 2001 Epidemiology of meningococcal disease in North America in *Meningococcal Disease. Methods and Protocols*, edited by A. J. POLLARD and M. C. J. MAIDEN. Humana Press Inc., Totowa, New Jersey.

POLLEY, S. D., W. CHOKEJINDACHAI and D. J. CONWAY, 2003 Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics* **165:** 555-561.

POOLMAN, J. T., P. A. VAN DER LEY and J. TOMMASSEN, 1995 Surface structures and secreted products of meningococci in *Meningococcal Disease*, edited by K. CARTWRIGHT. John Wiley & Sons Ltd, Chichester.

POSADA, D., K. A. CRANDALL and E. C. HOLMES, 2002 Recombination in evolutionary

genomics. *Annu. Rev. Genet.* **36:** 75-97.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING and F. B. P., 2002 *Numerical Recipes in C++. The Art of Scientific Computing.* Cambridge University Press, Cambridge.

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69:** 1-14.

PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN and M. W. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16:** 1791-1798.

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. *Genetics* **164:** 1667-1676.

PYBUS, O. G., M. A. CHARLESTON, S. GUPTA, A. RAMBAUT, E. C. HOLMES *et al.*, 2001 The epidemic behavior of the hepatitis C virus. *Science* **292:** 2323-2325.

PYBUS, O. G., A. J. DRUMMOND, T. NAKANO, B. H. ROBERTSON and A. RAMBAUT, 2003 The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol.* **20:** 381-387.

RABINER, L. R., 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77:** 257-286.

RAGHUNATHAN, P. L., S. A. BERNHARDT and N. E. ROSENSTEIN, 2004 Opportunities for control of meningococcal disease in the United States. *Annu. Rev. Med.* **55:** 333-353.

RAMBAUT, A., D. POSADA, K. A. CRANDALL and E. C. HOLMES, 2004 The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5:** 52-61.

RICH, S. M., M. C. LICHT, R. R. HUDSON and F. J. AYALA, 1998 Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A* **95:** 4425-4430.

ROBBINS, H., 1956 An empirical Bayes approach to statistics in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* edited by J. NEYMAN. University of California Press, Berkeley.

ROBBINS, K. E., P. LEMEY, O. G. PYBUS, H. W. JAFFE, A. S. YOUNGPAIROJ *et al.*, 2003 U.S. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.* **77:** 6359-6366.

RODRÍGUEZ, F., J. L. OLIVER, A. MARIN and J. R. MEDINA, 1990 The general stochastic

model of nucleotide substitution. *J. Theor. Biol.* **142:** 485-501.

RONQUIST, F., and J. P. HUELSENBECK, 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

ROSENSTEIN, N., O. LEVINE, J. P. TAYLOR, D. EVANS, B. D. PLIKAYTIS *et al.*, 1998 Efficacy of meningococcal vaccine and barriers to vaccination. *JAMA* **279:** 435-439.

RUBIN, D. B., 1984 Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12:** 1151-1172.

SAINUDIIN, R., W. S. WONG, K. YOGEESWARAN, J. B. NASRALLAH, Z. YANG *et al.*, 2005 Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J. Mol. Evol.* **60:** 315-326.

SCHIERUP, M. H., and J. HEIN, 2000 Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156:** 879-891.

SCHNEIDER, S., D. ROESSLI and L. EXCOFFIER, 2000 Arlequin version 2.000: a software for population genetic data analysis. University of Geneva, Geneva.

SCHWARTZ, B., P. S. MOORE and C. V. BROOME, 1989 Global epidemiology of meningococcal disease. *Clin. Microbiol. Rev.* **2 Suppl:** S118-124.

SELANDER, R. K., D. A. CAUGANT, H. OCHMAN, J. M. MUSSER, M. N. GILMOUR *et al.*, 1986 Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl. Environ. Microbiol.* **51:** 873-884.

SEXTON, K., D. LENNON, P. OSTER, S. CRENGLE, D. MARTIN *et al.*, 2004 The New Zealand Meningococcal Vaccine Strategy: a tailor-made vaccine to combat a devastating epidemic. *N.Z. Med. J.* **117:** U1015.

SHERIDAN, I., O. G. PYBUS, E. C. HOLMES and P. KLENERMAN, 2004 High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *J. Virol.* **78:** 3447-3454.

SHPAER, E. G., and J. I. MULLINS, 1993 Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses. *J. Mol. Evol.* **37:** 57-65.

SHRINER, D., D. C. NICKLE, M. A. JENSEN and J. I. MULLINS, 2003 Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.* **81:** 115-121.

SHRINER, D., R. SHANKARAPPA, M. A. JENSEN, D. C. NICKLE, J. E. MITTLER *et al.*, 2004 Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. *Genetics* **166:** 1155-1164.

SIERRA, V. G., C. CAMPA, L. GARCIA *et al.*, 1991 Efficacy evaluation of the Cuban vaccine VA-MENGOC-BC against disease caused by serogroup B *Neisseria meningitidis*, pp. 129-134 in *Neisseria 1990*, edited by M. ACHTMAN. Walter de Gruyter, Berlin.

SMITH, D. J., A. S. LAPEDES, J. C. DE JONG, T. M. BESTEBROER, G. F. RIMMELZWAAN *et al.*, 2004 Mapping the antigenic and genetic evolution of influenza virus. *Science* **305:** 371-376.

SMITH, N. H., J. MAYNARD SMITH and B. G. SPRATT, 1995 Sequence evolution of the *porB* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis* - evidence of positive Darwinian selection. *Mol. Biol. Evol.* **12:** 363-370.

SNAPE, M. D., and A. J. POLLARD, 2005 Meningococcal polysaccharide-protein conjugate vaccines. *Lancet Infect. Dis.* **5:** 21-30.

SNEATH, P. H. A., and R. R. SOKAL, 1973 *Numerical Taxonomy*. WH Freeman, San Francisco.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry* 3rd edition. WH Freeman, New York.

SORIANO-GABARRÓ, M., N. ROSENSTEIN and F. M. LAFORCE, 2004 Evaluation of serogroup A meningococcal vaccines in Africa: a demonstration project. *J. Health Popul. Nutr.* **22:** 275-285.

STEINHAUER, D. A., and J. J. SKEHEL, 2002 Genetics of influenza viruses. *Annu. Rev. Genet.* **36:** 305-332.

STEPHENS, M., 2003 Inference under the coalescent in *Handbook of Statistical Genetics* 2nd edition, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. John Wiley & Sons, Ltd, Chichester.

STEPHENS, M., and P. DONNELLY, 2000 Inference in molecular population genetics. *J. R. Stat. Soc. B* **62:** 605-655.

STEPHENS, M., and P. SCHEET, 2005 Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76:** 449-462.

STOLLENWERK, N., M. C. MAIDEN and V. A. JANSEN, 2004 Diversity in pathogenicity can cause outbreaks of meningococcal disease. *Proc. Natl. Acad. Sci. U.S.A.* **101:** 10229-10234.

STRIMMER, K., K. FORSLUND, B. HOLLAND and V. MOULTON, 2003 A novel exploratory method for visual recombination detection. *Genome Biol.* **4:** R33.

STUART, J. M., 2001 Managing outbreaks: the public health response in *Meningococcal Disease. Methods and Protocols*, edited by A. J. POLLARD and M. C. J. MAIDEN. Humana Press Inc., Totowa, New Jersey.

STUMPF, M. P. H., and G. A. T. MCVEAN, 2003 Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.* **4:** 959-968.

SU, X., M. T. FERDIG, Y. HUANG, C. Q. HUYNH, A. LIU *et al.*, 1999 A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286:** 1351-1353.

SUAREZ, D. L., D. A. SENNE, J. BANKS, I. H. BROWN, S. C. ESSEN *et al.*, 2004 Recombination resulting in virulence shift in avian influenza outbreak, Chile. *Emerg. Infect. Dis.* **10:** 693-699.

SUCHARD, M. A., R. E. WEISS, K. S. DORMAN and J. S. SINSHEIMER, 2002 Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Syst. Biol.* **51:** 715-728.

SUZUKI, Y., 2004 New methods for detecting positive selection at single amino acid sites. *J. Mol. Evol.* **59:** 11-19.

SUZUKI, Y., and T. GOJOBORI, 1999 A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16:** 1315-1328.

SWANSON, W. J., R. NIELSEN and Q. YANG, 2003 Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.* **20:** 18-20.

SWANSON, W. J., Z. YANG, M. F. WOLFNER and C. F. AQUADRO, 2001 Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. U.S.A.* **98:** 2509-2514.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-595.

TAVARE, S., D. J. BALDING, R. C. GRIFFITHS and P. DONNELLY, 1997 Inferring coalescence times from DNA sequence data. *Genetics* **145:** 505-518.

THORNE, J. L., H. KISHINO and J. FELSENSTEIN, 1991 An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:** 114-124.

THORNE, J. L., H. KISHINO and J. FELSENSTEIN, 1992 Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34:** 3-16.

TWIDDY, S. S., C. H. WOELK and E. C. HOLMES, 2002 Phylogenetic evidence for

adaptive evolution of dengue viruses in nature. *J. Gen. Virol.* **83:** 1679-1689.

URWIN, R., E. C. HOLMES, A. J. FOX, J. P. DERRICK and M. C. J. MAIDEN, 2002 Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. *Mol. Biol. Evol.* **19:** 1686-1694.

URWIN, R., and M. C. MAIDEN, 2003 Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.* **11:** 479-487.

URWIN, R., J. E. RUSSELL, E. A. THOMPSON, E. C. HOLMES, I. M. FEAVERS *et al.*, 2004 Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infect. Immun.* **72:** 5955-5962.

VOGEL, U., H. CLAUS and M. FROSCH, 2001 Capsular operons in *Meningococcal Disease. Methods and Protocols*, edited by A. J. POLLARD and M. C. J. MAIDEN. Humana Press Inc., Totowa, New Jersey.

WAKELEY, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.* **69:** 45-48.

WAKELEY, J., 2004 Metapopulation models for historical inference. *Mol. Ecol.* **13:** 865-875.

WAKELEY, J., and N. ALIACAR, 2001 Gene genealogies in a metapopulation. *Genetics* **159:** 893-905.

WALL, J. D., 2000 A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17:** 156-163.

WANG, J. F., D. A. CAUGANT, X. LI, X. HU, J. T. POOLMAN *et al.*, 1992 Clonal and antigenic analysis of serogroup A *Neisseria meningitidis* with particular reference to epidemiological features of epidemic meningitis in the People's Republic of China. *Infect. Immun.* **60:** 5267-5282.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7:** 256-276.

WEISS, G., and A. VON HAESELER, 1998 Inference of population history using a likelihood approach. *Genetics* **149:** 1539-1546.

WILKINSON, J. H., and C. REINSCH, 1971 *Handbook for Automatic Computation*. Volume II *Linear Algebra*. Springer-Verlag.

WILKINSON-HERBOTS, H. M., 1998 Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37:** 535-585.

WILSON, D. J., D. FALUSH and G. MCVEAN, 2005 Germs, genomes and genealogies. *Trends Ecol. Evol.* **20:** 39-45.

WILSON, I. J., M. E. WEALE and D. J. BALDING, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Statist. Soc. A* **166:** 155-201.

WINCKLER, W., S. R. MYERS, D. J. RICHTER, R. C. ONOFRIO, G. J. MCDONALD *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308:** 107-111.

WIUF, C., J. HEIN, 2000 The coalescent with gene conversion. *Genetics* **155:** 451-462.

WONG, W. S., Z. YANG, N. GOLDMAN and R. NIELSEN, 2004 Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041-1051.

WRIGHT, S., 1931 Evolution in Mendelian populations. *Genetics* **16:** 97-159.

WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. *Am. Nat.* **74:** 232-248.

YANG, W., J. P. BIELAWSKI and Z. YANG, 2003 Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J. Mol. Evol.* **57:** 212-221.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555-556.

YANG, Z., R. NIELSEN, N. GOLDMAN and A. M. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431-449.

YANG, Z., and W. J. SWANSON, 2002 Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19:** 49-57.

YAZDANKHAH, S. P., P. KRIZ, G. TZANAKAKI, J. KREMASTINOU, J. KALMUSOVA *et al.*, 2004 Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. *J. Clin. Microbiol.* **42:** 5146-5153.

ZANOTTO, P. M., E. G. KALLAS, R. F. DE SOUZA and E. C. HOLMES, 1999 Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* **153:** 1077-1089.

ZHOU, J., L. D. BOWLER and B. G. SPRATT, 1997 Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal Neisseria

species. *Mol. Microbiol.* **23:** 799-812.

ZHOU, J., and B. G. SPRATT, 1992 Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Mol. Microbiol.* **6:** 2135-2146.

ZHU, P., E. A. VAN DER, D. FALUSH, N. BRIESKE, G. MORELLI *et al.*, 2001 Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis. *Proc. Natl. Acad. Sci. U.S.A.* **98:** 5234-5239.

ZHUANG, J., A. E. JETZT, G. SUN, H. YU, G. KLARMANN *et al.*, 2002 Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J. Virol.* **76:** 11273-11282.

ZOLLINGER, W. D., J. BOSLEGO, E. MORAN, J. GARCIA, C. CRUZ *et al.*, 1991 Meningococcal serogroup B vaccine protection trial and follow-up studies in Chile. The Chilean National Committee for Meningococcal Disease. *NIPH Ann.* **14:** 211-212.