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

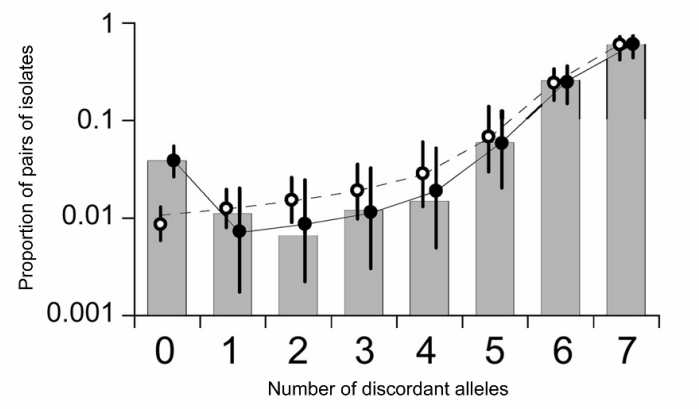

Figure 1 Allelic mismatch distribution for Czech carriage study (grey bars). The horizontal axis shows the number of loci at which a pair of isolates can differ (up to 7 for MLST), and the vertical axis the proportion of pairs that differ at that number of loci. Open circles show the fit under the standard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

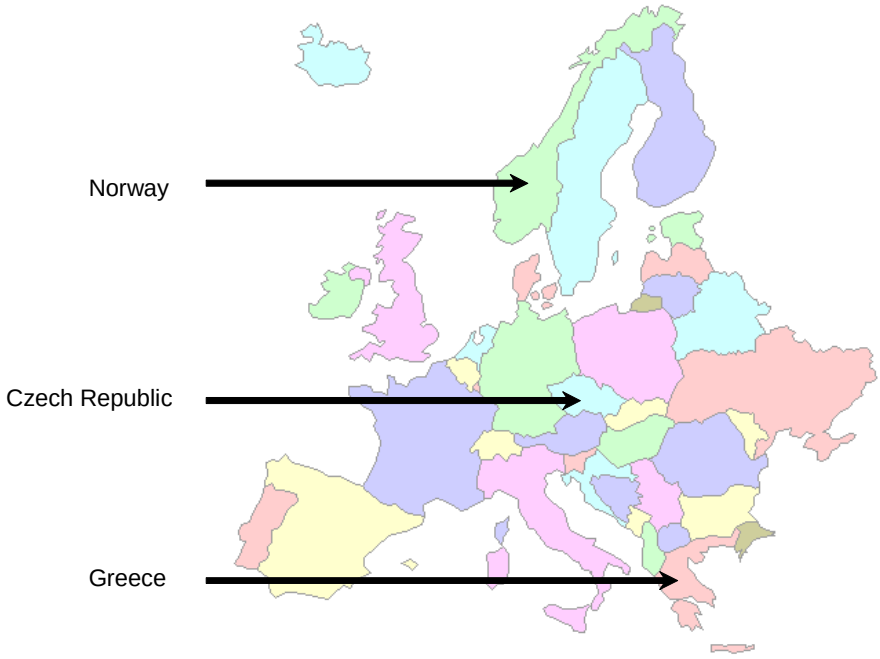

Figure 3 Map of Europe showing the location of the three countries sampled by Yazdankhah *et al.* (2004): Norway, Czech Republic and Greece. Distances between countries were measured from capital to capital (http://www.wcrl.ars.usda.gov/cec/java/lat-long.htm).

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

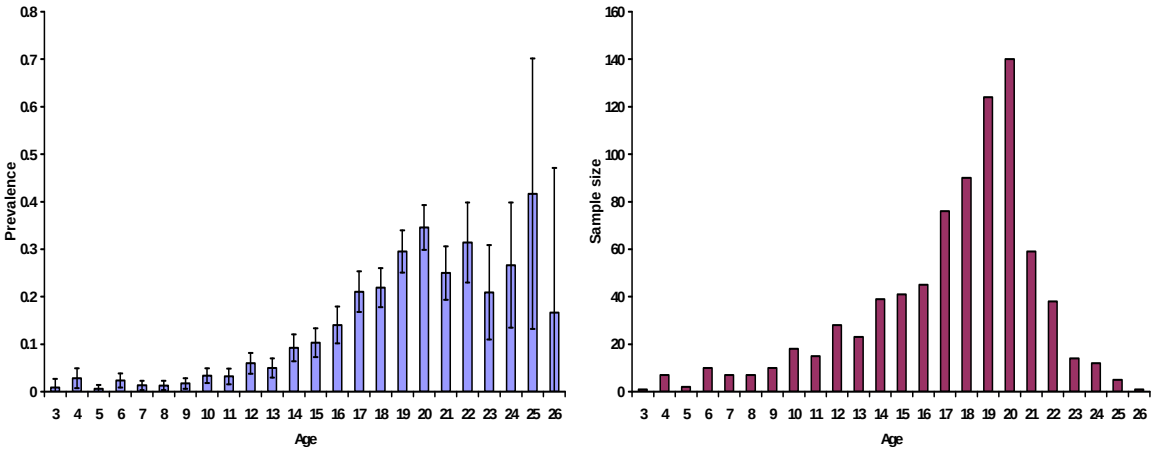

Figure 4 Left: Estimated prevalence for each age group in the Bavarian carriage study (Claus *et al.* 2005). The error bars show the approximate 95% confidence interval. Right: Sample size for each age group in the Bavarian carriage study.

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

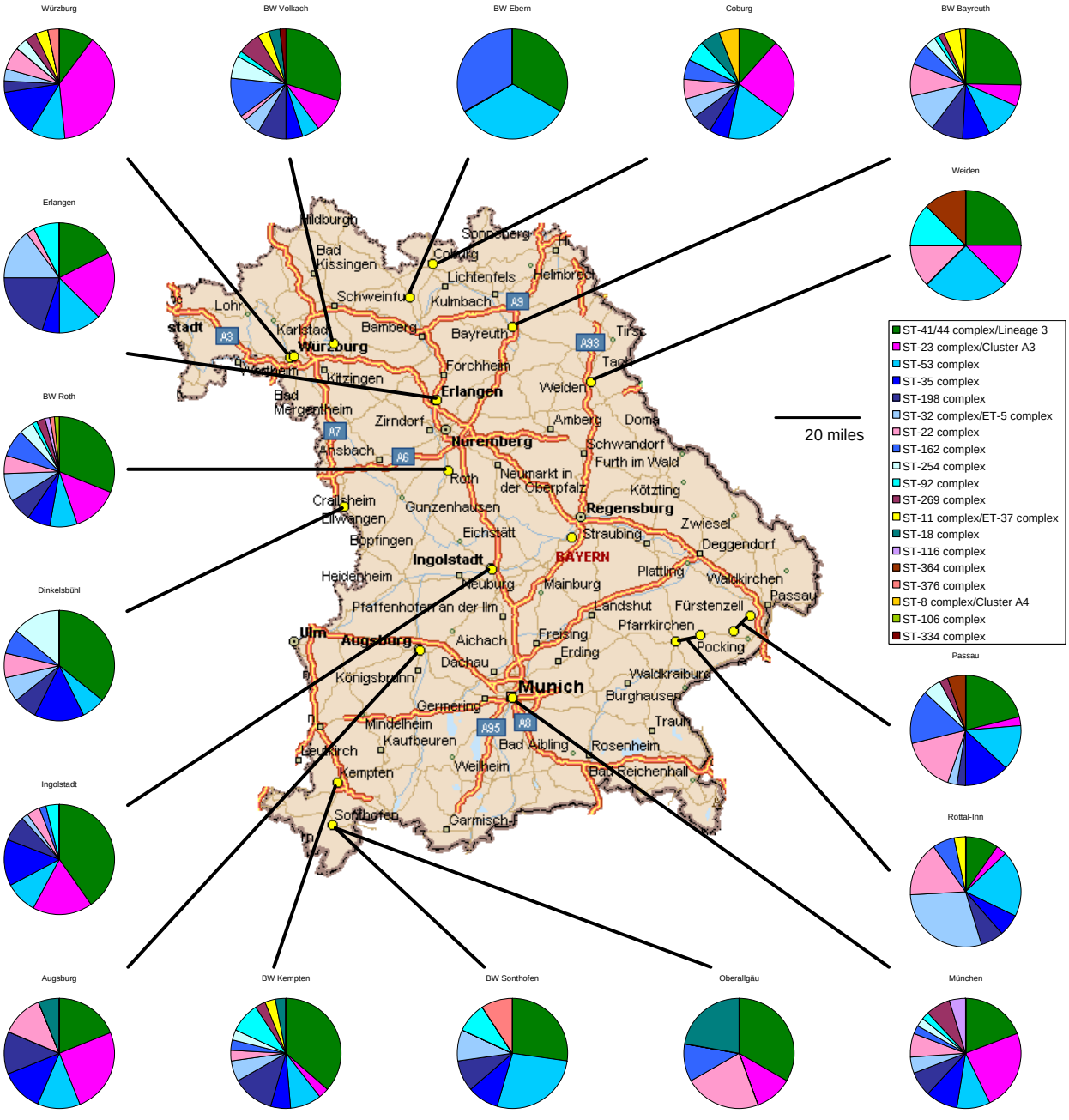

Figure 5 Genetic composition of meningococcal samples according to sampling location in Bavaria (Claus *et al.* 2005). Clonal complexes are colour-coded (see key). The principal clonal complexes are ST-41/44 complex (dark green), ST-23 complex (magenta) and ST-53 complex (sky blue). ST-11 complex, which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

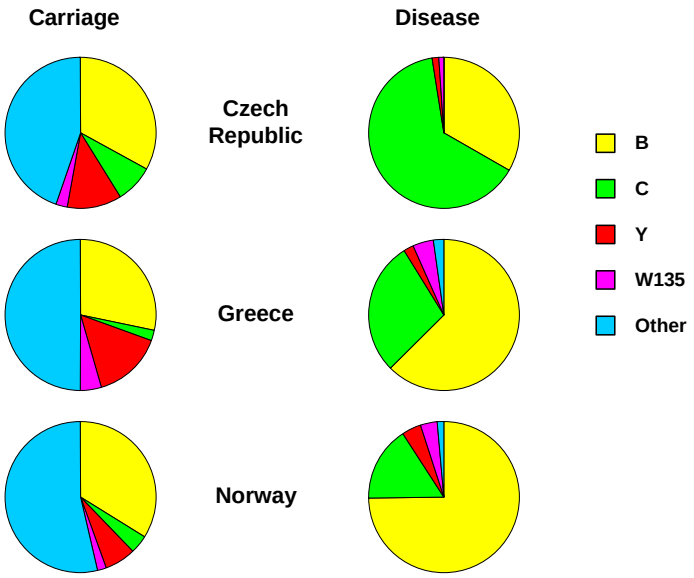

Figure 6 Distribution of serogroups in the Czech, Greek and Norwegian carriage and disease-causing isolate collections (Yazdankhah *et al.* 2004).

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.