CHAPTER SEVEN

Genetic Effects of Spatial Structure

CHAPTER SUMMARY

This chapter deals with the complexities of spatially structured populations. Such structure is the reality for many species of evolutionary interest. It arises when a species is divided into separate sub-populations (demes) with limited amounts of migration, or when individuals are distributed continuously over the species range, but disperse only over short distances.

We describe methods for measuring the genetic consequences of spatial subdivision by partitioning allele frequencies or diversity measures among different hierarchical levels, e.g., into differences between demes versus differences among individuals within them. These generate measures of divergence among populations relative to total variability, such as F_{ST}.

We explain how coalescent theory can predict the properties of neutral variation in structured populations. For the classic island and stepping-stone models, the mean coalescence time for a pair of alleles sampled from the same deme (and hence the expected within-deme diversity value under the infinite sites model) is the same as in a panmictic population, with an effective population size given by the sum of the N_e values for each deme. This property also applies to some more general migration models, including continuous populations (specifically, when migration is conservative, an important concept that we define). For alleles from different demes, between-deme diversity values are always higher than those within demes, but decrease with the migration rate. We describe situations under which a high proportion of total diversity is between demes (leading to high F_{ST} values), including the effects of deme extinction and re-colonization.

When there is a large number of demes, some general properties emerge. In particular, the place where an allele is sampled may be largely unrelated to where its ancestors were located. This can be understood in terms of the "scattering" and "collecting" phases of the coalescence of alleles sampled from a set of demes. Details of the migration process can then have surprisingly little importance in determining divergence patterns, explaining why effects of isolation by distance are often weak. With large deme numbers, population structure has little effect on the expected values of many of the properties of samples

of sequences, provided that each allele is sampled from a different deme, even if variants are subject to selection as well as drift.

In real organisms, population ranges may change, involving bottlenecks in population sizes, and species may increase or decrease in size over historical time. It is essential to take these possibilities into account when testing for selection, so we briefly discuss their effects.

When selection pressures are uniform in space, the details of migration patterns do not strongly affect the fixation probabilities of new, weakly selected semidominant mutations, if deme numbers are large or migration is conservative. The fixation probability of recessive or partially recessive deleterious mutations is, however, reduced by population structure, and the fixation of recessive favorable mutations is promoted. With weak selection and genetic drift, individual recessive and weakly deleterious mutations tend to be locally distributed, contributing to between-population heterosis.

Spatially uniform balancing selection opposes the effects of drift within demes, reducing F_{ST} relative to the neutral value, whereas local selective differences may increase F_{ST}. Comparisons between populations can thus potentially detect outlier loci that may be affected by selection. Comparisons between F_{ST} estimates for quantitative traits and neutral or nearly neutral molecular markers are also useful for detecting selection on phenotypes in structured populations.

Finally, differences in mean fitness among demes may allow inter-deme selection, which is sometimes proposed to be a major factor in adaptive evolution. However, it requires restricted conditions to be effective in opposing selection on individuals within populations.

> The breeding structure of natural populations thus is likely to be intermediate between the model of subdivision into partially isolated territories and that of local inbreeding in a continuous population.
>
> Sewall Wright (1940a)

INTRODUCTION

Apart from describing the interaction between migration and local differences in selection pressures in **Section 4.1**, we have so far assumed that populations are panmictic, i.e., there is no spatial subdivision, so that all individuals contribute equally to the entire population in the next generation, apart from any fitness differences or accidents of genetic drift. In reality, species are usually spread out over a wide area, or along a linear habitat, so that individuals from the same locality are more likely to contribute offspring to that locality than individuals from elsewhere. The extent to which individuals move from one place to another between birth and reproduction must affect the extent to which genetic differences will be maintained between different local populations.

This may have important consequences for the amounts and patterns of both neutral and selected genetic variants in the species, producing genetic differences between populations and affecting the total level of variability present in the species as a whole. Importantly, tests for selection based on patterns of DNA sequence polymorphism, of the type described in **Section 6.4.iii**, may be biased by departures from the predictions for a panmictic population, caused by spatial structure. In addition, local populations may differ genetically as a result of genetic drift. If variants are affected by selection as well as drift, the populations' mean fitnesses may differ, potentially allowing selection to operate at the level of populations (rather than on individuals, as we have so far assumed); it is important to examine this possibility, because of its implications for the mechanism of adaptive evolution.

In order to understand observed patterns of genetic variability in spatially subdivided populations, to see how methods for detecting and estimating selection are affected by subdivision, and to evaluate the evolutionary consequences of interactions between selection, migration, and drift, we need to extend our theoretical models and relate them to the data. This is the purpose of this chapter.

7.1. DESCRIBING POPULATION DIFFERENTIATION

7.1.i. Use of genetic markers to describe differences among populations

Here we provide an overview of the methods for quantifying genetic differentiation among populations of the same species. More detailed accounts of the statistical properties of the various estimators of the parameters described here can be found in Weir and Hill (2002) and Excoffier (2003), and reviews of their theoretical bases are given by Nagylaki (1998a) and Rousset (2003). Software packages such as Genepop (http://genepop.curtin.edu.au/) and DnaSP (http://www.ub.es/dnasp/) are available to implement these methods.

7.1.i.a. *Variances in allele frequencies and F_{ST}*

Differences between local populations of the same species can be described in many different ways, depending on the types of genetic variants being used. For variants such as allozymes (**Section 1.2.i**), a natural measure of the differences between a set of populations at a given locus is the variance in allele frequencies among them, at least when there are just two alleles per locus. Populations are often organized into hierarchical sets with increasing levels of subdivision, such as subspecies within a species, geographic races within subspecies, and local populations within races (**Figure 7.1**). Variances between groups at any level of such a hierarchy can then be estimated.

The use of variances suffers from the problem that their values depend on the mean allele frequencies at the loci. As discussed in **Section 5.1.i.c**, this problem can be overcome for a biallelic locus by dividing the variance in allele frequency by its maximum possible value, the product $\overline{p}(1 - \overline{p})$, where $\overline{p}$ is the mean allele frequency over the set of populations. When used to measure differentiation among a set of populations, this quantity is usually denoted by the symbol F_{ST}. Its mean over a set of loci is a good measure of differentiation between populations.

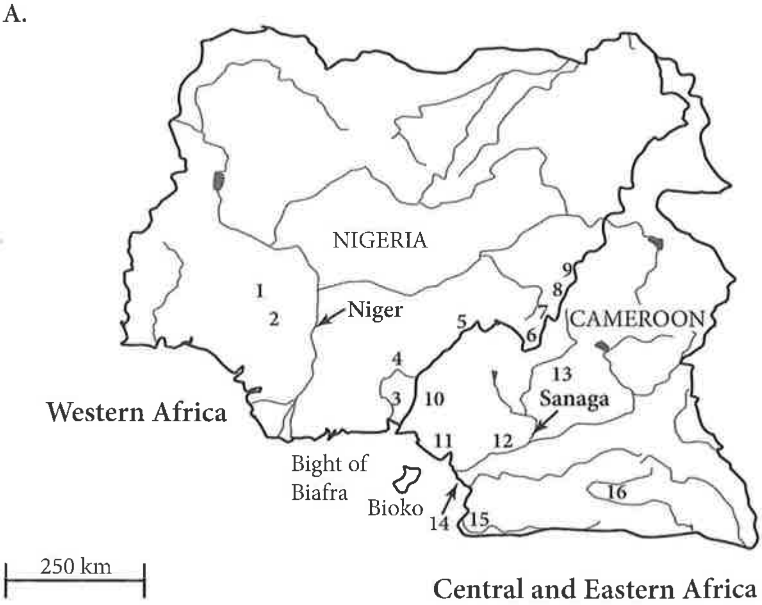

FIGURE 7.1 An example of a species (chimpanzees) whose populations are organized into hierarchical sets with increasing levels of subdivision. **A.** Locations from which individuals were sampled. **B.** Two possible interpretations (**i** and **ii**) of chimpanzee subspecies. **C.** Phylogenetic tree using a part of the mitochondrial genome, showing two major lineages (corresponding to interpretation **i**), and possible sub-lineages, suggesting interpretation **ii**. The two major lineages are estimated to have separated more than 500,000 years ago, and their present geographic ranges are in central and eastern Africa, and western Africa, respectively. Evidence for migration between these regions, across the Sanaga River, is limited. The numerical values on the nodes indicate the frequency with which these are supported by bootstrap resampling of the data (see Felsenstein 2004, Chapter 20). [Adapted from Figures 2, 4, and 6 of Gonder et al. (2006).]

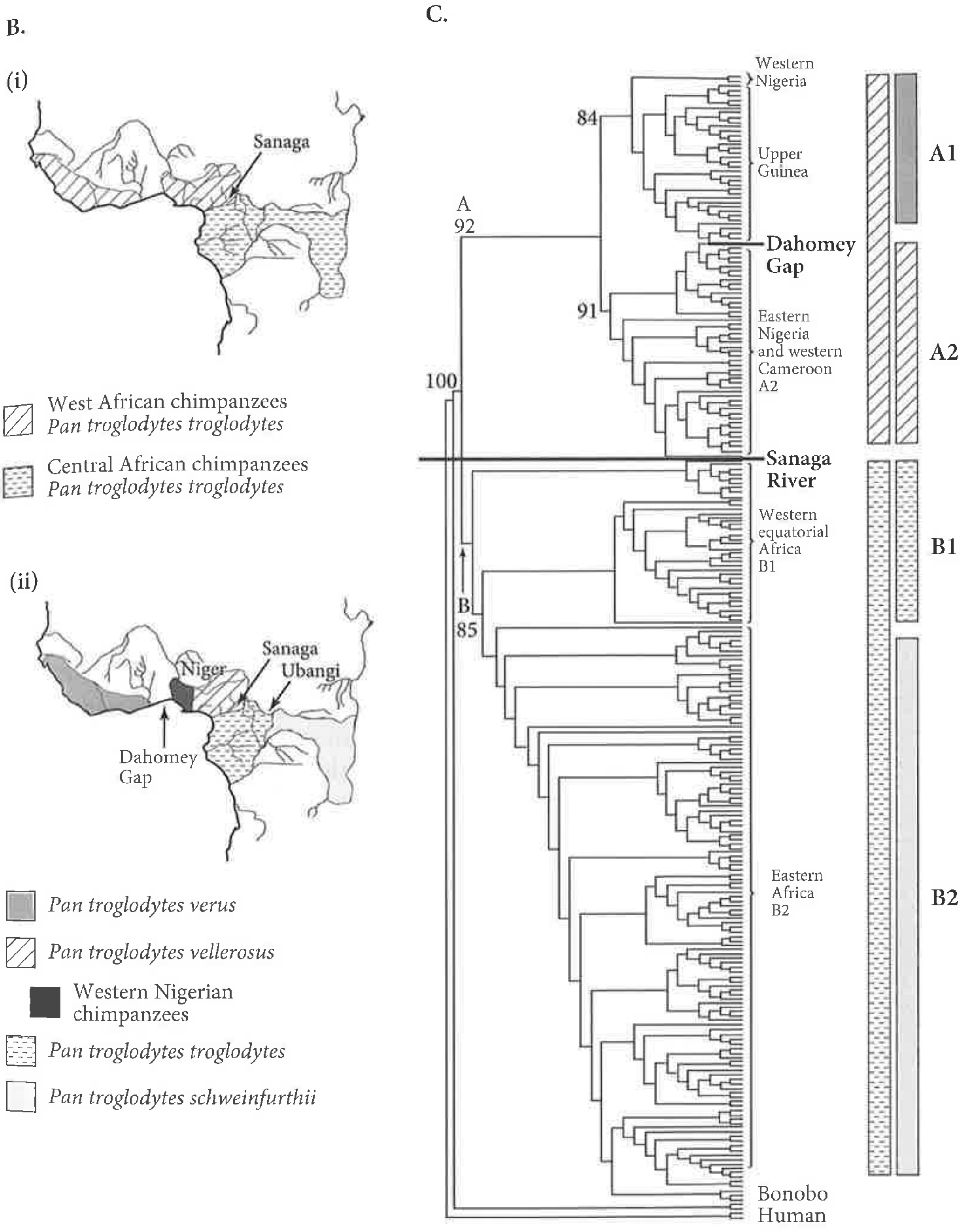

FIGURE 7.1 (Continued)

F_{ST} statistics can be estimated for all levels of a hierarchy of samples from successively finer geographic regions (Wright 1951), using a *nested* or *hierarchical* analysis of variance to estimate components of variance in allele frequency at a locus at successive levels of the hierarchy (Excoffier 2003). The lowest level of the hierarchy represents individuals within local populations or demes (**Section 4.1.i**). The departure of genotype frequencies from Hardy-Weinberg expectation at a biallelic locus within a deme can also be quantified by a measure that is similar to the inbreeding coefficient (**Section 1.3.ii**). This is denoted here by F_{IS}, where the subscripts indicate that this F statistic measures the departure of individual genotype frequencies within a local population from random mating expectation. As in **Section 5.1.i.b**, F is used here instead of f for the statistics describing all levels of a hierarchy, in order to distinguish these measures from the inbreeding coefficient based on identity by descent caused by matings between related individuals within a single population; F_{IS} is often referred to as the "fixation index" (Wright 1951). Unlike the inbreeding coefficient, F_{IS} can be negative if a population has an excess of heterozygotes (caused, for example, by selection or assortative mating).

7.1.i.b. *Gene diversities for allelic variants*

The method just described does not allow for the possibility of multiple alleles at the loci of interest, unless we adopt the expedient of simply considering one variant versus all alternatives, thereby throwing away information (which is undesirable). Microsatellite loci, for example, usually have large numbers of alleles (**Section 1.2.iii.c**), and are often used to study population differentiation. A commonly used method that avoids this problem is to use the gene diversities introduced in **Section 1.2.i.b**. Pooling over all samples from a species (i.e., using the frequencies of all alleles in the entire sample), we can estimate the species-wide diversity, H_T, for a set of loci, using the standard formula (**Box 1.2**). For the next level of the hierarchy, such as subspecies, we can calculate diversity for each subspecies and determine the mean over subspecies. Denoting this by H_S, it seems reasonable to define an analog of F_{ST} by:

$$G_{ST} = \frac{(H_T - H_S)}{H_T} \qquad (7.1)$$

G_{ST} describes the difference between the level of diversity in the species as a whole and the mean within-subspecies diversity, relative to the overall diversity. It is thus a natural measure of the extent to which subspecies are genetically differentiated (Nei 1973). This can be extended to successive levels of a hierarchy by computing the values of the mean diversities within successive levels and obtaining the corresponding G_{ST} statistics (Nei 1973; Excoffier 2003). For loci with two alleles, it can be shown that the theoretical values of F_{ST} and G_{ST} are equivalent (**Problem 7.1**).

For microsatellite loci, some information is lost by using allelic diversity, because this ignores the information about differences in alleles' repeat numbers, which one should take into account, because they tend to increase more linearly than diversity with genealogical divergence; the variance of repeat lengths among alleles is therefore often used to measure variability (**Section 5.1.iii.f**). This can be extended to differentiation between populations by estimating the relevant mean squared differences in repeat lengths among pairs of alleles, sampled from the same and different populations. This provides an analog of F_{ST} and G_{ST}, denoted by R_{ST} (Goldstein et al. 1995; Slatkin 1995b).

7.1.i.c. *Nucleotide site diversity*

If DNA sequence information is used, we could use the gene diversity method just described to characterize population differentiation, replacing allele frequencies by haplotype frequencies (**Section 1.2.v.a**). This is often done, especially in the extensive literature on mitochondrial gene variation (Excoffier et al. 1992). However, this approach loses information about the extent of sequence differences, and it is better to treat each nucleotide site as an observational unit and to define an analog of G_{ST}, K_{ST}, at this level (Hudson et al. 1992). The values of π_T and π_S are estimated for the set of populations in question, where the nucleotide site diversity π for a sequence or set of sequences (**Section 1.2.iii.c**) replaces gene diversity in **Equation 7.1**. This gives:

$$K_{ST} = \frac{(\pi_T - \pi_S)}{\pi_T} \tag{7.2}$$

The statistical significance of K_{ST} can be assessed by resampling. To do this, variants at individual sites are reassigned randomly to populations, so as to compare the observed K_{ST} value with the distribution of K_{ST} over many randomizations that assume no subdivision of the species (Hudson et al. 1992). A similar procedure can be performed with an alternative measure, S_{nn}, which measures the extent to which related sequences are found in the same population (Hudson 2000).

7.1.i.d. *Other methods*

In addition to these widely used methods, many other procedures have been devised for quantifying genetic differentiation between populations (e.g., Holsinger 1999; Song et al. 2006). One approach is to assign individuals to distinct groups, based on their multi-locus genotypes, and to use Bayesian statistical inference to find the best assignment (Pritchard et al. 2000; Dawson and Belkhir 2001; Falush et al. 2003; Corander et al. 2004; Guillot et al. 2005); see http://pritch.bsd.uchicago.edu/structure.html, http://www.genetix.univ-montp2.fr/partition/partition.html, and http://www.abo.fi/fak/mnf/mate/jc/smack_software_eng.html. This does not require prior knowledge of the individuals' spatial origins, but estimates the number of distinct groups into which individuals appear to fall. Unlike the methods described above, however, the results cannot easily be related to the underlying population genetic processes.

7.1.ii. Empirical patterns

A very large amount of data has been accumulated on genetic variation among populations of animals and plants. The results range from low levels of differentiation in mobile organisms, such as the fruitfly *Drosophila pseudoobscura* (Schaeffer and Miller 1992) or humans (Lewontin 1972a; Akey et al. 2002), which have mean G_{ST} or K_{ST} values of around 0.10, to levels approaching 1 in highly self-fertilizing species of plants, where gene flow between populations due to pollen dispersal is likely to be very limited, because ovules receive little pollen from other individuals (Hamrick and Godt 1996; Charlesworth 2003). **Figure 7.2** shows this effect in a plant species.

It is often assumed that statistics like K_{ST} and G_{ST} reflect the amount of migration between populations, but, as we will see in **Section 7.2** below, they are also strongly affected by the effective sizes of local populations, reflected in the level of variability within populations. For example, **Equation 7.2** shows that, if $\pi_S = 0$, $K_{ST} = 1$. High levels of the divergence statistics are indeed often associated with low diversity within local populations, which suggests low effective population sizes (Jarne and Staedler 1995; Hamrick and Godt 1996; Charlesworth 2003), as can be seen in **Figures 7.2** and **7.3**. **Figure 7.3** shows a comparison between an outcrossing and a highly selfing species of the nematode worm *Caenorhabditis*. Here, migration depends on movements of individuals, not gametes, and so self-fertilization should not directly affect migration. In this case, the much higher K_{ST} value in the inbreeding species is probably due to the reduced effective population size caused by selfing (see **Sections 5.2.iii.a** and **8.3.i**).

7.2. THE THEORY OF NEUTRAL VARIATION WITH SPATIAL STRUCTURE

Now that the relevant measures have been defined and illustrated with examples from nature, we next explain how they are expected to behave. Much of this theory deals with F_{ST} and related statistics, but in order to understand their properties it is necessary to examine the behavior of diversity measures at the different hierarchical levels. We begin with neutral variation, which must be understood before dealing with variants that are under selection, which we consider in **Section 7.3**.