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

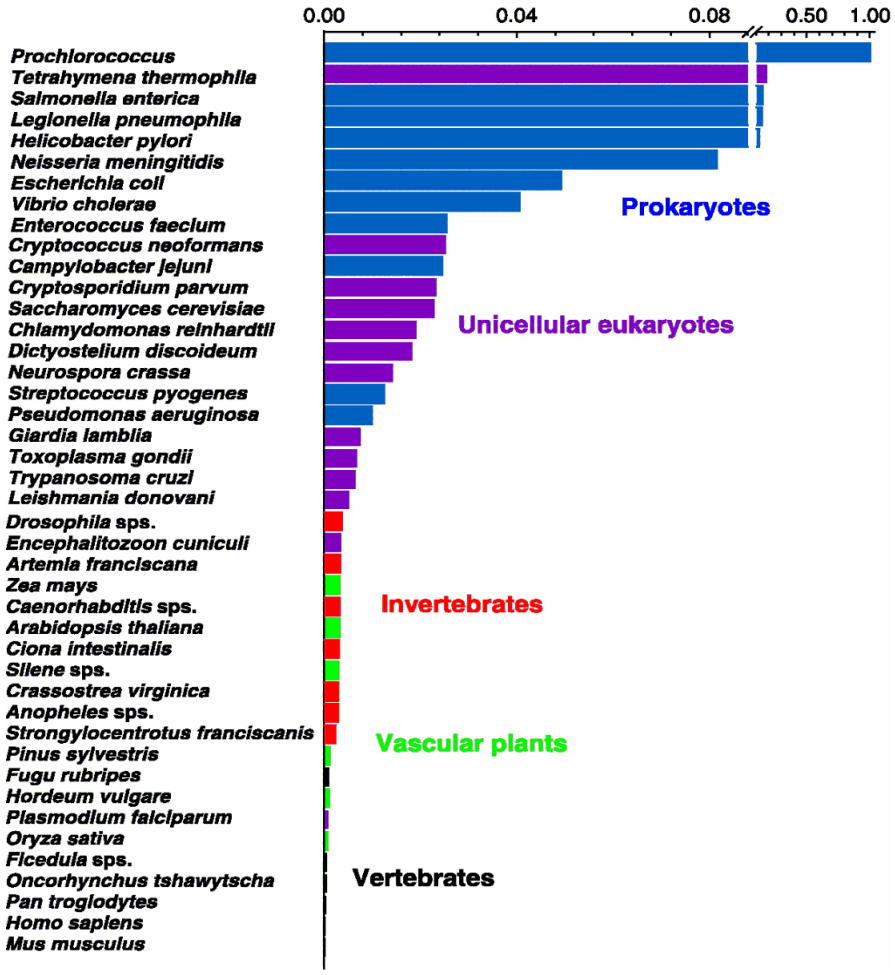

Figure 1 Estimates of the population mutation rate θ/P for different taxa. Source: Lynch and Conery (2003).

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*

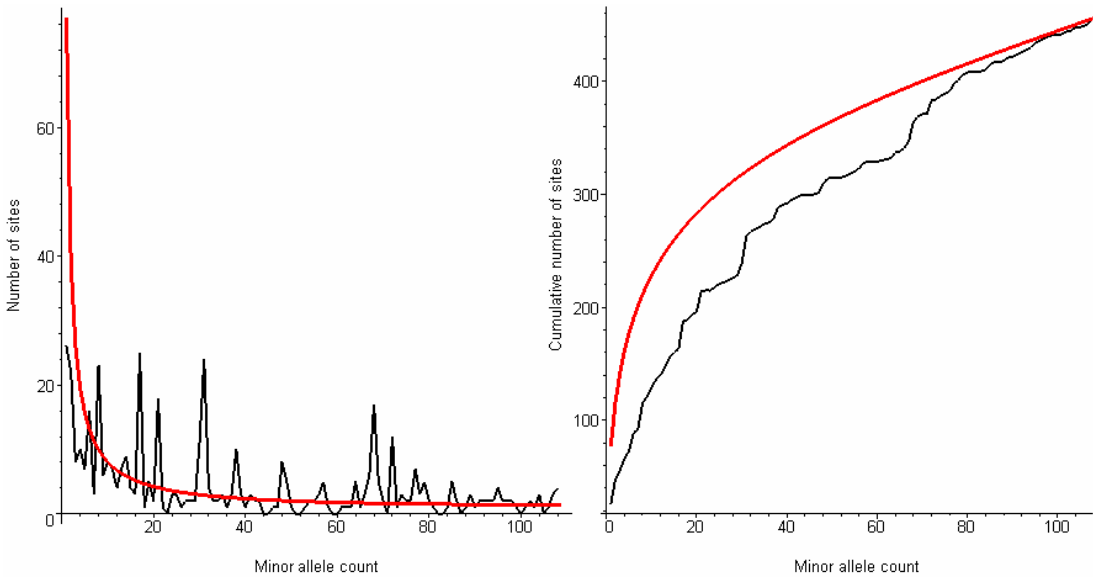

Figure 2 Observed distribution of minor allele count across biallelic segregating sites. In both figures the red line indicates the neutral expectation. Left: plot of the number of sites with a given minor allele count. Right: plot of the cumulative number of sites with a given minor allele count.

(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

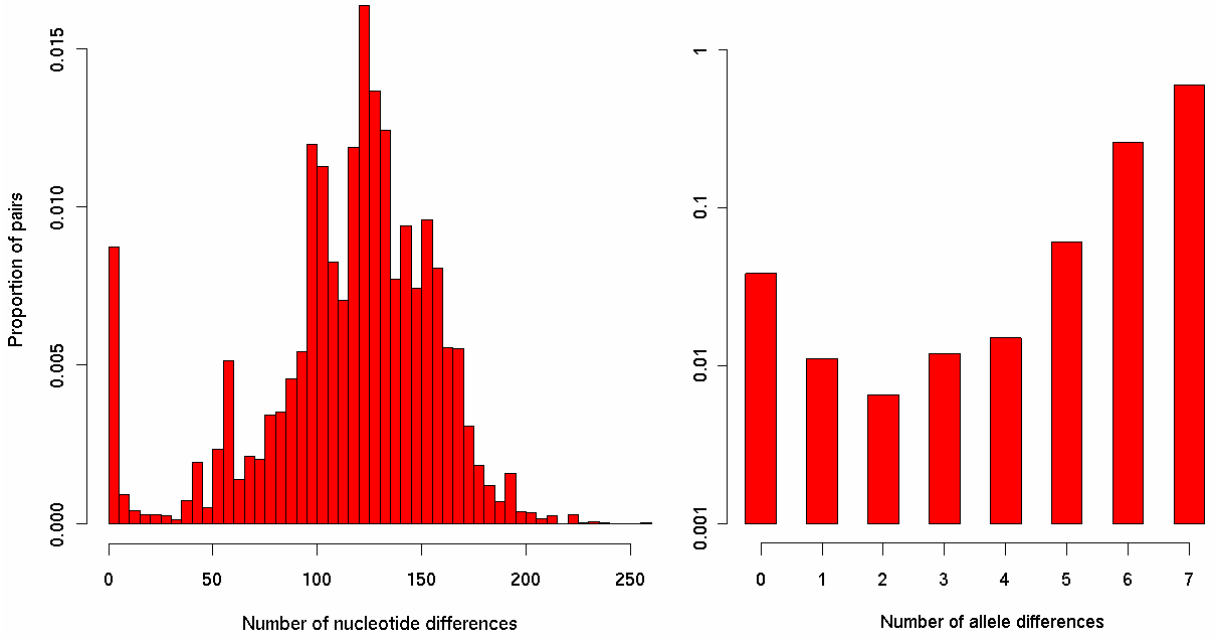

Figure 3 Mismatch distributions for isolates sequenced at the seven MLST loci. Left: histogram of the number of nucleotide differences between pairs of isolates, out of 3284 bp in total. Right: bar chart of the number of allele differences between pairs of isolates, out of 7 loci.

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

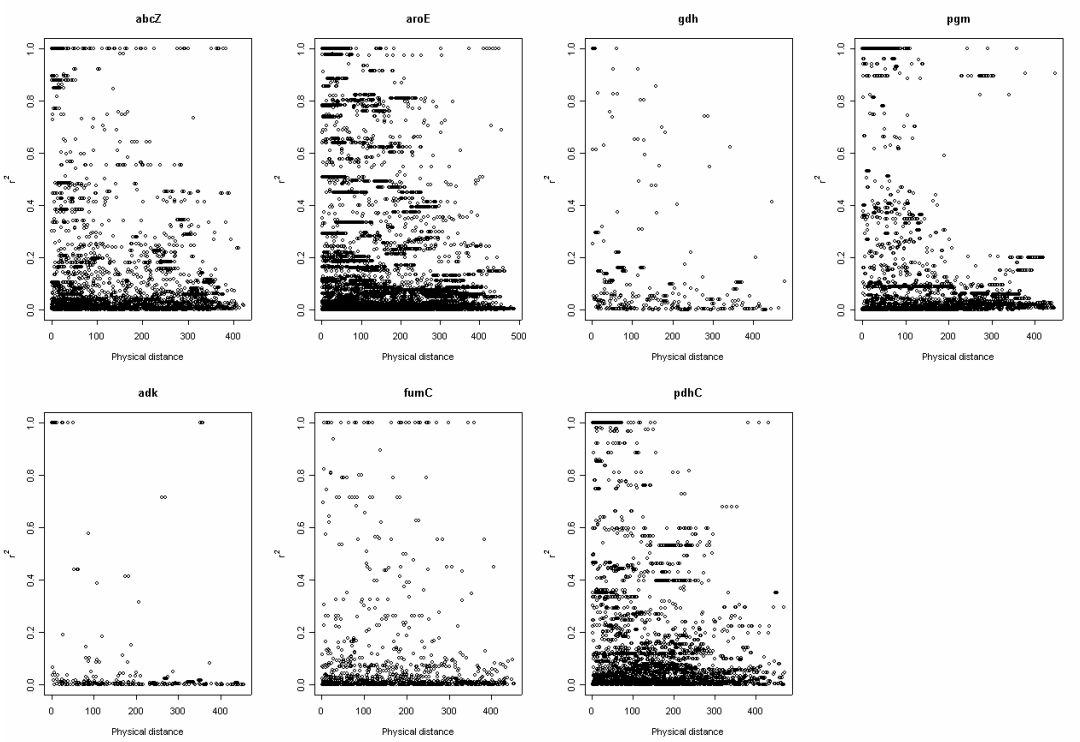

Figure 4 Breakdown in linkage disequilibrium, as measured by r^2, with physical distance in each of seven housekeeping loci from the Czech carriage study.

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

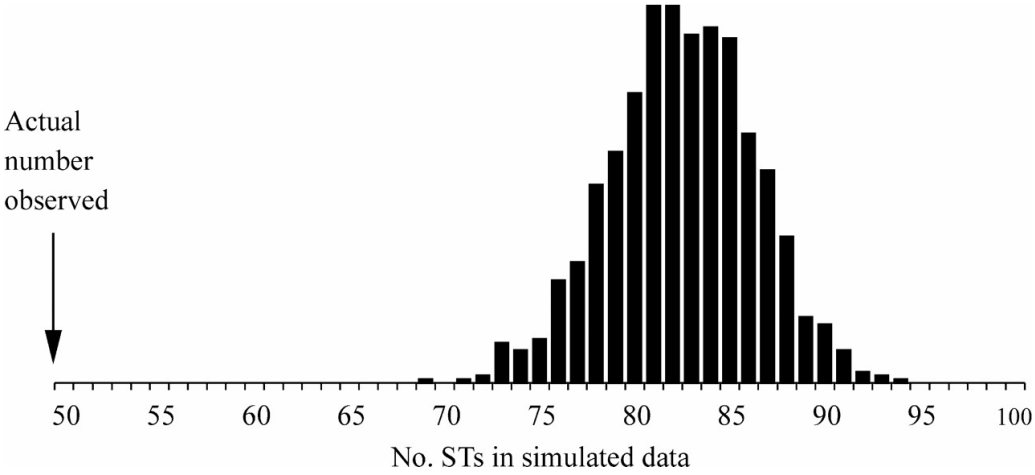

Figure 5 Null distribution of the number of unique haplotypes (STs) under the parameters estimated by LDhat for a sub-sample of 100 sequences[3]. The observed number was 50, which is outside the range of simulated values.

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.

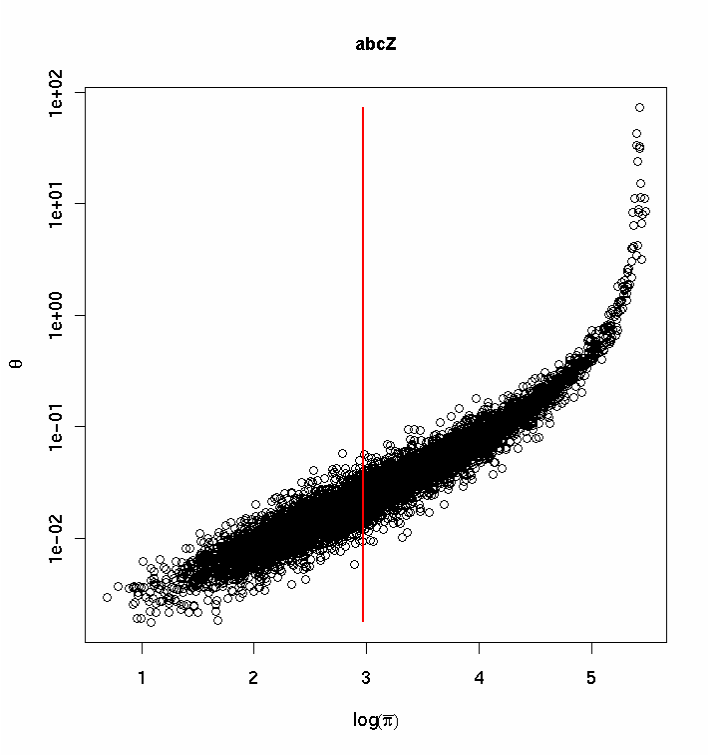

Figure 6 Joint posterior of $f(s_1(\mathbf{X}),\theta|S)$ for *abcZ*, with θ on a log scale. The red line indicates the observed value $S_1 = 2.98$. Locfit (Loader 1996) is used to estimate the conditional density of $f(\theta| s_1(\mathbf{X}) = S_1,S_2,S_3)$ along this line. See Figure 7.

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

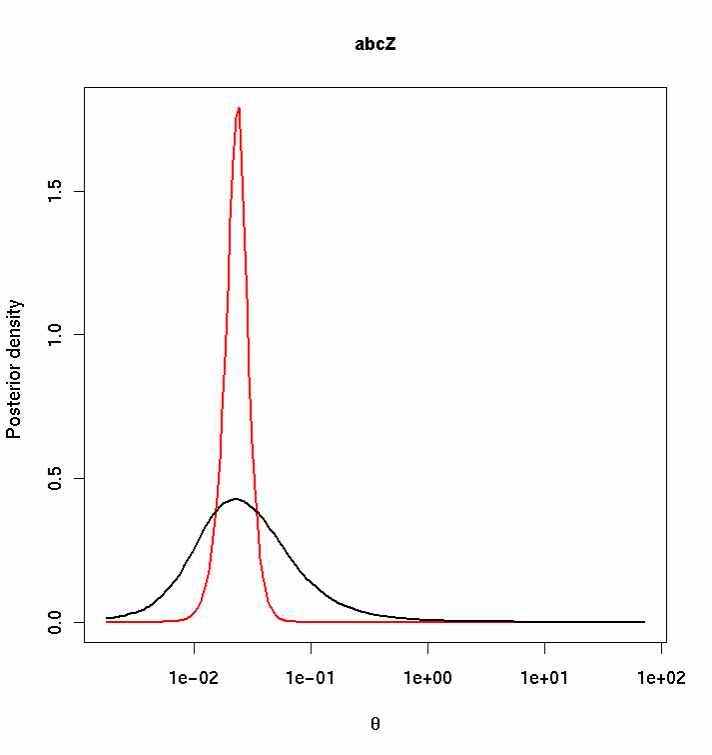

Figure 7 Black line: posterior of $f(\theta|S)$. Red line: posterior of $f(\theta|s_1(\mathbf{X})=S_1,S_2,S_3)$. Both were fit using locfit (Loader 1996). θ is on a log scale. By conditioning on the observed value S_1, a much tighter posterior is obtained.

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.

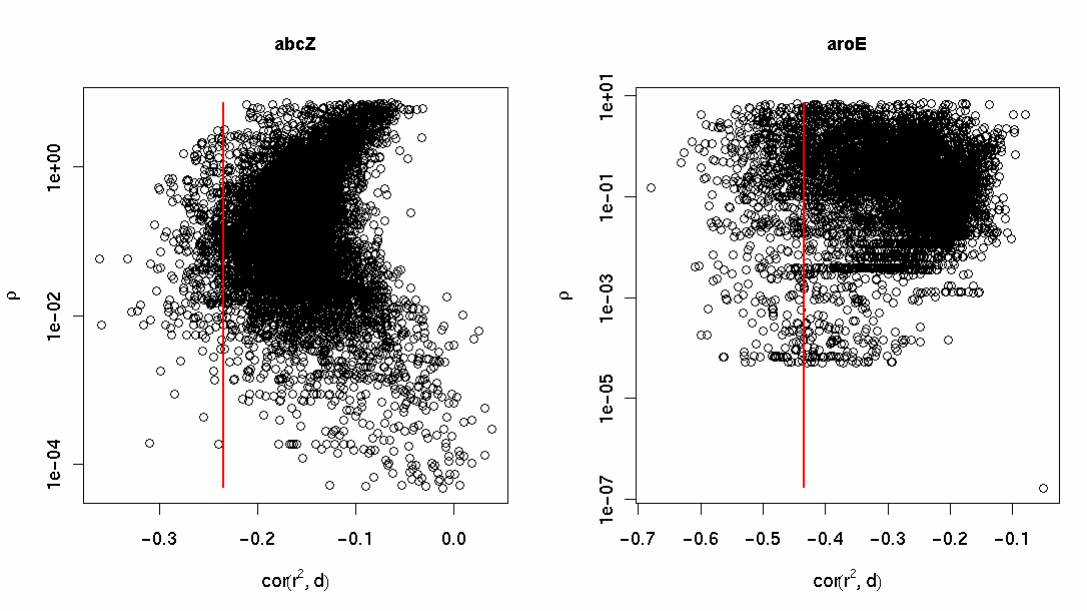

Figure 8 Scatterplots of $f(s_3(\mathbf{X}),\rho|S)$ for *abcZ* and *aroE* (ρ is on a log scale). The observed values of S_3 are marked with red lines. For *abcZ* the chain has mixed well, although the density is concentrated at smaller values of $s_3(\mathbf{X})$ than that observed. For *aroE* the observed 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

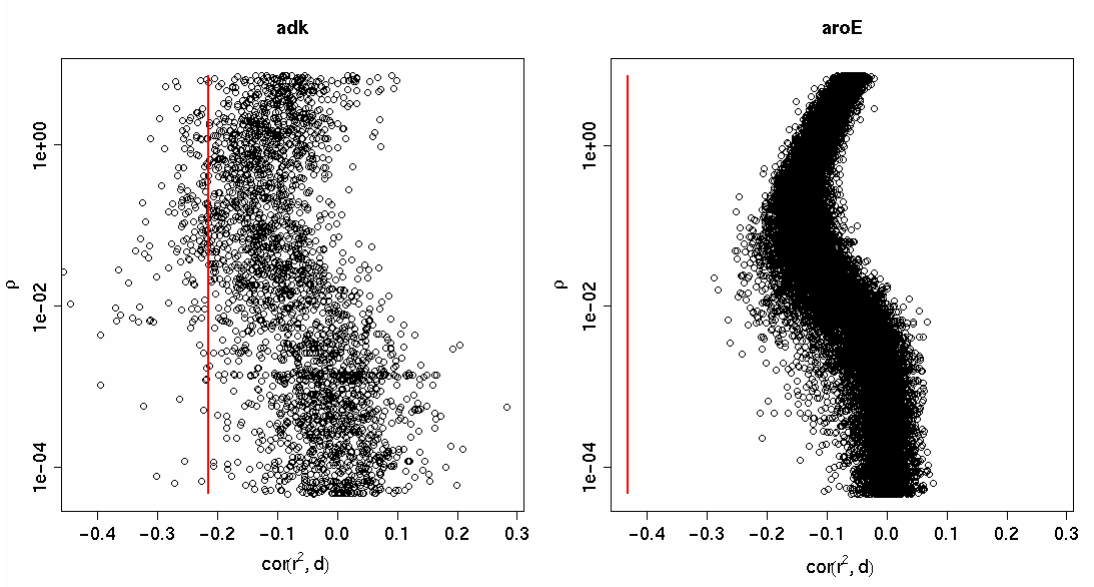

Figure 9 Cross-validation reveals discrepancies between the observed $S_3 = \mathrm{cor}(r^2,d)$ and that predicted by a model fit using S_1 and S_2. For loci *adk* and *aroE*, $f(s_3(\mathbf{X}),\rho|S_1,S_2)$ is plotted, with the observed value of S_3 indicated by the red line. For *adk*, $p^* = 0.120$,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

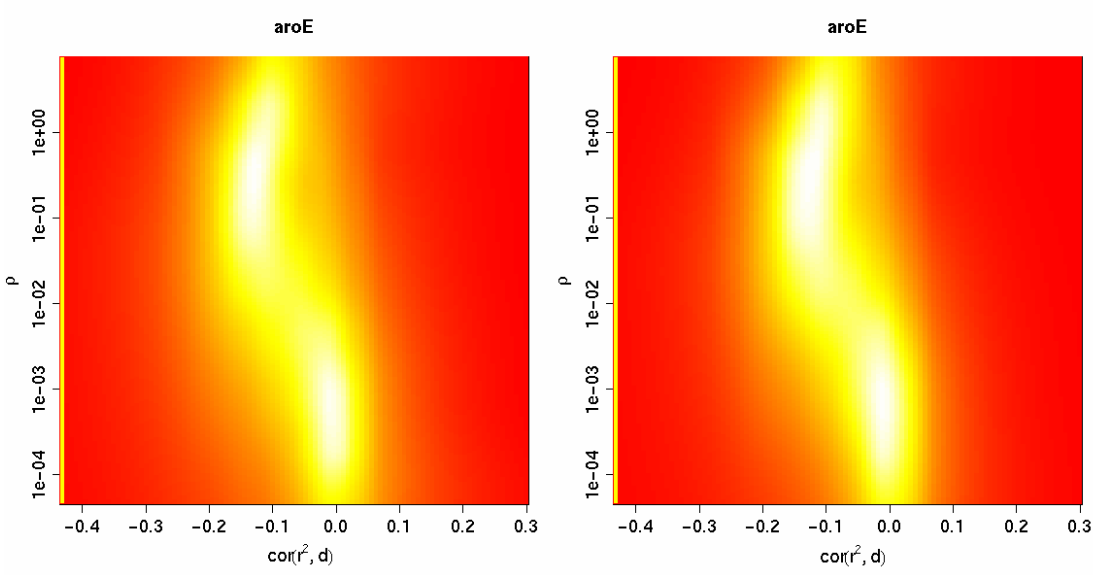

Figure 10 Locfit has been used to estimate $f(s_3(\mathbf{X}),\rho|S_1,S_2)$ and $f(s_3(\mathbf{X}),\rho|\ s_1(\mathbf{X})=S_1,\ s_2(\mathbf{X})=S_2)$ (left and right images respectively) for *aroE*. Note that ρ is on a log scale. More intense colours (closer to white) indicate higher posterior density.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