Chapter 4

Genetic Analysis of Multiple Populations

Why Study Multiple Populations?

In Chapter 3 we learned that by quantifying the genetic diversity of single populations, we can gain considerable insight into processes as diverse as bottlenecks, reproduction and natural selection. We must understand these process before we can interpret genetic data in an ecological context, but the genetics of populations are influenced by both intra-population and inter-population processes, and therefore the next step is for us to investigate how gene flow (the transfer of genes from one population to another) influences the evolution of populations and species. We will start this chapter by looking at how we can quantify population subdivision and gene flow. However, we cannot predict the effects of gene flow without adding genetic drift and natural selection into the equation, because it is these three processes that jointly determine the extent to which populations will genetically diverge or converge. In the second part of this chapter we shall therefore revisit genetic drift and natural selection, but this time the emphasis will be on the ways in which they interact with gene flow to influence the genetic structure of populations across a species' range.

Quantifying Population Subdivision

With the exception of rare or endangered species, which have limited distributions, virtually every species investigated to date has revealed some level of genetic differentiation between populations. A lack of population differentiation would mean that all populations had the same allele frequencies. This would be possibly only if the entire species constituted a single group of randomly mating individuals, which is highly unlikely to occur in species that comprise multiple populations.

Genetic Distance

One way in which we can measure the genetic similarly of two populations is by estimating the genetic distance between them. There are many different ways in which this can be done, one of the most common being Nei's (1972) standard genetic distance, *D*. To calculate this, we first need to know Nei's measure of genetic identity (*I*), which reflects the genetic similarity of populations. For a given locus, this is calculated as:

$$I = \frac{\sum_{i=1}^{m} (p_{ix} p_{iy})}{\left[\left(\sum_{i=1}^{m} p_{ix}^{2} \right) \left(\sum_{i=1}^{m} p_{iy}^{2} \right) \right]^{0.5}} \quad \textbf{(4.1)}$$

in which p_{ix} is the frequency of allele i in population x, p_{iy} is the frequency of allele i in population y, and m is the number of alleles at the locus (see Box 4.1).

Box 4.1 Calculating Nei's genetic distance

The data in Table 4.1 represent the allele frequencies of a single allozyme locus (phosphoglucomutase, or Pgm) in two populations of the freshwater zooplankton *Daphnia pulex* in Illinois and Indiana, USA.

Table 4.1 Allele frequencies at the *Pgm* locus in two *D. pulex* populations (data from Crease et al., 1990)

Pgm allele	Illinois population	Indiana population
F	0.146	0.491
M	0.818	0.106

Pgm allele	Illinois population	Indiana population
S	0.036	0.403

We will denote Illinois as population *x* and Indiana as population *y*. Since $I = \sum_{i=1}^{m}(p_{ix}p_{iy})/[(\sum_{i=1}^{m}p_{ix}^2)(\sum_{i=1}^{m}p_{iy}^2)]^{0.5}$, we first need to calculate the sum of the products of the allele frequencies as follows:

$$\sum_{i=1}^{m}(p_{ix}p_{iy}) = (0.146)(0.491) + (0.818)(0.106) + (0.036)(0.403) = 0.173$$

We then need to calculate the sums of the squared allele frequencies, which for the Illinois population is:

$$\left(\sum_{i=1}^{m}p_{ix}^2\right) = 0.146^2 + 0.818^2 + 0.036^2 = 0.692$$

and for the Indiana population is:

$$\left(\sum_{i=1}^{m}p_{iy}^2\right) = 0.491^2 + 0.106^2 + 0.403^2 = 0.415$$

Therefore:

$$I = 0.173/[(0.692)(0.415)]^{0.5} = 0.173/\sqrt{[(0.692)(0.415)]}$$
$$= 0.173/0.536 = 0.323$$

Because $D = -\ln I$, in this case the genetic distance between the two populations is:

$$D = -\ln(0.323) = 1.13$$

Values of *I* range from zero to one. Once calculated, they can be used to obtain Nei's *D* as follows:

$$D = -\ln I \quad \textbf{(4.2)}$$

Values of *D* range from zero to infinity. If two populations have similar allele frequencies, in other words if $p_{ix} \approx p_{iy}$, genetic similarity (*I*) between the two will approach 1, and genetic distance (*D*) will approach zero. At the other extreme, if two populations have no alleles in common, *I* will be zero and *D* will be infinity.

F-statistics

Perhaps the most common method for quantifying the genetic differentiation between populations is based on ***F*-statistics**, which were developed by Wright (1951). *F*-statistics use inbreeding coefficients to describe the partitioning of genetic variation within and among populations. *F*-statistics can be calculated at three different levels. Note that in describing these levels we will follow the convention set by the literature underlying the theory of *F*-statistics, and refer to spatially discrete samples as subpopulations instead of populations; elsewhere we will revert to the more common practice of referring to discrete breeding units as populations. The first *F*-statistic, $\boldsymbol{F_{IS}}$, measures the degree of inbreeding within individuals relative to the rest of their subpopulation. This reflects the probability that two alleles within the same individual are identical by descent, and is the same as the inbreeding coefficient *F* that was introduced in Chapter 3. It is calculated as:

$$F_{IS} = \frac{H_S - H_I}{H_S} \quad \textbf{(4.3)}$$

in which H_I is the observed heterozygosity in a subpopulation at the time of investigation (*individual* heterozygosity), and H_S is the heterozygosity that would be expected if the subpopulation was in HWE (*subpopulation* heterozygosity).

The second *F*-statistic is $\boldsymbol{F_{ST}}$ (also known as the fixation index), and this provides an estimate of the genetic differentiation between subpopulations; in other words it tells you how genetically similar two populations are to one another. It is a measure of the degree of inbreeding within a subpopulation relative to the total population (total population here meaning all of the subpopulations combined), and reflects the probability that two alleles drawn at random from within a subpopulation are identical by descent. F_{ST} is calculated as:

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad \textbf{(4.4)}$$

in which H_S is the same as in equation (4.3) and H_T is the expected heterozygosity of the *total* population (i.e. the relevant subpopulations, combined). The third F-statistic, which is used much less frequently than the other two, is $\boldsymbol{F_{IT}}$. This provides an overall inbreeding coefficient for an individual by measuring the heterozygosity of an individual relative to the total population. F_{IT} is therefore influenced by both non-random mating within a subpopulation (F_{IS}), and population subdivision (F_{ST}), and is calculated as:

$$F_{IT} = \frac{H_T - H_I}{H_T} \quad \textbf{(4.5)}$$

in which H_T and H_I are the same as in equations (4.3) and (4.4). The relationship between the three statistics is:

$$F_{IT} = F_{IS} + F_{ST} - (F_{IS})(F_{ST}) \quad \textbf{(4.6)}$$

Because F_{ST} measures the extent to which populations have differentiated from one another, it is the F-statistic with which we are most concerned in this chapter.

Box 4.2 Calculating *F* statistics

We will now use some hypothetical data from two scarlet tiger moth subpopulations to calculate F-statistics. Recall from Chapter 3 that moths with white spotting were assigned the genotype AA, moths with little spotting were aa, and moths with intermediate spotting were the heterozygotes, Aa. The frequencies of the alternative genotypes from two subpopulations are given in Table 4.2, and from these data we first need to calculate H_I, H_S and H_T.

Table 4.2 Frequencies of alternative wing genotypes in two hypothetical subpopulations of the scarlet tiger moth

	No. of individuals			Genotype frequency		
	AA	*Aa*	*aa*	*AA*	*Aa*	*aa*
Subpopulation 1	352	63	12	0.824	0.148	0.028
Subpopulation 2	312	77	27	0.750	0.185	0.065

First we calculate H_I, which is the average observed heterozygosity across populations:

$$H_I = \frac{0.148 + 0.185}{2} = 0.167$$

Next we calculate the subpopulation heterozygosity, H_S, which is the heterozygosity that would be expected if the subpopulations were in HWE. We know that under HWE, heterozygosity is equal to $2pq$, therefore we first need to calculate p and q. In subpopulation 1 there are 427 moths and therefore 854 alleles, so:

$$p = \frac{(2)(352) + 63}{854} = 0.898$$

$$q = \frac{(2)(12) + 63}{854} = 0.102.$$

In subpopulation 2 there are 416 moths and therefore 832 alleles, so:

$$p = \frac{(2)(312) + 77}{832} = 0.843$$

$$q = \frac{(2)(27) + 77}{832} = 0.157$$

The expected heterozygosity is therefore:

$$2pq = 2(0.898)(0.102) = 0.183 \text{ in subpopulation 1, and}$$

$$2pq = 2(0.843)(0.157) = 0.265 \text{ in subpopulation 2.}$$

This means that

$$H_S = \frac{0.183 + 0.265}{2} = 0.224$$

Copyrighted material

Next, we calculate H_T by using the average allele frequencies from the two subpopulations to calculate the expected heterozygosity if the total population was in HWE. In this case:

$$p = \frac{0.898 + 0.843}{2} = 0.871$$

$$q = \frac{0.102 + 0.157}{2} = 0.130$$

Therefore

$$H_T = 2(0.871)(0.130) = 0.226$$

We can now calculate inbreeding within subpopulation 1 (in the same way that F was calculated in Chapter 3) as:

$$F_{IS} = (H_S - H_I)/H_S = (0.183 - 0.148)/0.183 = 0.191$$

and the inbreeding due to population differentiation as:

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{0.226 - 0.224}{0.226} = 0.009$$

and inbreeding due to both inbreeding within, and differentiation among, populations as:

$$F_{IT} = \frac{H_T - H_I}{H_T} = \frac{0.226 - 0.167}{0.226} = 0.261$$

Interpreting F_{ST}

If two populations have identical allele frequencies they will not be genetically differentiated and therefore F_{ST} will be zero. At the other extreme, if two populations are fixed for different alleles then F_{ST} will be equal to one (see Table 4.3 for some examples of F_{ST} values). Within that range, F_{ST} values of 0–0.05 are generally considered to indicate little genetic differentiation; values of 0.05–0.25 indicate moderate genetic differentiation; and values over 0.25 represent pronounced levels of genetic differentiation. However, these are only approximate guidelines, and in reality even very low levels of F_{ST} may represent important levels of genetic differentiation. The significance of F_{ST} estimates is based on a permutation procedure which shuffles genotypes among populations thousands of times, and an F_{ST} value is calculated for each permutation. The P-value of the test is based on the number of times that these F_{ST} values are equal to or larger than the one that is calculated from the actual data set.

Table 4.3 Some examples of F_{ST} values that have been calculated for various taxa. Collectively, these examples show that geographical distance is a poor predictor of population differentiation

Species	Distance (km)	F_{ST} (or analogue)[1]	Molecular marker	Reference
Pitcher's thistle (*Cirsium pitcheri*)	2–435	0.112–0.594	Microsatellites	Gauthier *et al.* (2010)
Sweet vernal grass (*Anthoxanthum odoratum*)	0.2–0.4	0.015–0.18	AFLPs	Freeland *et al.* (2010)
Freshwater mussels (*Unio pictorum*)	0.8–100	0.0017–0.0273	AFLPs	Zieretz *et al.* (2010)
Atlantic salmon (*Salmo salar*)	5–164	0.001–0.072	Microsatellites	Primmer *et al.* (2006)
Tuco-tuco (a rodent; *Ctenomys rionegrensis*)	24–44	0.185–0.984	Microsatellites	Kittlein and Gaggiotti (2008)
Parasitoid wasp (*Nasonia vitripennis*)	2–5000	−0.025–0.185[2]	Microsatellites	Grillenberger *et al.* (2009)
Pygmy rabbit (*Brachylagus idahoensis*)	0.5–32	0.005–0.054	Microsatellites	Estes-Zumpf *et al.* (2010)

[1]See Box 4.3 for analogues of F_{ST}.
[2]Negative F_{ST} values should be interpreted as no genetic differentiation between two populations, and reflect imprecision of the algorithm used to estimate the value.

In many species the amount of gene flow between populations is inversely proportional to the geographic distances between them because individuals are most likely to disperse to nearby sites. This is known as **isolation by distance** (IBD; Wright, 1943), and is most commonly assessed by regressing log-transformed estimates of gene flow between pairs of populations against the appropriate log-transformed geographic distances. The significance of this relationship can be assessed using a **Mantel test**, which tests for a correlation between genetic and geographic distances. The Mantel test is appropriate for this comparison because it does not assume that the population pairwise comparisons are independent. The slope and the intercept of the IBD regression can be used to test the strength of the relationship. A metanalysis by Jenkins *et al.* (2010) identified a number of trends with respect to patterns of IBD (Table 4.4).

Table 4.4 The likelihood of finding a pattern of isolation by distance is influenced more by some factors than by others. Data from Jenkins et al. (2010)

Factor	Significant?	Comments
Sample size	Yes	Data from >9 populations needed to achieve >50% probability of significance; data from >24 populations needed to achieve >90% probability of significance
Metabolism	Yes	Ectotherms more likely than endotherms to show pattern of IBD
Taxa	Yes	In a comparison of amphibians, reptiles, birds, fish, mammals and arthropods, some were more likely to follow IBD than others
Broad taxonomic divisions	No	Neither plants nor animals more likely to show pattern of IBD
Active versus passive dispersal	No	Active and passive dispersers equally likely to show pattern of IBD
Molecular markers	No	Likelihood of finding IBD did not depend on the type of molecular marker used
Genetic distance estimator	No	Neither F_{ST}, F_{ST} analogue, nor Nei's D more likely to show pattern of IBD
Plant habits	No	Neither herbs, shrubs, nor trees more likely to show pattern of IBD

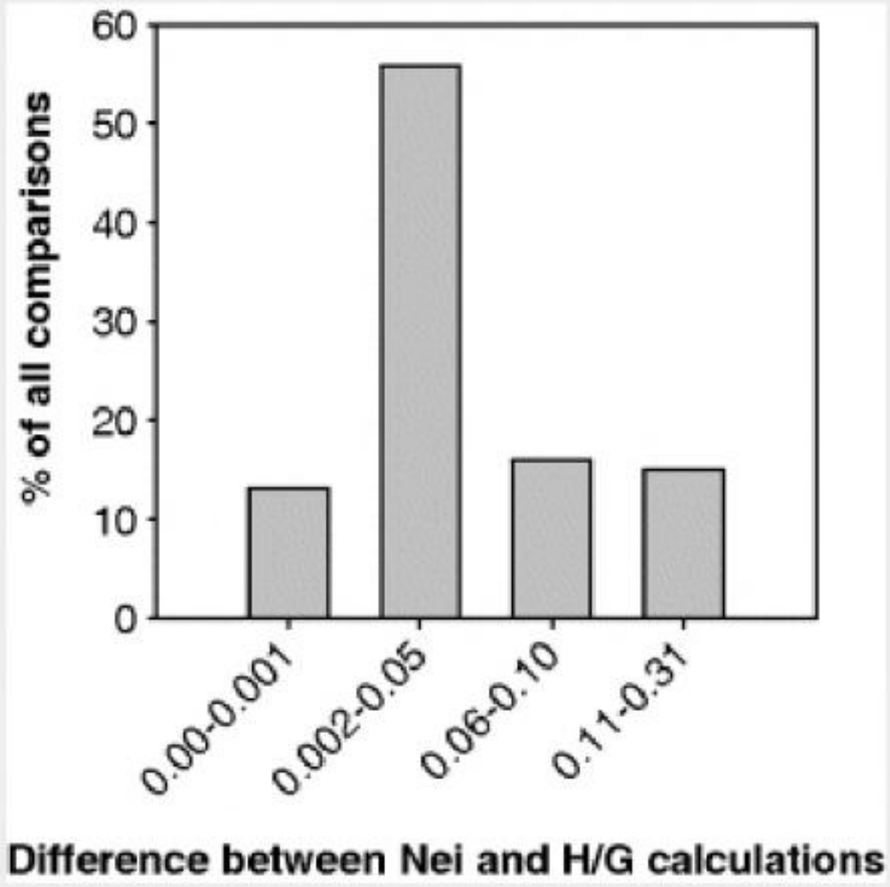
Box 4.3 Analogues of F_{ST}

F_{ST} as developed by Wright (1951) was the original method for estimating population differentiation from allele frequencies, but since then several variations have been developed that each measures population differentiation in a slightly different way. One of these is $\boldsymbol{G_{ST}}$, which was developed by Nei (1973). G_{ST} is equivalent to F_{ST} when there are only two alleles at a locus. In the case of multiple alleles, G_{ST} is equivalent to the weighted average of F_{ST} for all alleles. A similar measurement is **theta** ($\boldsymbol{\theta}$; Weir and Cockerham, 1984), which is often preferred because it takes into account the effects of uneven sample sizes and the number of sampled populations. More recently, $\boldsymbol{R_{ST}}$ was developed by Slatkin (1995) specifically for the analysis of microsatellite data. R_{ST} differs from the other measurements because it assumes a stepwise mutation model (SMM; Chapter 2) which may be more realistic for microsatellite data than the infinite alleles model (IAM). A recent paper by Jost (2008) suggested that when researchers are using highly polymorphic markers to examine differentiation among populations they should not use G_{ST} or its analogues because when diversity is high this measure will approach zero (no differentiation). This will occur even if there are no alleles shared between populations. Jost (2008) proposes a new measure D to overcome this problem (calculated in the program SMOGD http://www.ngcrawford.com/django/jost/). There is currently an interesting discussion about the best ways to measure diversity and differentiation (see Heller and Siegismund, 2009; Jost, 2009; Ryman and Leimar, 2009). One place to access these and similar types of discussions is the Nature Network population genetics forum (http://network.nature.com/groups/popgen/forum/topics).

It is beyond the scope of this text to provide a comprehensive discussion on the many different measures of population differentiation. The theoretical literature surrounding these measurements is copious, and often opaque; interested readers are referred to the recommended reading at the end of this chapter. Nevertheless, even if few of us can understand all of the mathematics that underlie estimates of genetic differentiation, it is important to understand that these estimates will often vary depending on which method is used. This may be particularly important if we are comparing the results of multiple studies, for example we may be interested in whether populations of perennial plants show greater differentiation than populations of annual plants, but such a comparison will be valid only if it is based on estimates of genetic differentiation that were all calculated in the same way. This was illustrated by a study that compared 227 measurements of G_{ST} that had been calculated by either the method of Nei (1973), or the method of Hamrick and Godt (1990). The main difference between the two methods is that in the former, G_{ST} is calculated from the mean values of H_T and H_S values after they have been averaged across all loci, whereas in the latter, G_{ST} values are calculated separately for all loci and then averaged. A comparison of the two methods revealed that although results were often comparable, G_{ST} values in 15% of all studies differed by >0.10 (Culley *et al.*, 2002; Figure 4.1).

Material com direitos autorais

Figure 4.1 Range of differences between G_{ST} values calculated using both Nei's (1973) and Hamrick and Godt's (1990) methods (H/G), based on 227 comparisons. Data from (Culley et al., 2002).



Some generalizations are emerging about the link between ecology and population differentiation. For example, a recent review of 40 species of lowland neotropical rain forest birds, each of which have populations isolated on either side of the partial barriers that are presented by the Andes, the Amazon River and the Madeira River, revealed that canopy species had statistically lower levels of genetic differentiation across the geographical barriers compared to understorey birds (Burney and Brumfield, 2002) (Figure 4.2). However, for the most part it can be extremely difficult to predict the expected level of differentiation between any pair of populations, because the range of population divergence is remarkable both within and among species.

Figure 4.2 Ruddy pigeon (Patagioenas subvinacea), a neotropical canopy bird. A review of lowland neotropical rainforest species identified canopy species as having significantly lower levels of genetic differentiation across partial geographical barriers compared to understorey species (Burney and Brumfield, 2002). Photo attributed to Michael Woodruff.

Non-*a Priori* Identification of Populations

Material com direitos autorais

F_{ST}, and similar measures, assume that the groups being compared are populations, but this may not always be the case; as noted at the beginning of Chapter 3, defining the boundaries of populations can be challenging. Grouping animals into populations that do not reflect the true population can have severe consequences. For example, harvest limits for game animals are often determined based on population size estimates of animals living within a certain area or Management Unit (MU; see Chapter 88). The boundaries of each MU are usually determined by jurisdictional or political processes, not by biological characteristics. Imagine a situation in which there are two populations that occupy two MUs (MU1 and MU2). One population occupies all of MU1 and half of MU2. If a researcher sampled individuals from the two MUs, then grouped them by MU (because they don't realize there is substructure), they would likely determine that the genetic differentiation between MUs was 0. This is because one half of the individuals in MU2 are, in fact, the same as MU1 animals. The harvest limit would be set based on the total population size of the two MUs, which could result in the smaller population in MU2 being over-harvested. To avoid this problem, and to reduce the amount of bias that sampling and grouping has on the determination of population structure, non-*a priori* methods have been developed to identify populations. *A priori* knowledge is known or assumed beforehand (e.g. in the example above, that MUs are synonymous with populations), whereas non-*a priori* methods do not make that assumption.

By necessity, non-*a priori* methods are individual-based. One of the most common non-*a priori* analyses to identify population structure has been the Bayesian method of Prichard *et al.* (2000) (more on Bayesian methods later in this chapter). Their program, STRUCTURE (Prichard *et al.*, 2000, Falush *et al.*, 2003, Hubisz *et al.*, 2009), implements the Bayesian method by calculating the likelihood that the data can be divided into a number of clusters (K), with each cluster trying to maximize Hardy–Weinberg equilibrium and linkage equilibrium. As these are characteristics of a true theoretical population, the most likely K should correspond to the correct number of populations that have been sampled. In fact, the running of STRUCTURE and subsequent interpretation of the results can be complex but a large literature has built up to guide researchers through the process. Although it is currently the most common, STRUCTURE is not the only non-*a priori* method available to researchers; BAPS (Corander *et al.*, 2006) and Geneland (Guillot *et al.*, 2005) also cluster individuals into groups that conform to theoretical expectations of a population using Bayesian methods.

Several non-Bayesian methods have also been developed to determine population structure from genetic data, specifically those using multivariate methods such as principle components analysis (PCA) (Eigensoft: Patterson *et al.*, 2006; adegenet: Jombart, 2008), multidimensional scaling for SNP data (AW-clust: Gao and Starmer, 2008), and frequency correspondence analysis (Genetix: Belkhir *et al.*, 1999). Other advances are programs that incorporate geographic information (Geneland: Guillot *et al.*, 2005; sPCA: Jombart *et al.*, 2008; TESS: Guillot, 2009). When using all of these methods and programs, it is critical that the experimental design phase is carefully planned because results can be as biased as using *a priori* approaches (Box 4.4). As an example, Schwartz and McKelvey (2009) simulated on a landscape individuals that had a high probability of mating with neighbours, and after sampling these landscapes in ways typical of field studies, they used STRUCTURE to determine the most likely number of clusters from the data. In general, they identified more clusters (between four and seven) than actually existed in the data (only one). In an empirical study, this might lead the researchers to overestimate the number of populations in the study area.

Box 4.4 Experimental design

In the past, experimental design has played a minor role in molecular ecology research (population genetics specifically) because the samples that are used were collected opportunistically. For example, a researcher who is interested in the patterns of genetic variation within a species may focus their sampling effort where they know the species occurs, in addition to soliciting samples from researchers working in other parts of the species' range (research sampling, c.f. Schwartz and McKelvey 2009). This usually results in a data set in which sample sizes are variable among locations, and may be skewed towards a certain sex, age class or time period. At best, this strategy can limit the analyses that can be conducted; at worse, it can bias the results. Therefore, experimental design is needed before a study is conducted.

What is the importance of experimental design? The experimental design process is the planning stage that allows the researcher to determine the most efficient way to obtain an answer to their research question. It attempts to ensure that the results of a hypothesis-testing exercise arise because of a real ecological pattern, and not by chance or some source of bias (e.g. introduced by the researcher). Entire textbooks have been devoted to experimental design and a full treatment of this subject is beyond the scope of this text. However, we can point out the key elements of a well-designed experiment, and raise some questions that you will need to ask yourself in order to devise a sound sampling scheme.

1. Establish your research question and from that create a hypothesis that can be tested. A testable hypothesis is one that can be negated (you can reject the null hypothesis) and for which there is a single alternative hypothesis. This step will involve a thorough investigation of the research that has gone on before in the system in which you are interested.

2. Determine how you will test your hypothesis. This step identifies the 'ideal' situation in which the only concern is identifying the best way to test your hypothesis. In this step you should be especially concerned with determining if the test results will allow you to reject the null hypothesis. A number of decisions can be encompassed in this phase of the design process.

a. Experimental approach. Designs can be broadly grouped into two approaches: observational and experimental. Observational designs are those in which the researcher uses natural variation to determine a correlation between variables

and then to infer a cause and effect relationship. Experimental designs are those in which the experimenter manipulates one variable in the system to determine the effect on other variables. Both approaches have strengths and weaknesses, with experimental approaches generally being able to provide clearer results by reducing the number of variables that influence the outcome. The approach that you choose is often dictated by the system you are studying. For example, if you are studying large mammals, they are not easily manipulated so the observational approach is more appropriate in many cases. In some systems, experimental approaches can be implemented. For example, Resnick *et al*. (1990) were able to artificially manipulate stream environments to change the predation pressure and thus the selection pressure on the guppies (*Poecilia reticulate*) that lived there. This resulted in heritable changes in life history traits such as age at first reproduction, number of offspring and life span.

b. Marker selection. Determining which genetic markers you will use is key to a successful project and your choice of marker will vary with the temporal and spatial scale of your project (see also Chapter 2 for discussion about markers). For example, if your research question is about evolution of species in a family, then you will need to select markers (typically gene sequences) that are evolving at a relatively slow rate; alternatively, if population differentiation is your research focus then relatively rapidly evolving markers such as microsatellites will be more appropriate. In addition to selecting the appropriate type of markers, the number of markers is important. If you are sequencing, how many genes will be sampled? Do they represent the various genomes present in the species (i.e. mitochondrial, nuclear, chloroplast; see Chapter 2)? If you are using microsatellites or SNPs, how many loci are needed, and how variable do these loci need to be?

c. Sample size. Consider what sample size you require in order to reject your null hypothesis with confidence. This number will vary based on your research question but should include an evaluation of the number of individuals to be sampled, the number of locations to sample and the number of samples per location. Where possible try to conduct a power analysis to determine these values. A power analysis is a statistical analysis that can be used to determine the number of samples or replicates that are needed to detect a significant difference when it exists. Power analyses are difficult to do when you know little about your study species; therefore, it is often useful to carry out a preliminary study to get an idea of marker and sample variability. See Box 2.4 for more detailed directions regarding sample sizes.

d. Sampling methods. This step often involves a pilot study to ensure that you can collect the samples or data that you need. You may plan to study a rare species but a pilot study may show that your sampling method will not actually detect the species in question. For example, Ebert *et al*. (2009) conducted a pilot study to determine if their methods of non-invasively collecting hairs from wild boars (*Sus scrofa*) provided an unbiased sample of the population. Their results indicated that adult and sub-adult animals were not sampled equally, which could bias the results of larger studies. For some studies, this step will also involve determining the optimal sampling strategy. For example, in landscape genetics studies, collecting samples in a randomized way can reduce the chances that spatial autocorrelation will confound the results. Consider also the natural history of your study species; for example, if you are sampling a plant that can reproduce both clonally and sexually, are you aiming to maximize the number of genotypes that you sample, or to maximize the number of clonal replicates?

3. In some situations, your ideal design is the one that is carried forward. However, many times compromises need to be made because of various limiting factors. Part of the design process is to evaluate these constraints in an 'efficiency' step. In this step, you will trade off the time, effort, techniques and money available with the number of sampling sites, samples, genes or markers to a point where you can efficiently test your hypothesis without compromising the results.

4. Seek input regarding the overall feasibility of your research design. There are few things more annoying in research than to submit to a journal a report of all your hard work, only to be told by the reviewers that your sample sizes were inadequate for your conclusion, or that your sampling method was biased!

Quantifying Gene Flow

One of the most important determinants of genetic differentiation among populations is gene flow, which is essentially the movement of genes between populations. However, before looking at ways to quantify gene flow, we have to understand exactly what constitutes gene flow, and how it differs from dispersal and migration. Dispersal generally refers to the movement of individuals or propagules between discrete locations or populations, whereas migration refers to periodic movement to and from a specific geographical area, often occurring seasonally and along a consistent route. While either dispersal or migration must precede gene flow, neither result in gene flow unless individuals successfully reproduce once they have arrived at their new location. Nevertheless, dispersal is sometimes used as a surrogate for gene flow, and we need to be aware of this potential limitation when evaluating different methods of quantification. Understanding the movements of individuals, gametes (e.g. pollen) and genes between populations is fundamental to studies of ecology and population genetics because, as we shall see later in the chapter, gene flow can profoundly influence a host of relevant variables including population size, genetic diversity, local adaptation, and ultimately speciation. Unfortunately, quantifying the movements of organisms and their genes is seldom straightforward. In this section we shall look at several methods that are commonly used to estimate ongoing dispersal and gene flow. These methods can be divided into three main categories: direct, indirect and **assignment tests**.

Direct Methods

The colonization of new habitats, such as the invasion of new ponds by freshwater zooplankton, or the appearance of weeds in a recently cleared habitat, provides direct, incontrovertible evidence for dispersal, although such observations of dispersal impart very little information about the extent or frequency of movements between established populations. A direct method that can provide more detailed estimates of dispersal is mark–recapture. Individuals are first marked in one of a number of different ways that include rings on birds' legs, or paint markings on insect wings or carapaces (Figure 4.3). Direct evidence of dispersal is obtained when marked individuals are later recaptured some distance from their source population. A major drawback to this approach is that individual marking is a very time consuming exercise and often yields limited returns because many of the marked individuals will never be recaptured.

Figure 4.3 A ruddy darter (Sympetrum sanguineum) that can be identified from the label on its wing as individual K15. This dragonfly was marked using indelible ink and subsequently recaptured as part of a study on the dispersal behaviour of odonates (Conrad *et al.*, 1999). Photo provided by Kelvin Conrad.

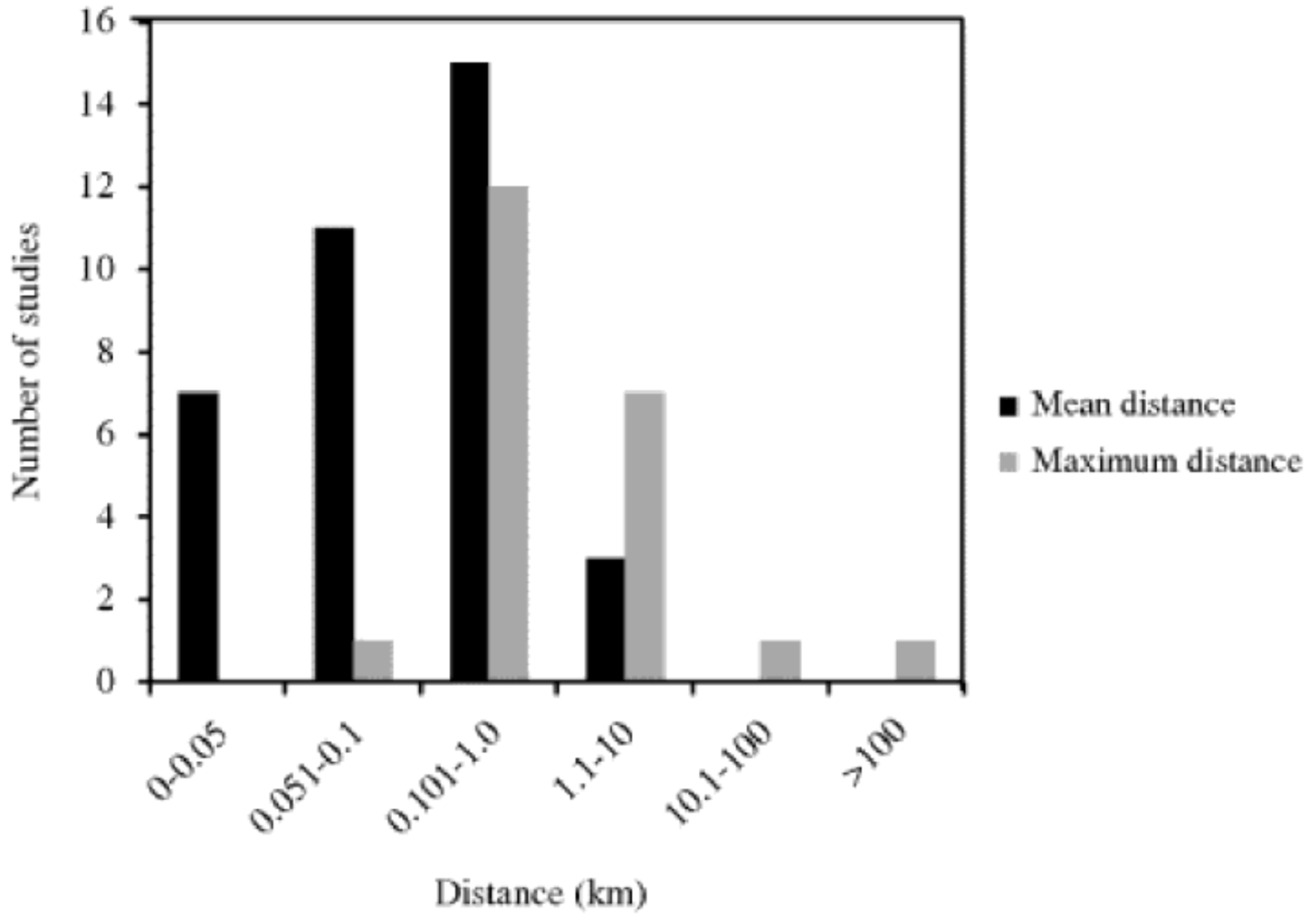
Even more intensive, although potentially a greater source of information, is the use of radio tracking or, in some cases, radar tracking. Radio transmitters have been used to monitor migratory movements of a number of taxa including birds, fish, and butterflies (Figure 4.4). Although the expense and laboriousness associated with radio tracking means that relatively few individuals can be monitored in this way, it does have the strong advantage of providing data on all movements, not just on the sites of recapture. In addition, global positioning technology and miniaturization is now allowing researchers to use satellite telemetry to track animals. These systems are still expensive and time consuming to deploy but provide 24-hour GPS positions and in some cases the instruments can also collect environmental (e.g. temperature) or biometric (e.g. heart rate) data.

Figure 4.4 A radio transmitter is surgically inserted into a pallid sturgeon (Scaphirhynchus albus), an endangered species in the Mississippi river drainage system. Researchers hope that radio transmitters will help them better understand and protect this species by tracking their movements and hopefully identifying possible spawning areas. Photo attributed to the US Geological Survey.

Material com direitos autorais

Radio tracking and mark–recapture studies provide estimates of dispersal but not gene flow, because they provide no information on whether or not immigrants will reproduce. Other direct measures of dispersal that do accurately reflect gene flow are based on parentage studies, which compare the genotypes of offspring and potential parents to determine how far seeds, pollen and other propagules have travelled. Parentage studies have been used most commonly in plants, and are generally most straightforward when seeds are collected from a maternal plant. The pollen parent can be identified by matching half the seed's genotype to the maternal plant, and then determining which potential pollen donor supplied the other half of the seed's genotype. Quantifying seed dispersal from parentage studies is also possible, but can be more challenging because in this case neither parent is known, and so the potential number of combinations of paternal and maternal plants is potentially much larger. Parentage studies have collectively provided some surprising insights into the frequency of long-distance pollen dispersal, having shown that it is more common than was previously thought in both wind-pollinated and animal-pollinated species (reviewed in Ashley, 2010; Figure 4.5). This has obvious implications for the population genetics of trees, and is also relevant to management and conservation. For example, one study of the Amazonian tree *Dinizia excelsa* focussed on trees that were isolated in forest fragments or in pasture land, and found that isolated trees produced on average three times as many seeds as trees in denser populations. They also found that isolated trees received pollen from multiple donor trees covering distances of hundreds of metres; in this case, the pollinators were introduced African honeybees which are more frequent pollinators in disturbed habitat than the native bees (Dick, 2001).

Figure 4.5 Summary of mean and maximum pollen-dispersal distances from multiple studies, all based on genetic parentage analysis. Data from Ashley (2010) and references therein.



Although parentage-based methods of inferring dispersal have been primarily employed in studies of plants, they can in theory be applied to any taxonomic group. One study of coral-reef damselfish (*Stegastes partitus*) genotyped 751 fish from 11 sites around the Bahamas and, despite the large population size and substantial area covered, were able to identify two parent-offspring pairs, both of which provided evidence of self-recruitment (larvae that remain at, or return to, their natal population). These two instances of self-recruitment based on such a small proportion of sampled fish suggests that self-recruitment occurs often, at least in the two populations from which it was identified through parent-offspring pairs (Christie *et al*., 2010).

Estimating gene flow by comparing the genotypes of offspring and putative parents can provide a comprehensive picture of dispersal, but this approach is impractical or impossible in large populations or in species that often travel long distances. In fact, a common denominator to all direct studies of dispersal is that they are costly, time consuming and can generally be conducted on only a relatively small number of individuals. Furthermore, with the exception of parentage analysis, direct evidence for dispersal does not necessarily translate into evidence of gene flow. For these reasons, many researchers have turned in recent years to indirect methods of estimating gene flow, and we shall look at the most common of these in the next section.

Indirect Methods

After Wright (1951) developed *F*-statistics, he went on to demonstrate that there is a simple relationship between the genetic divergence of two populations, measured as F_{ST}, and the amount of gene flow between them, which is given as:

Copyrighted material

$$F_{ST} = \frac{1}{4N_e m + 1} \quad (4.7)$$

in which N_e is the effective size of each population, and m is the migration rate between populations; in other words $N_e m$ is the number of breeding adults that are migrants. From equation (4.7), we can calculate $N_e m$ as:

$$N_e m = \frac{\frac{1}{F_{ST}} - 1}{4} \quad (4.8)$$

In the scarlet tiger moth example that was given earlier in this chapter, the F_{ST} value for the two populations was 0.009. This would translate into a very high indirect estimate of gene flow, because $N_e m = (1/0.009 - 1)/4 = 27.5$ individuals migrating between the populations each generation.

$N_e m$ is a popular method for estimating gene flow from genetic data, in part because of the ease with which it can be calculated. There are, however, several problems associated with estimates of $N_e m$. For one thing, it is based on what is known as the island model of population structure. This model assumes no selection or mutation, and an infinite number of populations, all of which are the same size, and which have an equal probability of exchanging migrants. $N_e m$ also assumes that populations are in migration-drift equilibrium, which occurs when the increase in genetic differentiation caused by drift is equal to the decrease in genetic differentiation caused by gene flow. Populations reach equilibrium only when population sizes and migration rates remain more or less constant; equilibrium is therefore negated by a number of demographic processes including recent range expansions, habitat fragmentation and population bottlenecks. Natural populations seldom, if ever, meet all of the assumptions that surround the calculation of indirect estimates of gene flow from F_{ST} (Whitlock and McCauley, 1999). Nevertheless, both empirical (Bohonak, 2010) and simulated (Slatkin and Barton, 1989) data indicate that calculations of $N_e m$ can be informative even if some of the basic assumptions have been violated.

Assignment Tests

A more recently developed suite of methods for quantifying dispersal from genetic data is based on what are known as assignment tests. These assign individuals to their most likely population of origin by comparing their genotypes to the genetic profiles of various populations. This approach differs from the indirect methods of assessing gene flow ($N_e m$) because it identifies individuals that have dispersed from their natal population, as opposed to comparing overall genetic similarities between populations. The original assignment tests use a maximum likelihood method to calculate the probabilities that a given genotype arose from alternative populations based on the allele frequencies in those populations (Paetkau *et al.*, 1995). Individuals are then assigned to the population from which they have the highest probability of originating, unless all of the probabilities are low in which case the true natal population may not have been sampled. This approach assumes that all populations are in Hardy–Weinberg equilibrium and that the loci being characterized are in linkage equilibrium.

In recent years, various modifications have made to improve the accuracy of assignment tests. One refinement has been the application of a **Bayesian** method (Pritchard *et al.*, 2000; Wilson and Rannala, 2003). The Bayesian approach to statistical inference is based on subjective statements of probability. Data are assumed to be fixed, and prior information is used to test the likelihood that various parameters can explain the data. Unlike classical statistics, which typically provide a single P value, Bayesian statistics often provide multiple probabilities, and this means that numerous scenarios (such as several candidate populations of origin) can be compared simultaneously. Simulation studies and applications to actual data sets have found that the Bayesian approach often outperforms the original frequency-based method (Cornuet *et al.*, 1999; Manel *et al.*, 2002), although this approach does assume that the true population of origin has been sampled, which may not be the case.

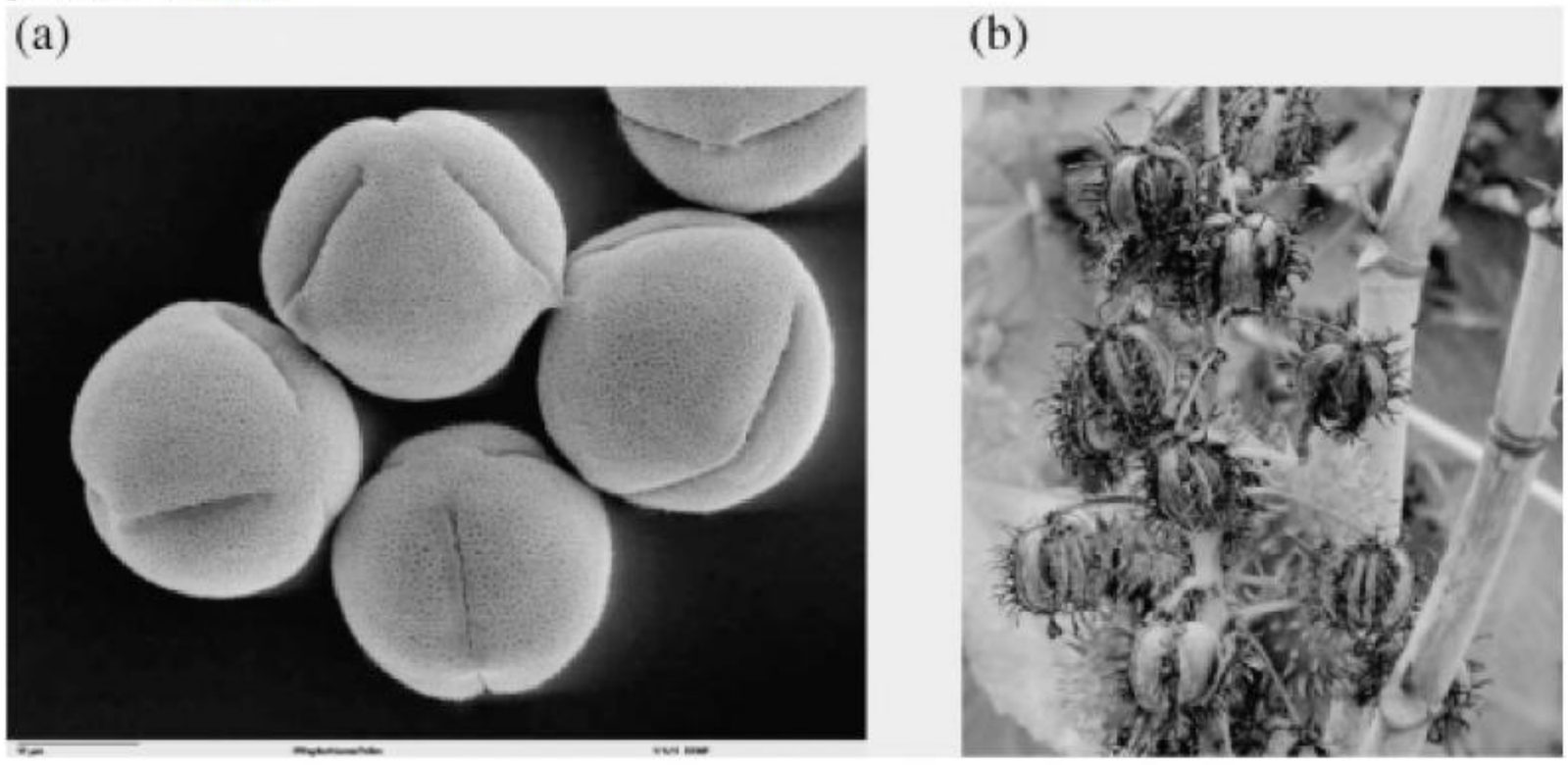
One advantage of assignment tests is that they allow researchers to identify the fate of individuals who are moving between populations. This distinction proved important in a study of the marbled murrelet (*Brachyramphus marmoratus*), a threatened seabird that nests in old-growth forests along the west coast of North America, and which in recent years has experienced a decline in numbers throughout much of its range. A population in central California is now geographically separated from the nearest substantial population, and while the population size has remained constant at approximately 600 individuals over the last few years, reproductive rates have declined substantially. This discrepancy can be at least partially explained by the results of assignment tests, which classified 7% (19 of 270) of murrelets in the central California population as migrants; however, only 10.5% of murrelets that were classified as migrants were part of parent–offspring pairs, compared to 45.4% of birds that were classified as residents (Peery *et al.*, 2010). This means that migrants are less likely to reproduce than residents, and has important implications for the long-term survival of a population that is seemingly bolstered by immigrants, but whose growth rate is actually declining.

What Influences Gene Flow?

Copyrighted material

There are both advantages and disadvantages to dispersal. Dispersing individuals may benefit by avoiding inbreeding, locating a new site with relatively few competitors, or escaping from pathogens, parasites or predators. On the other hand, they may be unable to locate a suitable new site or mate, or they may be preyed upon *en route*. Even if a seemingly suitable new habitat is located, at which conspecifics can be found, success is not guaranteed. In some populations of the damselfly *Calopteryx splendens*, males discriminated strongly against immigrant females. When resident males did mate with immigrant females, the females were more likely to have come from populations with similar predation pressures, and similar densities of conspecifics. These data suggest that in this species, gene flow (but not necessarily dispersal) is most likely to occur between ecologically similar populations (Wellenreuther *et al.*, 2010). From an evolutionary perspective, dispersal should occur only if the benefits outweigh the risks, and this cost–benefit analysis is part of the reason why the frequency, mode and distance of dispersal vary tremendously between species. However, this general statement is further complicated by the fact that numerous species have different dispersal mechanisms throughout their life cycle (e.g. pollen versus seeds (Figure 4.6), or swimming larvae versus flying adults), and therefore the benefits and risks to dispersal should also vary over time.

Figure 4.6 Many species have multiple dispersal mechanisms. (a) Pollen grains of the castor oil plant are dispersed by the wind, whereas (b) castor oil seeds are dispersed by ants (seed photo attributed to Dalgial).Raw castor beans are toxic to humans due to the presence of ricin.



Barriers to Dispersal

A species' ability to disperse is not dictated solely by its ability and propensity to move from one site to another: physical barriers may also impede dispersal and gene flow. Barriers to dispersal that may be relatively easy to recognize on land include water bodies and mountain ranges which limit the distribution of many ground-dwelling species, and dry land which is a barrier for many aquatic species (Figure 4.7). Major rivers, for example, are important barriers to gene flow for eastern chipmunks (*Tamias striatus*) (Chambers and Garant, 2010). Other important barriers that are less obvious to us include ocean currents and patterns of circulation, which can explain a lot of the genetic differentiation that exists among marine populations (White *et al.*, (2010a)). The cold-water upwelling close to the tip of South Africa, for example, has provided a barrier to long-term gene flow in a number of marine species including shallow water sea urchins in the genus *Tripneustes* (Lessios *et al.*, 2003). On both land and in sea, some barriers are particularly hard to identify because they are not represented by any obvious physical structure. These include things like gradients in nutrient concentrations or salinity, which can generate conditions to which populations or species are not adapted. The Rio de la Plata estuary in Uruguay, for example, hinders the dispersal of pelagic larvae of the crab *Armases rubripes* because the salinity and temperature are suboptimal for larval development and survival (Luppi *et al.*, 2003).

Figure 4.7 Aerial view of Death Valley, within the Mojave Desert. This desert includes the hottest, driest locations in the USA, and because it is uninhabitable to most species, it is often a barrier to dispersal. Photo attributed to Robert Campbell.

Copyrighted material

Landscape Genetics

The recent increase in the amount of genetic data that can be easily generated has allowed molecular ecologists to look for new ways to connect environmental data and genetic data in order to examine gene flow. Earlier in this chapter, we examined various methods to measure dispersal and gene flow among populations. These methods work best when the populations are differentiated to some degree; however, we can learn a lot about gene flow by using an approach called **landscape genetics**, which combines population genetics, spatial statistics and landscape ecology (Manel *et al.*, 2003; Stofer *et al.*, 2007; Holderegger and Wagner, 2008). By combining the techniques of each of these fields, researchers seek to quantify the influence that landscape features and environmental variables have on the distribution of allele frequencies among populations. The 'landscape' in landscape genetics refers to the area that the organism of interest is utilizing; for most animals this will consist of a number of different habitat types of varying suitability. If a population lives in a homogeneous landscape, we would expect no population structure (panmixia). If the habitat is homogeneous, but larger than the dispersal distance of an individual, then we might expect a pattern of isolation by distance to develop. However, most species exist as numerous populations that are connected by dispersal and which occupy landscapes that consist of different habitat types. Each habitat type may be easier or harder for individual animals to cross, and therefore gene flow will not be equal throughout the landscape. The aim of landscape genetics is to understand the relationship between habitats and gene flow (Figure 4.8).

Figure 4.8 One goal of landscape genetics may be to determine the effects that forest fires have on the distribution of alleles among populations of forest-dwelling animals. Photo attributed to Mark Wolfe.

Landscape genetics has a similar goal as phylogeography (Chapter 6), but can be distinguished by its finer scale of resolution: phylogeography is concerned with temporal scales that span hundreds to thousands of years (evolutionary time scales), while landscape genetics is interested in recent and contemporary gene flow (ecological time scales). In terms of spatial scales, phylogeography commonly examines the entire species' range, while landscape genetics is primarily concerned with landscapes that populations occupy at the time of investigation.

The data that are collected for landscape genetics studies include genetic data from highly variable markers (AFLPs, microsatellites, SNPs, etc.) and environmental data (habitat type, topological features, etc.). These data are linked by collecting the geographic position of each genotyped individual. Geographic Information Systems (GIS) can then be used to map each individual and genotype onto each

Copyrighted material

landscape variable. In fact, landscape genetics relies heavily on GIS technology to model the habitats that organisms and their genes must travel through.

The most common way to compare genetic data to other data is to create a pairwise matrix of genetic distances, and then to compare this to the pairwise matrix of geographic or environmental distances using Mantel tests. As discussed previously in this chapter, the simplest comparison is between a geographic distance matrix (using the straight-line distance between points) and a genetic distance matrix such as pairwise F_{ST} values or a relatedness measure. If there is a significant positive correlation between genetic and geographic distances, this indicates that there is a pattern of isolation by distance (Wright, 1943). This simple model assumes that all habitat patches between the points are equally suitable for the animal to cross, which is not likely to be the case. For example, because freshwater fish cannot move across land, a distance matrix that limits potential paths between locations to rivers should be better at describing the genetic relationships among fish captured at different locations than a straight-line, overland distance matrix.

To further improve on these models, the landscape can be divided into habitat types and each habitat type can be assigned a value, or cost, based on the suitability of that habitat for the species. The distance matrix can then be calculated as a series of least-cost-paths between points. These represent the cost (as a distance) for an animal to move between two points through different habitats. Using this method, researchers can infer the patterns of gene flow in a landscape by determining which set of habitat costs will best describe the observed genetic distances (i.e. by finding the best correlation between genetic and cost distances). For example, Vignieri (2005) studied the Pacific jumping mouse (*Zapus trinotatus*), which is primarily found in riparian habitats in alpine areas. In order to evaluate how this species disperses through mountains between watersheds, Vignieri created four distance matrices: straight-line (Euclidian) distances, river distances, overland (Euclidean distance factoring in elevation) distances, and habitat path distances. The habitat path model gave lower costs to riparian areas and lower elevation areas. Vignieri found that the highest correlation was with the habitat path model, which suggests that Pacific jumping mice move between watersheds by following rivers and streams as well as by crossing non-riparian habitats.

Landscape genetics can be used to detect barriers in a landscape that are relevant to a specific species. This is especially important for conservation and management, because anthropogenic factors increasingly fragment the landscape but the impact of fragmentation can be difficult to assess. Landscape genetics has been used in a number of instances to infer where these barriers occur. Pérez-Espona *et al.* (2008) evaluated the influence of a number of natural and anthropogenic landscape features on the genetic structure of red deer (*Cervus elaphus*) (Figure 4.9). They created a number of cost–distance matrices for various natural and man-made landscape features, and then correlated these with genetic distances. Their results indicated that sea lochs, mountain slopes, roads, and fenced-forest boundaries disrupted gene flow, whereas inland lochs and rivers facilitated gene flow (Pérez-Espona *et al.*, 2008).

Figure 4.9 Red deer (Cervus elaphus). Landscape genetics has determined that in this species, inland lochs and rivers facilitate gene flow, whereas sea lochs, mountain slopes, roads and fenced-forest boundaries reduce gene flow (Pérez-Espona *et al.*, 2008). Photo attributed to Mehmet Karatay.

Landscape genetics approaches have been applied to a wide variety of research fields, as well as to conservation. From an evolutionary perspective, landscape genetics provides a way to link changes in a single population with the overall evolution of a species, by revealing how alleles spread among individuals and populations. On a more practical level, landscape genetics has been used to study disease transmission. For example, Criscone *et al.* (2010) found that infections of the roundworm (*Ascaris lumbricoides*) within a small population of humans consisted of multiple infection foci in spite of the possibility of transfer from a single source that might be expected based on human (host) movements. They also found that certain households were at the centre of infection and that roundworms from nearby houses were more related to each other than to roundworms from houses that were farther away. They also determined that the infecting households tended to be stable over time, and suggested that this was due to the persistence of *Ascaris* eggs in the environment coupled with the tendency of the human population to defecate within 43 m (range: 2–200 m) of their household (Criscone *et al.*, 2010).

Copyrighted material

The development of landscape genetic approaches and techniques is an active area of research that will continue into the 21st century. New developments will include novel ways of calculating distance matrices in order to model gene flow (e.g. by incorporating circuit and network theory); new molecular markers that become available through next generation sequencing technologies; and novel applications of landscape genetics. The first step in the development of this field is finding ways to obtain increasingly precise explanations for current patterns of gene flow, while the second step will be using that information to predict future patterns of gene flow in response to things like anthropogenic stresses or climate change.

Metapopulations

In the absence of barriers to dispersal, some species that live in fragmented habitat regularly disperse between suitable sites, and this may lead to what is known as a **metapopulation**. The concept of a metapopulation dates back to the 1930s, although the actual term is attributable to Levins (1969) who presented an explicit model of a metapopulation. Levins' model, which we now know as a classical metapopulation, refers to a 'population of populations' that exist in a balance between extinction and recolonization, and are linked to one another by ongoing dispersal and gene flow.

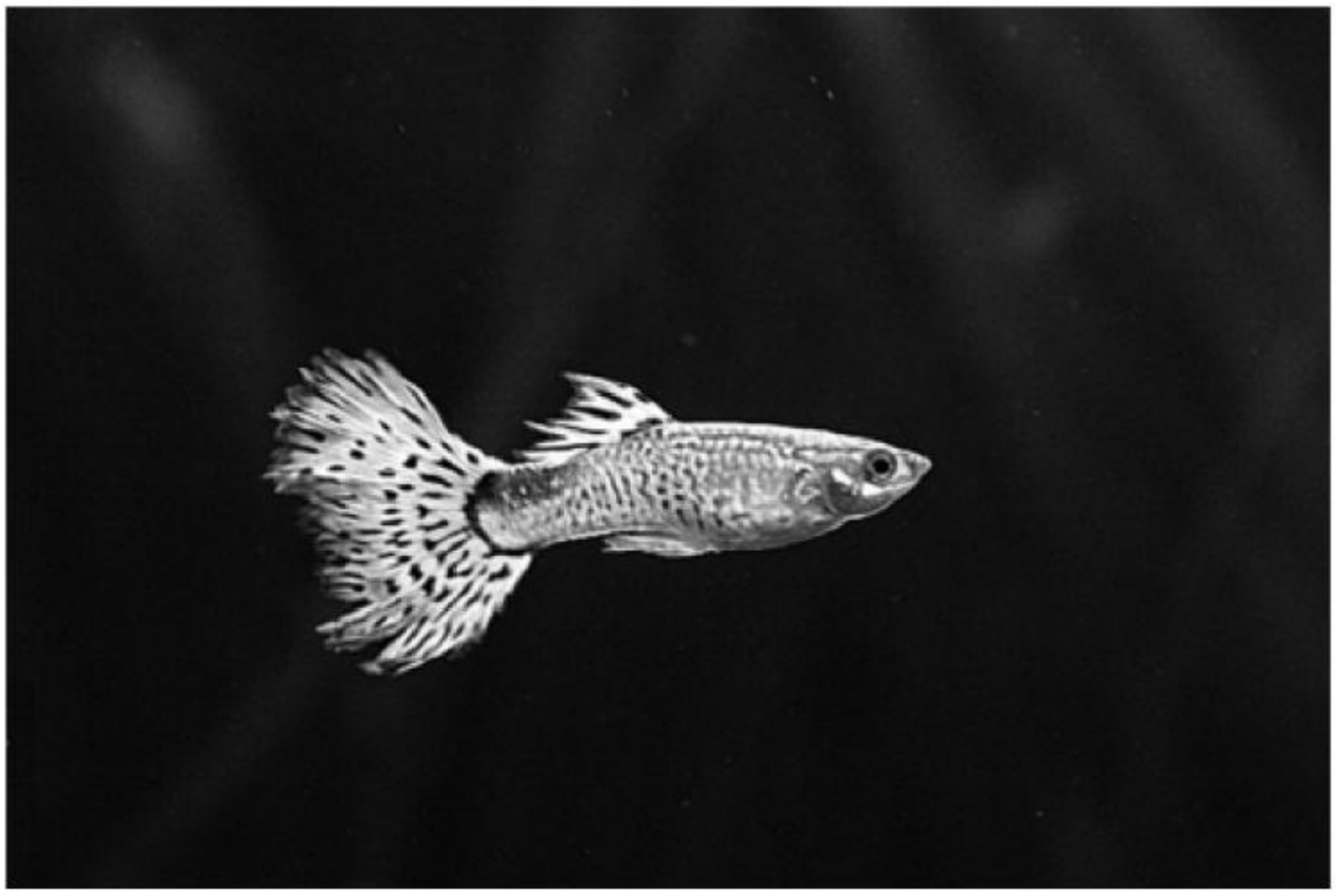
The process of repeated extinctions and recolonizations can affect the genetic structure of a metapopulation in different ways. Extinctions (sink populations) and recolonizations (from source populations) will often be accompanied by population bottlenecks which, as we know from Chapter 3, lead to reduced N_e, accelerated genetic drift and depleted genetic variation. N_e will be substantially reduced if all extant populations within a metapopulation descended relatively recently from a single ancestral population. Extinctions and recolonizations will also affect the extent to which populations are genetically differentiated from one another. Depending upon the balance between genetic drift and gene flow, differentiation among local populations may be enhanced if the founding of a new population is accompanied by a pronounced bottleneck, or diminished if gene flow remains high.

As the concept of a metapopulation has become more commonplace in the literature, several variations have emerged. A slight modification of the original concept, which required all subpopulations to have an equal probability of going extinct at any given time, is the island–mainland model. This describes a metapopulation in which a small number of regional populations (mainlands) remain relatively stable over time and act as sources of (re)colonizers for new populations (islands). More recently, the term has often been broadly adapted to refer simply to a series of conspecific populations that are connected to one another by dispersal. In accordance with this modified view, in which localized extinctions are not necessarily a prerequisite, a number of alternative metapopulation models have been created. There is, however, some debate in the literature over whether the latter simply represents a series of subdivided populations.

Microsatellite data were used to study guppies (*Poecilia reticulata*) (Figure 4.10), a freshwater fish living in the headwater floodplain area of the Caroni drainage system in Trinidad. Although populations in the uplands had high levels of genetic differentiation and low levels of genetic diversity, they bore the genetic signatures of long-term stability (e.g. no evidence of recent bottlenecks). Microsatellite-based evidence for ongoing gene flow between the headwater and lower tributaries demonstrated that large, fast-flowing rivers are not a barrier to gene flow among guppy populations. Downstream-biased migration and gene flow has led to lowland populations with relatively low levels of genetic differentiation and high levels of genetic diversity; in addition, migration among lowland populations was often substantial. Collectively, these data led the authors to conclude that guppies in the Caroni drainage system can be considered as a source–sink metapopulation, with the downstream populations representing a 'super sink' that continues to receive immigrants from rivers (sources) that are further upstream (Barson *et al.*, 2009). Although this may not be a 'classic' metapopulation, because there was no evidence of repeated extinctions and recolonizations, ongoing dispersal and the identification of source and sink populations allow this system to be identified as a metapopulation in the slightly looser sense of the word.

Figure 4.10 A guppy (Poecilia reticulata). In the Caroni river drainage system in Trinidad, this species lives in a source-sink metapopulation.

Material com direitos autorais



Interspecific Interactions

So far in this chapter we have focused on some of the studies of *intra*specific populations that have helped to unravel patterns of dispersal and gene flow. We shall now look at some of the ways in which *inter*specific relationships can also influence movements of both plants and animals. Interspecific dispersal can be facilitated in a number of ways, including host–parasite interactions; hitchhiking of seeds on mammal fur or bird feathers (Figure 4.11); transportation of pollen by insects, birds and bats; predator–prey interactions; and seed consumption. Another, less obvious, way in which interspecific interactions can influence gene flow is through hybridization.

Figure 4.11 A barbed seed trapped in the white fur of a cat. Many plants have seeds designed to get caught in animal fur as a means for dispersal. Photo attributed to Andrew Dunn.

Hybridization

Hybridization between divergent but related species can create a new avenue of gene flow between otherwise isolated genomes. The consequences of hybridization can be incredibly profound in terms of generating genetic diversity, producing novel ecological traits, and providing new pathways for adaptive evolution. In sunflowers for example, hybridization between the annual species *Helianthus annuus* and *Helianthus petiolaris* gave rise to three new hybrid species in North America: *Helianthus anomalus*, *Helianthus deserticola*

Copyrighted material

and *Helianthus paradoxus* (Rieseberg, 1991). The three hybrid species are adapted to novel habitats that are not usually colonized by either parental species (Rieseberg *et al*., 2003). There are many other examples in which gene flow between species leads to changes in species ecology and adaptive evolution.

Once thought of as an anomaly, interspecific hybridization is now known to be widespread in many different taxonomic groups. We do not know how many species hybridize in nature, although the frequency of interbreeding species is estimated to be quite high in a number of taxonomic groups including vascular plants (25%), European butterflies (12%), fruit flies (1%), birds (10%), British ducks (25%) and European mammals (6%); (Grant and Grant, (1992b); Mallet, 2005). Probably even more widespread is the hybridization that occurs within species following the interbreeding of 'individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more heritable characters' (Harrison, 1990). Many of the examples in this section will refer to interspecific hybridization, but while reading this you should keep in mind that much of the theory is equally applicable to hybridization between individuals from intraspecific populations or lineages.

In wild populations, an increasing number of hybrids have been identified in recent years from genotypic data. As we saw in Chapter 2, this is possible because hybridization leads to the introgression of alleles across species boundaries, a process that often results in cytonuclear disequilibrium (although recall from Chapter 2 that hybrids can be identified solely from nuclear data). Introgression (gene flow from one genome to another), combined with the preponderance of hybrids in virtually every taxonomic group, means that hybridization has become a relatively common explanation for unexpected distributions of alleles. Substantial introgression can occur even when hybridization is infrequent. Hybrids between native red deer (*Cervus elaphus*) and introduced Japanese sika deer (*C. nippon*) on the island of Argyll in Scotland were identified from mtDNA sequences and microsatellite alleles (Goodman *et al*., 1999). Hybridization between the two species is uncommon, and therefore introgression of alleles is rare at any one locus; however, where the two species come into contact with each other, up to 40% of deer carry alleles that have apparently been transferred from one species to the other.

Although we can quantify rates of hybridization between different species, it is not always easy to estimate the resulting gene flow and ecological consequences. Whether or not hybridization is very common, the ability of gene regions to pass from one species to another is often determined by many factors including genetic architecture, the fitness value of particular gene regions, and genotype-by-environment interactions. The application of **genomics** technologies, which will be discussed in greater detail in the next chapter, has greatly increased our understanding at a molecular level of intra- and interspecific gene flow.

The iris species *Iris fulva* and *I. brevicaulis* are sympatric in many regions throughout the southern United States. Hybridization between the two species is extremely limited because they flower at different times, have different sets of pollinators and grow in different habitat types. In addition, when hybrids do form, their viability and fertility is generally lower than those of the parental species (Arnold, 1994). It is therefore not surprising that F_1 hybrids are only very rarely formed in nature, although hybrid zones can be found where these two species co-occur, and substantial introgression has been observed around these zones. Martin *et al*. (2005) found that a number of alleles from *I. fulva* are associated with increased survivorship in hybrid individuals. Similarly, several alleles from *I. brevicaulis* are associated with high survivorship in hybrids, while one allele was associated with low survivorship. This study also showed that introgressed alleles from *I. fulva* can lead to either earlier or later flowering (depending on the locus and the allele), which can reduce reproductive isolation between the two species by leading to more synchronized flowering times (Taylor *et al*., 2009). When artificial hybrids were produced by back-crossing an F_1 hybrid to both parental species, *I. fulva* alleles were over-represented compared to *I. brevicaulis* alleles, indicating that gene flow is more likely to occur from *I. fulva* to *I. brevicaulis* than vice versa (Tang *et al*., 2010). Indeed, parallel findings have also been made in natural hybrid zones, which show high levels of introgression from *I. fulva* to *I. brevicaulis* (Arnold, 1994). Combined, these results provide evidence that some parts of the *Iris* genome are more permeable to introgression than others, and factors such as genetic architecture and natural selection play a large role in determining which gene regions are transferred from one species to another.

Scascitelli *et al*. (2010) reached similar conclusions with regard to hybrids between the sunflowers *Helianthus annuus* and *H. debilis*. The authors analysed several natural populations of *H. annuus* ssp. *annuus*, *H. debilis* and the hybrid subspecies, *H. annuus* ssp. *texanus*, using 88 microsatellite loci distributed across the genome. Analysis over all loci showed that between 14% and 27% of individuals in each of the populations represent recent immigrants from one of the other taxa. Migration rates between the parental taxa and the hybrid lineage are higher than migration rates between the two parental taxa, so the hybrid lineage acts as a bridge for gene flow between the two parental taxa. Overall, introgression from *H. annuus* to *H. debilis* was greater than vice versa. However, the authors found one locus at which there was no introgression from *H. annuus* to *H. debilis*, but very high levels of introgression from *H. debilis* to *H. annuus*. They also found several small genomic regions that showed low levels of migration and higher than expected levels of differentiation, which indicates that these regions are not subject to gene flow between the two species.

Box 4.5 Hybridization and genetic diversity

As we have seen, hybridization can have a large impact on genetic diversity *within* species, because it can result in the transfer of genes, or even large sections of the genome, from one species to another. Provided there are no fitness costs associated with these

episodes of genetic introgression (and there often are), hybridization can increase the genetic diversity of populations and species. This genetic reshuffling can have both positive and negative consequences; gene flow between genomes from different species can lead to the break-up of epistatic gene complexes (complexes of genes that work together, but are physically separate in the genome), which results in a decrease in fitness following a process known as hybrid breakdown. In contrast, recombination can also lead to the fixation of new combinations of beneficial alleles, and the expression of high fitness traits in hybrid individuals. Random recombination in a large number of hybrid offspring can create a heterogeneous pool of hybrid genotypes, such that hybrid individuals vary greatly with respect to each other and to parental individuals in terms of fitness. In many cases, genotypic and phenotypic variation in a hybrid population will not only be greater than that found in each parent population, but will also exceed the combined variation of both parent populations; this is known as **transgressive segregation**. At times this genetic rearrangement may result in stable populations with unique phenotypes, and when this happens a new species may emerge.

The extremely high levels of genetic diversity in populations of the outcrossing monkey flower *Mimulus guttatus* have been attributed to long-term introgression of genes from the selfing congeneric *M. nasutus* into *M. guttatus* (Sweigart and Willis, 2003). Another example of this phenomenon has been found in populations of Darwin's ground finches (genus *Geospiza*) (Figure 4.12), which experience regular bottlenecks that are often caused by environmental extremes such as drought or excessive rainfall. Despite these bottlenecks, populations show little evidence of depleted genetic diversity, and this seems to be at least partly attributable to ongoing hybridization among all six species (Freeland and Boag, 1999). The generation of genetic diversity following hybridization has been cited as one reason for the rapid evolutionary change that has led to **adaptive radiation** after species colonize new environments, two examples of this being Darwin's finches on the Galápagos Archipelago, and African cichlid fish in Lake Malawi (Seehausen, 2004). The acquisition of new genes through introgression can also be of practical concern, because bacteria frequently exchange antibiotic resistance genes. In an unusual case, there is evidence that some of these bacterial genes have even crossed a kingdom boundary and been incorporated into a fungal genome (Penalva *et al*., 1990).

Figure 4.12 Genetic diversity within Geospiza magnirostris, the large ground finch, is maintained in part by regular interspecific hybridization (Freeland and Boag, 1999). Source of figure: Darwin, C. R. ed. 1839. Birds, Part 3 No. 4 of The zoology of the voyage of H.M.S. Beagle, by John Gould. Edited and superintended by Charles Darwin. London: Smith Elder and Co.

Population Differentiation: Genetic Drift and Natural Selection

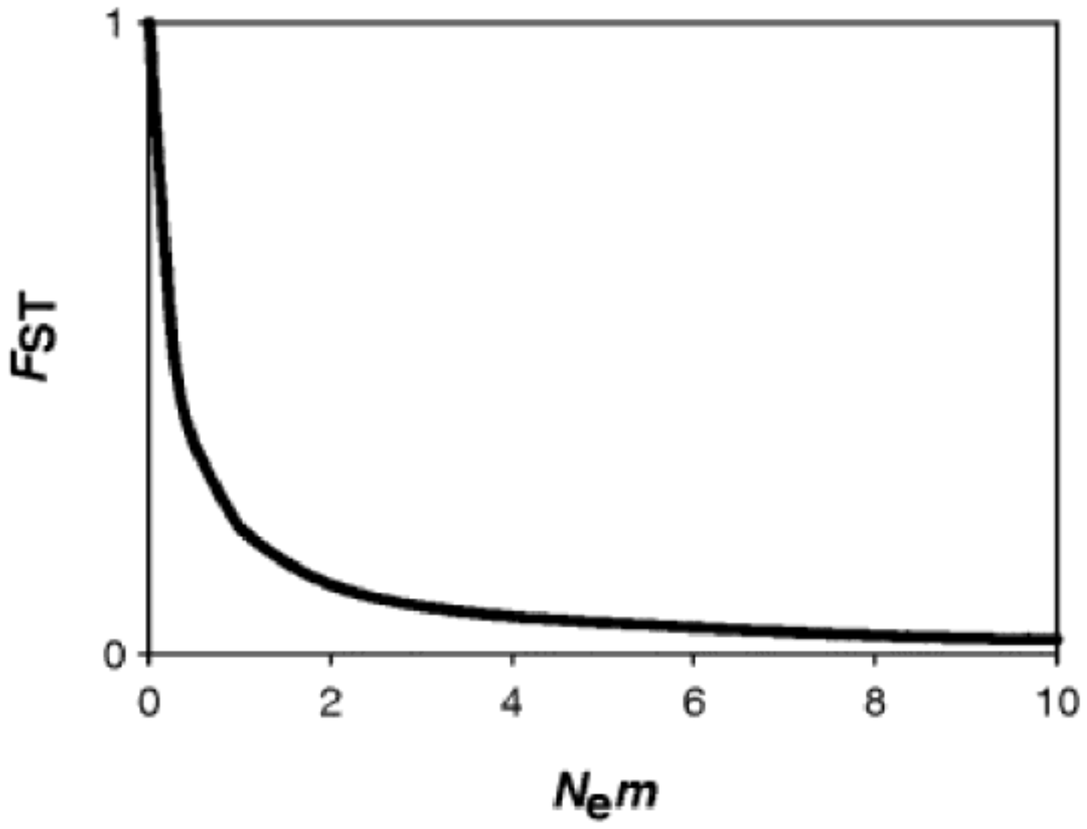
Gene flow is undoubtedly one of the most important processes in population genetics. In the absence of gene flow, a combination of mutations and genetic drift will cause populations to genetically diverge from one another, whereas in the presence of gene flow, populations can be held together as interconnected units that collectively contribute to a species' evolution. Nevertheless, gene flow does not necessarily make populations genetically homogeneous; populations may regularly exchange immigrants and still maintain a significant level of genetic divergence. Population differentiation in the face of ongoing gene flow has been one of the most intensely debated subjects in evolutionary biology. How much gene flow is necessary to prevent differentiation by genetic drift? How is local adaptation possible if individuals are regularly dispersing among sites? What are the relative roles of drift and selection in promoting population differentiation? These questions are of fundamental importance to the evolution of populations and the cohesiveness of species. In this section we will first look at how gene flow and genetic drift interact, although this will be a relatively short discussion as genetic drift was covered in some detail in Chapter 3. We will then devote the rest of this section to the seemingly paradoxical co-existence of gene flow and local adaptation.

Gene Flow and Genetic Drift

Copyrighted material

In the absence of gene flow, conspecific populations will generally diverge from one another as a result of genetic drift. However, very little gene flow is necessary to reduce the rate of genetic drift and thereby prevent substantial population subdivision. Consider once more the relationship between gene flow and population differentiation (F_{ST}). Earlier in this chapter we learned that when populations are in equilibrium, $F_{ST} = 1/(4N_em + 1)$. From this equation we can see that although populations are maximally divergent when $N_em = 0$ and $F_{ST} = 1$, even small increases in gene flow (N_em) will markedly reduce population differentiation (F_{ST}) (Figure 4.13). With only one migrant every fourth generation ($N_em = 0.25$), F_{ST} will be reduced to 0.5. If one migrant moves between a pair of populations every generation ($N_em = 1$), then $F_{ST} = 0.20$. Because F_{ST} is a measure of the degree of inbreeding within a local population relative to the collective population, and reflects the probability that two alleles drawn at random from within a population are identical by descent, an F_{ST} value of 0.2 means that the local population is only 20% more inbred than the collective population even though gene flow is still very low. In fact, theory predicts that N_em values as low as one migrant per generation may be sufficient to prevent the differentiation of populations by genetic drift (Wright, 1931).

Figure 4.13 A comparison of the genetic differentiation between populations (F_{ST}) and the accompanying indirect estimates of gene flow (N_em), based on the relationship $F_{ST} = 1/(4N_em + 1)$. F_{ST} values are rapidly reduced with even low levels of gene flow.



Because gene flow reduces the rate of genetic drift, it stands to reason that, all else being equal, isolated populations will have a lower N_e, a higher rate of drift and lower genetic variation compared to populations that receive immigrants (Table 4.5). Even low levels of immigration can introduce new genotypes and increase the effective size, and hence genetic diversity, of local populations. The importance of gene flow to the genetic diversity – and hence the survival – of threatened or endangered species was illustrated by a study of Morelet's crocodile (*Crocodylus moreletti*) in Belize, Central America. In the 1960s, Morelet's crocodile declined to critically low levels because of intense hunting pressure, although populations now seem to be recovering. Despite this recent bottleneck, microsatellite analysis of crocodiles from seven populations in north-central Belize revealed surprisingly high levels of within-population genetic diversity ($H_e = 0.49$), plus low levels of among-population genetic differentiation. Both of these findings are at least partially attributable to high levels of migration (approximately 5 migrants per generation among localities). One of the populations included in this study, New River Lagoon, is located in a relatively large and undisturbed area, and it appears that this population is acting as a genetic reservoir for the region. Regular dispersal from the New River Lagoon to other sites is increasing genetic diversity and reducing genetic drift within smaller regional populations (Dever *et al.*, 2002).

Table 4.5 Some of the ways in which gene flow (and lack thereof) can influence population genetics. Gene flow tends to be positively correlated with genetic diversity and N_e, and negatively correlated with the rate of genetic drift within populations, and the extent of genetic differentiation among populations

	Connected populations (gene flow)	**Isolated populations (no gene flow)**
N_e	↑	↓
Genetic drift	↓	↑
Genetic diversity	↑	↓
Population differentiation	↓	↑

Gene Flow and Local Adaptation

Genetic drift is not the only force that causes populations to diverge from one another. Environments are not homogeneous, and therefore different genotypes will be selected for in different parts of a species' range; this is known as local adaptation. A commonly cited example of local adaptation is heavy metal tolerance in plants and fungi. Suilloid ectomycorrhizal fungi (*Suillus luteus* and *S. bovines*), for example, were found to be highly zinc-tolerant when collected within 5 km of a zinc smelter, whereas those collected at least 15 km away from the pollution source were zinc-sensitive (Colpaert *et al.*, 2004). As with genetic drift, adaptation of populations to contrasting environmental conditions is a diverging force, but unlike drift, adaptation can occur even when gene flow is relatively high. The co-existence of these diverging and homogenizing forces has led to one of the most challenging questions in ecological genetics: how can adaptation occur in the face of gene flow?

The answer to this question involves a trade-off between the amount of gene flow and the strength of selection pressure acting on the trait in question. Selection pressures can be quantified on the basis of relative fitness levels. The phenotype (and its underlying genotype) that leaves the largest number of viable descendants within a population is considered to have a **fitness value** (w) of one. Because all other genotypes have a lower fitness, their fitness values must be less than 1. Consider a hypothetical population in which genotype A has the highest fitness ($w = 1$), and the relative fitness of genotype a is 0.75. This means that for every 100 A genotypes that survive, only 75 a genotypes will survive. A related measure is the **selection coefficient** (s), which is the corollary of w and a measurement of the reduced probability of survival. It is calculated as:

$$s = 1 - w \quad (4.9)$$

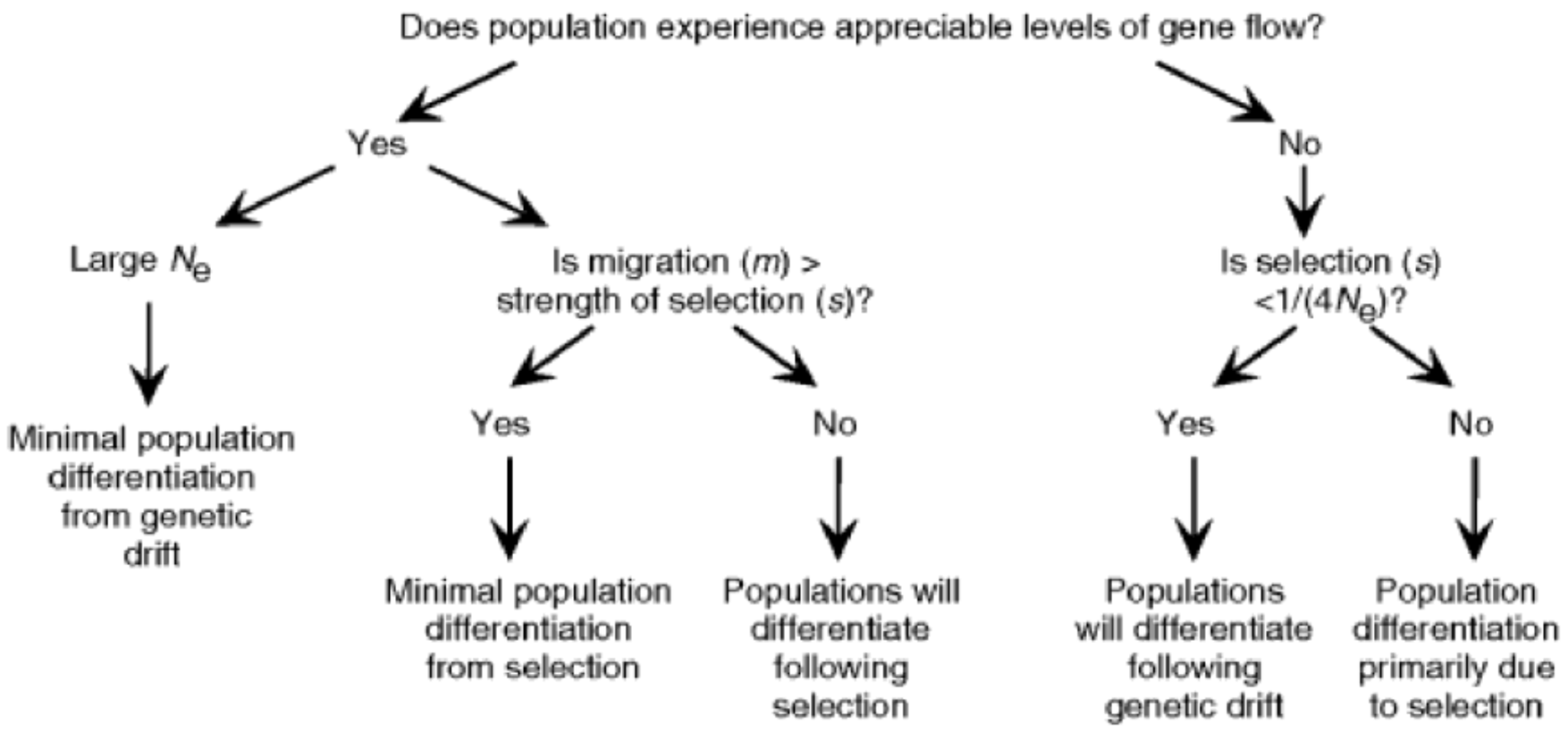
The selection coefficient against genotype a is calculated as $s = 1 - 0.75 = 0.25$, in other words survival of a genotypes is predicted to be 25% lower than A genotypes in each generation. For any given genotype, the higher the selection coefficient, the stronger the selection pressure against that particular genotype will be, and the more likely that genotype is to become extinct. One of the best known examples of selection coefficients is industrial melanism in the peppered moth *Biston betularia*. Although most populations were light coloured before the industrial revolution, dark forms of the moth became favoured in polluted areas because they are better camouflaged against blackened trees. The selection coefficient against light coloured moths in these regions has been estimated at 0.32, making their chances of survival approximately 32% lower than dark-coloured moths (Haldane, 1924).

If selection pressure is high enough, local adaptations may persist despite high levels of gene flow (Endler, 1977). This has been demonstrated in a number of species including the Dominican anole (*Anolis oculatus*), a morphologically variable member of the family Iguanidae that is endemic to Dominica. Morphological diversity in this species, which includes variable tail depth, scale size and body size, reflects adaptation to environmental variables that include rainfall and type of vegetation, both of which vary markedly across the island. Populations therefore show substantial levels of morphological differentiation, but at the same time microsatellite data revealed surprisingly high levels of gene flow, with only 7.8% of population pairs revealing F_{ST} values that were significantly higher than zero (Stenson *et al.*, 2002). In another example, populations of Galápagos lava lizards (*Microlophus albemarlensis*) showed no evidence of genetic differentiation at six out of seven microsatellite loci, despite having differences in anti-predator behaviour that most likely reflect local adaptation (Jordan *et al.*, 2005). The anole and lava lizard studies provide just two of the numerous examples in which local adaptation has been maintained despite high levels of ongoing gene flow.

Drift Versus Selection

Because gene flow can impede local adaptation only if the migration rate (m) exceeds the strength of selection (s), it follows that if gene flow is low, even relatively weak selection pressures can accelerate the divergence of two populations from one another. This, however, is only part of the story; we must also remember the potential contribution of genetic drift to population differentiation. So how do drift and selection interact? Key here is the size of populations (N_e). If selection is strong relative to the population size, in other words if s is much greater than $1/(4N_e)$, then the effects of drift will be negligible. If, however, s is much less than $1/(4N_e)$, then changes in allele frequencies will be largely attributable to genetic drift, although mutations can also make a contribution. The interactions of gene flow, drift and selection are summarized in Figure 4.14.

Figure 4.14 Genetic differentiation of populations depends on a combination of gene flow, effective population size and natural selection.



There are a number of reasons why we may wish to know whether drift or selection is a more important driver of population differentiation. For one thing, natural selection may mean that populations in one part of a species' range may have much lower fitness if transplanted elsewhere, and this is an important consideration in reintroduction programmes in conservation biology (see Chapter 8). In practice, however, it is often extremely difficult to obtain estimates of s and N_e in wild populations, meaning that a comparison of these two variables is seldom a practical way of assessing the relative importance of selection. One alternative method for identifying local adaptation is through reciprocal transplant experiments, in which individuals are sampled from one or more populations and then transplanted to other sites of interest. Individuals are then monitored as they develop at both their home site and in the sites to which they were relocated. At each population the relative fitness values of native and introduced individuals are compared, and a higher fitness in native plants provides evidence for local adaptation. Common garden experiments can also be conducted, in which individuals from different populations are grown in a common garden environment (e.g. greenhouse or environmental chamber) in which environmental variables (e.g. nutrients or water) can be manipulated, and the survival and reproduction of individuals from different populations can be monitored and compared. Both reciprocal transplant and common garden experiments have been used to great effect in many studies of plants and fungi, although they tend to be impractical (reciprocal transplant) or very challenging (common garden) when studying more mobile species. Molecular ecologists therefore often rely on other methods for inferring patterns of natural selection and local adaptation, some of which we will outline below.

Patterns of Molecular Evolution

We know from Chapter 3 that one way to infer selection is to compare sequences from different populations and calculate the proportion of synonymous and non-synonymous substitutions (d_N/d_S ratio). Synonymous substitutions do not alter the encoded amino acid and are therefore usually neutral, whereas non-synonymous substitutions are much more likely to cause phenotypic change. Patterns of substitution can therefore provide a way of identifying natural selection, because non-synonymous substitutions should be proportionately higher in selected genes. Table 4.6 lists some examples in which the ratio of non-synonymous to synonymous substitutions has been used to infer natural selection.

Table 4.6 Examples of genes in which selection has been inferred from a nonsynonymous/synonymous substitution ratio that is greater than one. After Ford (2000) and references therein

Taxa	**Gene**	**Function**
Salmonids	Transferrin	Resistance to bacterial infection
Humans, mice, fish	Major histocompatibility complex (MHC)	Immunity
Flowering plants (Solanaceae)	S-locus system	Self-incompatibility
Filamentous fungi	het-c locus (heterokaryon incompatibility)	Regulation of self/nonself-recognition during vegetative growth
Marine gastropods	Lysin genes	Proteins used by sperm to create a hole in the egg vitelline envelope
Sea urchins	Binding genes	Proteins that attach sperm to eggs during fertilization
Drosophila	Acp26Aa	Male ejaculate protein
House mouse	Abpa	Androgen-binding protein
Plasmodium	Surface protein genes	Detection of parasites by hosts

Urheberrechtlich geschütztes Material

One advantage to inferring selection from substitution ratios is that this method allows us to compare two categories of nucleotide substitutions along a single stretch of DNA, as opposed to comparing substitution rates from two or more gene regions that may have had different evolutionary histories. However, there are also some weaknesses associated with this method. For one thing it is not infallible; patterns of substitution will not reveal all instances of selection, in part because, as we saw in Chapter 1, sometimes only a single nucleotide change can have significant phenotypic consequences. Furthermore, we can use this method only if we know which gene has been selected for, and identifying candidate genes can be difficult. In the next section we will look at how a more general approach based on the comparison of data from multiple markers may give us clues about local adaptation.

Discordant Genetic Differentiation

Migration and drift are expected to have approximately equal effects on all neutral loci, whereas the effects of selection will vary between neutral and non-neutral loci. We therefore expect all neutral loci to show similar levels of genetic divergence among populations, whereas non-neutral loci (or loci linked to non-neutral loci) are expected to show anomalous levels of divergence. These anomalous levels may be either unusually high or unusually low, depending on the type of selection that the relevant genes have been subjected to: directional selection will increase population differentiation if different alleles are selected for in different populations, whereas stabilizing selection can decrease population differentiation if the same alleles are selected for in multiple populations. By comparing multiple measures of population differentiation that are each based on a different locus, researchers may discover a marker that shows unusual levels of differentiation and which therefore may indicate a genetic region that is under selection.

The ease with which data can now be collected simultaneously for many markers means that studies of discordant genetic differentiation have increased in recent years. A growing number of these studies are based on a genome scanning approach, which means that hundreds of markers are used to obtain a broad sample from the entire genome (as opposed to a handful of microsatellite loci), and this increases the likelihood that markers linked to genes under the influence of natural selection will be captured. In one such study (Herrera and Bazaga, 2008), 341 polymorphic AFLP markers were scored from 188 individuals of the violet *Viola cazorlensis* (Figure 4.15), which is an endemic species that occurs in disjunct populations distributed throughout the limestone mountains of south-eastern Spain. Different populations exhibit significant differences in flower morphology including lengths of the floral peduncle, floral spur and the upper, middle and lower petal blades. However, based on initial studies, it was not clear whether differences in floral traits were the result of genetic drift or divergent selection. The authors found that nine loci provided extremely high measures of F_{ST} compared to all other markers, and these were flagged as markers for candidate adaptive genes. The F_{ST} values calculated from the remaining, presumably neutral, loci indicated low levels of differentiation between populations (mean pairwise $F_{ST} = 0.0719$), indicating that gene flow between populations is quite high, which is not entirely surprising because the primary pollinator of this violet species is a hawk moth that migrates over long distances. Interestingly, the authors found correlations between allele frequencies and floral traits at four out of the nine outlier loci, indicating that these four loci are linked to genes that control floral traits, which in turn are likely under the influence of divergent selection. This provided evidence that floral divergence is a result of selection rather than genetic drift, although the authors were not able to propose any selective mechanisms that drive floral divergence in this plant species.

Figure 4.15 A flowering violet (Viola cazorlensis). In this species, floral divergence appears to be influenced by natural selection (Herrera and Bazaga, 2008).

Copyrighted material

Discordant genetic markers are often used to identify seemingly non-neutral genes, although caution must be taken when interpreting these types of data, because variation in F_{ST} among different loci can sometimes result from stochastic variation in genetic drift. This randomly generated variance of F_{ST} among loci may be particularly high if population sizes have fluctuated over time, because the loss of rare alleles during recent bottlenecks may impact some loci more than others. False positives may also arise if neutral markers are linked to markers that are under selection (although as we shall see in the next chapter, having a marker linked to an adaptive gene can be useful). There are now statistical algorithms that allow us to calculate error rates based on our own data sets, in order to estimate the likelihood that we will erroneously identify a locus as an outlier. In other words, we can quantify the likelihood that we will mistake a neutral locus for a locus under selection. The previous example of violets used a very restrictive P value of 0.0005 for identifying outlier loci, which was not expected to yield any false positives, or erroneously identify outlier loci. Had the authors used P values of 0.01 or 0.05, they would have detected a greater number of outlier loci (or loci expected to be under the influence of selection), but they would have also had a high rate of false positives (i.e. 3 and 15 loci would have erroneously been identified as outliers using the latter two P values, respectively).

Care must also be taken when comparing F_{ST} measures that are based on different types of genetic markers, because markers often mutate at different rates. The effects of mutation on F_{ST} should be negligible if populations are in drift-migration equilibrium, but rapidly evolving markers may yield relatively high F_{ST} values if new mutations are not dispersed between populations rapidly enough to attain equilibrium between gene flow and genetic drift. When all F_{ST} values are higher in one set of markers (e.g. all microsatellites) compared to another set of markers (e.g. all allozymes), discrepancies are more likely to be attributable to different mutation rates than to natural selection. To date, numerous studies have identified outlier loci as candidate genes under selection, and a growing number have proceeded to the next step, which is to identify the function of candidate genes (Piertney and Webster, 2010) (Table 4.7).

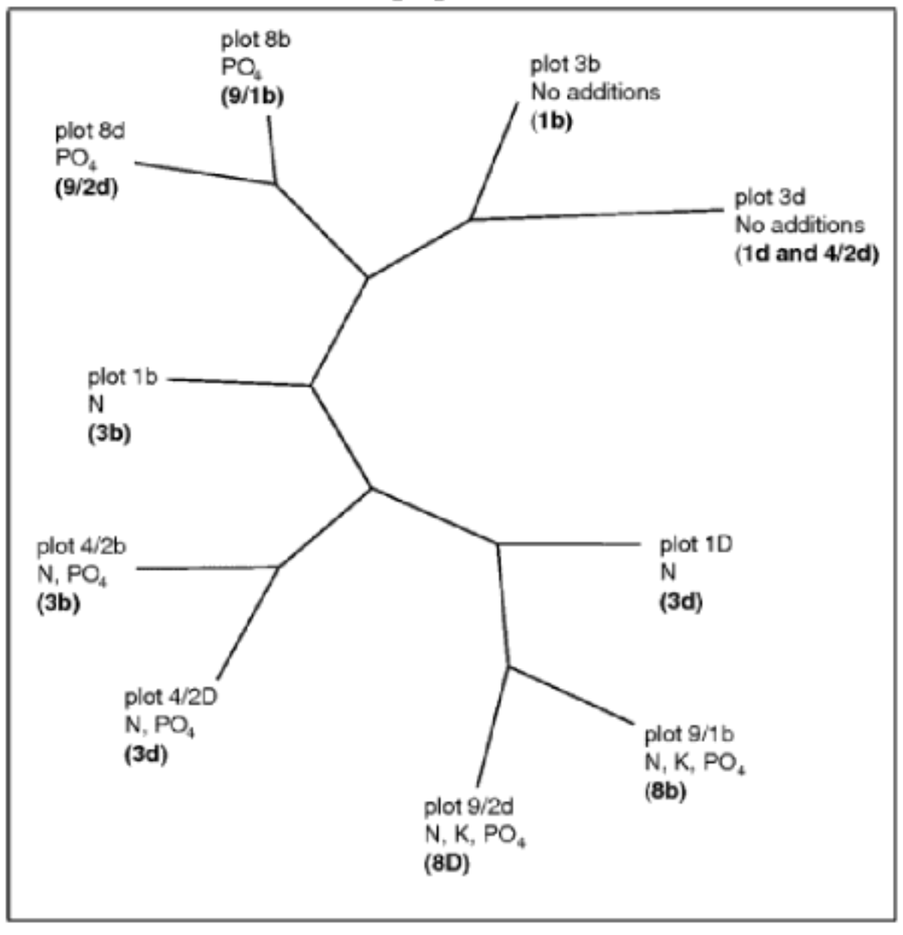
Table 4.7 Some of the candidate genes and the phenotypic traits that they influence in natural populations of non-model species. Adapted from Piertney and Webster (2010)

Copyrighted material

Candidate gene	Species	Phenotypic trait	Reference
Calmodulin (CALM1)	Darwin's finches (*Geospiza* spp.)	Beak morphology	Abzhanov *et al.* (2006)
cGMP-dependent protein kinase (KGP1)	Honey bee (*Apis mellifera*)	Foraging behaviour – division of labour	Ben-Shahar (2005)
Pantophysin (PAN1)	Cod (*Gadus morhua*)	Growth	Case *et al.* (2006)
Heat shock protein (HSP70)	Silver sea bream (*Sparus sarba*)	Adaptation to changing salinity	Deane and Woo (2004)
Phosphoglucose isomerase (PGI)	Glanville fritillary butterfly (*Melitaea cinxia*)	Dispersal	Haag *et al.* (2005)
Thyroid hormome receptor alpha	Ambystomatid salamanders	Timing of metamorphosis	Voss *et al.* (2003)
Wingless	Heliconius butterflies	Wing patterning	Kronforst *et al.* (2006)
Tyrosine related protein kinase I (TRYP1)	Soay sheep (*Ovis aries*)	Coat colour polymorphism	Gratten *et al.* (2007)

On a final note, a growing number of studies have shown that if divergent selection is sufficiently strong to promote reproductive isolation between populations, this creates a barrier to gene flow that can result in the genome-wide differentiation of populations following genetic drift. This will lead to an inverse correlation between gene flow and the adaptive divergence of populations, and thus a positive association between the phenotypic divergence and molecular genetic differentiation of populations, a pattern that is known as isolation by adaptation (IBA) (Nosil *et al.*, 2009; Thibert-Plante and Hendry, 2009). In other words, populations will show genome-wide divergence and not simply divergence at those loci that are under selection. This pattern was identified in a semi-natural experiment in which adjacent sweet vernal grass (*Anthoxanthum odoratum*) populations were growing on plots that had received different nutrient additions for up to 150 years. Despite their proximity, populations had differentiated from one another at outlier loci (identified as markers that may be in, or linked to, genes that are under selection), and also across a wider survey of putatively neutral (AFLP) loci (Figure 4.16). This was interpreted as evidence that the selection pressures from different nutrient regimes in adjacent plots were strong enough to cause reproductive isolation between populations, despite the high potential for gene flow, and this in turn has led to neutral genetic differentiation as a result of genetic drift (Freeland *et al.*, 2010).

Figure 4.16 Tree showing the genetic similarities between Anthoxanthum odoratum populations that had been exposed to different nutrient additions for 150 years. The nearest neighbours on the tree are genetically most similar to one another. Each branch is labelled with a plot number (e.g. 9/1b), with the nutrients that had been added (nitrogen – N, phosphate – PO_4, potassium – K), and with the geographically nearest neighbour (written in bold under the nutrient additions). The general trend is that the genetically most similar populations are those that received the same nutrient additions and not those that are physically closest to each other. This, along with the identification of candidate outlier loci, was interpreted as evidence that adaptation to nutrients has driven the genetic differentiation of these populations (Freeland et al., 2010).

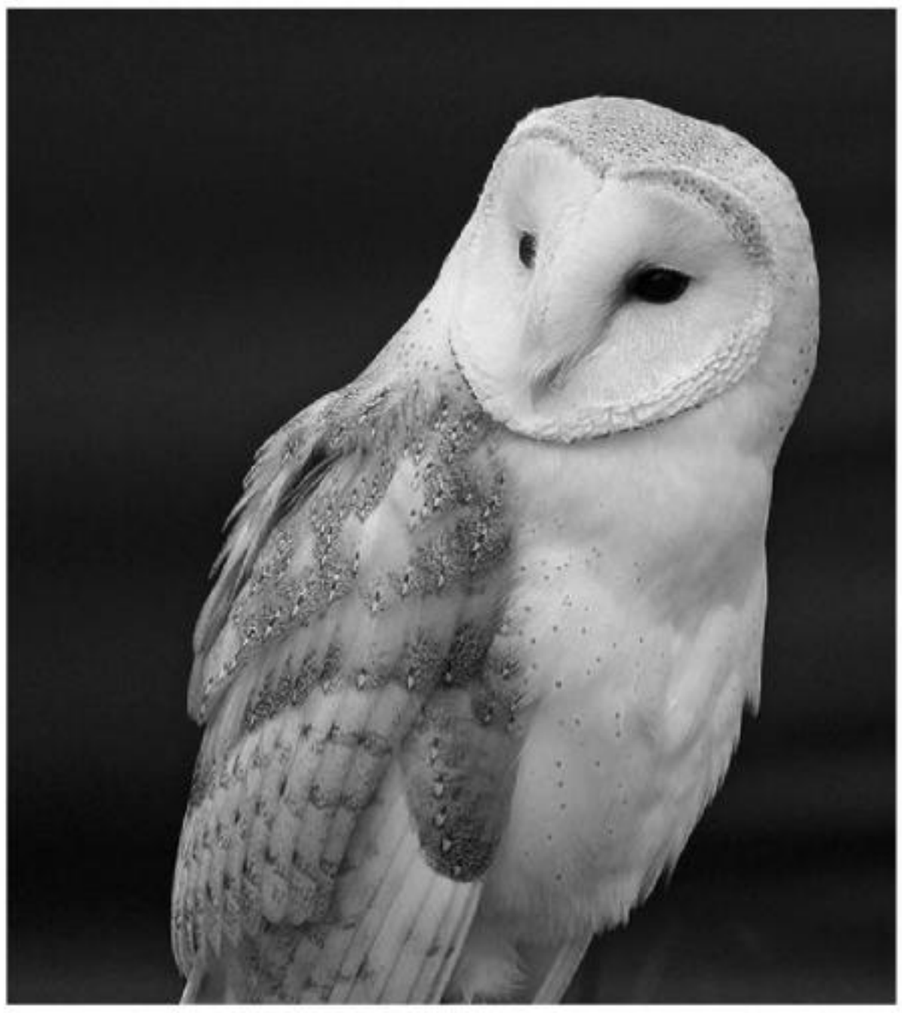


Copyrighted material

Clinal Variations in Allele Frequencies: F_{ST} versus Q_{ST}

Another way to look for evidence of natural selection is to examine changes in allele frequencies along geographical **clines**. These are caused by a progressive change in one or more environmental variables, for example a change in photoperiod along a latitudinal gradient. If particular alleles are associated with environmental conditions then this may be indicative of selection. In Europe, the barn owl (*Tyto alba*) (Figure 4.17) exhibits a cline in its plumage coloration, in which the underside of its body varies continuously from white in the south-west to reddish-brown in the north-east (Roulin, 2003). To determine whether or not this cline was neutral or adaptive, Antoniazza *et al.* (2010) first quantified population differentiation of plumage coloration by calculating Q_{ST} between 18 populations across a maximum distance of 2395 km. Q_{ST} is a calculation that is based on **quantitative traits**, which are traits that are influenced by several different genes. Many important traits are quantitative, including height, weight and measures of reproductive fitness such as clutch size in birds and time to flowering in plants. These contrast with **qualitative traits** which are discrete and controlled by one or a few loci, such as the round versus wrinkled seeds that were made famous by Mendel's experiments. Quantitative traits are controlled by a suite of genes known as **quantitative trait loci** (QTL), and Q_{ST} can be calculated on the basis of variation at these traits (Box 4.6).

Figure 4.17 A barn owl (Tyto alba). Plumage coloration varies across Europe and this seems to be a result of natural selection (Antoniazzi *et al.*, 2010).



In a cline, in which gene flow is expected to occur between populations, selection will act only on those QTL that underlie an adaptive phenotypic trait; neutral alleles are expected to show non-adaptive differentiation that will be influenced by a combination of gene flow and genetic drift. Therefore, if QTL are based on neutral quantitative traits then Q_{ST} and F_{ST} values should show similar levels of population differentiation (assuming that F_{ST} is based on neutral molecular markers); if, however, QTL are under selection then they should show either higher (divergent selection) or lower (stabilizing selection) levels of population differentiation compared to F_{ST}. In European barn owl populations, overall plumage colour Q_{ST} (which was actually calculated as P_{ST}; see Box 4.6) was 0.353, whereas overall F_{ST} based on microsatellite markers was only 0.011 (Antoniazza *et al.*, 2010). This discrepancy strongly suggests that there is an adaptive basis to the cline of plumage coloration in European barn owls. Note, finally, that comparisons of Q_{ST} and F_{ST} are not used solely to investigate clinal variation: they can be used to confirm or refute the possibility of adaptive quantitative traits among any pairs of populations. More examples of Q_{ST} and F_{ST} comparisons are given in Table 4.8.

Table 4.8 Some examples of studies that have inferred natural selection on quantitative traits from comparisons of Q_{ST} and F_{ST}

Copyrighted material

Species	Average Q_{ST}	Average F_{ST}	Suggested explanation	Reference
Endemic Mediterranean plant (*Brassica insularis*)	0.023 (seedlings) 0.087 (adults)	0.213	Stabilizing selection on phenotype	Petit *et al.* (2001)
Scots pine (*Pinus sylvestris*)	0.80	<0.02	Direction selection for date of bud burst	Karhu *et al.* (1996)
Great snipe (*Gallinago media*)	0.416	0.018	Directional selection for leg length	Saether *et al.* (2007)
Sea rocket (*Cakile maritima*)	0.53–0.86	0.155	Directional selection for multiple morphological traits	Gandour *et al.* (2008)
Wild barley (*Hordeum spontaneum*)	0.55–0.90	0.108–0.207	Directional selection for multiple morphological traits	Volis *et al.* (2005)

Box 4.6 Calculating Q_{ST}

Quantitative trait loci (QTL) are often strongly influenced by environmental effects, for example time to flowering in plants will depend on external factors such as temperature and photoperiod as well as the relevant genes. The variance in quantitative traits can therefore be partitioned into the amount of variation that is due to genetic factors, which is known as the genotypic variance ($I = \sum_{i=1}^{m}(p_{ix}p_{iy})/[(\sum_{i=1}^{m}p_{ix}^2)(\sum_{i=1}^{m}p_{iy}^2)]^{0.5}$), and that due to environmental factors, which is known as the environmental variance (σ_g^2). Differentiating between the contributions made by environmental and genetic factors to the phenotypic variation of quantitative characters can be problematic, because many of these characteristics are phenotypically plastic. In laboratory settings, breeding experiments and pedigree analyses can be used to good effect, but these are seldom practical for wild populations. Heritability can also be estimated from parent-offspring regressions, which provide a comparison of parent and offspring phenotypes. If parent and offspring phenotypes consistently show a strong similarity to one another regardless of environmental influences, then genotype variance (and hence heritability) is high, and environmental variance is low. Once the genetic (heritable) component of the QTL has been estimated, the genotypic variance of QTLs can be partitioned within and between populations. This is designated Q_{ST}, and is comparable to F_{ST} because it represents the degree to which populations are genetically differentiated. Q_{ST} is calculated as:

$$Q_{ST} = \frac{\sigma^2_{g(between)}}{(\sigma^2_{g(between)} + 2\sigma^2_{g(within)})} \quad \textbf{(4.10)}$$

in which σ_e^2 is the amount of genotypic variance between populations, and $\sigma^2_{g(between)}$ is the amount of genotypic variance within populations.

Although Q_{ST} should be based on estimates of genotypic variance, there are logistical reasons why such estimates are not always possible, in which case researchers often estimate Q_{ST} based on a set of assumptions. It has been proposed that these surrogates of Q_{ST} be differentiated from 'true' Q_{ST} estimates by referring to them as P_{ST}, for phenotypic or pseudo-Q_{ST} (Saether *et al.*, 2007) (although the validity of using P_{ST} values to study wild populations has been questioned by some; Pujol *et al.*, 2008). In the barn owl study referred to in the text, for example, estimates of Q_{ST} were made from plumage coloration, which depends on melanin, and although a previous study had shown that this has a strong heritable component (Mundy, 2006), the authors who compared plumage Q_{ST} to genetic F_{ST} did not test heritability of plumage coloration in their study, and were therefore using a measure that may be better described as P_{ST} (Antoniazza *et al.*, 2010).

Overview

Broadly speaking, gene flow maintains populations as cohesive units, whereas genetic drift and natural selection can cause populations to diverge and eventually split into distinct species. Throughout this chapter we have been primarily concerned with relatively recent adaptive and non-adaptive changes within populations. We will continue to explore recent and historical adaptive changes in the next chapter, where we will learn how genomics techniques can be applied to ecological questions.

Chapter Summary

- Most species consist of multiple populations. The extent to which these populations are differentiated from one another can be quantified in a number of ways including Nei's standard genetic distance (D).
- F-statistics use inbreeding coefficients to partition genetic variation within and between populations. F_{IT} reflects the total amount of inbreeding in a population, F_{ST} reflects the amount of inbreeding in a population that is due to the differentiation of subpopulations, and F_{IS} reflects the amount of inbreeding that occurs within subpopulations. F_{ST} is a widely used measure of how genetically distinct (sub)populations are from one another.
- The amount of gene flow between populations is one important determinant of genetic differentiation. Dispersal and gene flow can be quantified using direct methods such as radio-tracking; indirect methods, which provide estimates of $N_e m$; and assignment tests.
- Dispersal patterns vary widely across and within taxonomic groups. Dispersal is particularly complex in plants, fungi and invertebrates because many species within these groups have different dispersal mechanisms throughout their life cycle, and may also disperse through time following prolonged diapause.
- Dispersal does not necessarily lead to gene flow. Gene flow can be determined from both direct and indirect assessments.
- Barriers to dispersal will promote the genetic differentiation of populations. Landscape genetics combines environmental and genetic data to infer patterns of gene flow.
- Species that experience relatively high levels of ongoing gene flow between subpopulations may be described as living within a metapopulation. In a classical metapopulation structure, subpopulations also experience repeated extinctions and recolonizations.
- Hybridization between closely related species can lead to new opportunities for gene flow. Rates of hybridization do not necessarily reflect rates of gene flow, because different genomic regions are likely introgressed from one species to another at different rates, which are determined by factors that include genetic architecture and natural selection.
- Genetic drift and natural selection can both cause populations to diverge. Gene flow tends to homogenize populations, although divergence between populations can occur despite ongoing gene flow if selection (s) is strong enough. Conversely, if the migration rate (m) exceeds the strength of selection (s) then local adaptation will be hindered by the continued introduction of alleles from other populations.
- Selection can be inferred from the proportion of non-synonymous substitutions, from discrepancies in the levels of differentiation that are revealed at different loci, or from clinal variations in allele frequencies.
- Q_{ST} measures the genetic variance of quantitative trait loci (QTLs) within and among populations. If Q_{ST} is either more or less than F_{ST} then this too may be evidence of selection.

Useful Websites and Software

- GenAlEX, a user-friendly package for population genetic analysis that runs within Microsoft Excel, and includes features for researchers and teachers (e.g. sample data sets and tutorials): http://www.anu.edu.au/BoZo/GenAlEx/w
- Adegenet, R package for genetic analysis: http://adegenet.r-forge.r-project.org/
- MatSAM – a spatial analysis software to detect candidate loci for selection: http://www.econogene.eu/software/sam/default.asp
- Geneland – Bayesian algorithm for determining population structure using both genetic and spatial data: http://www2.imm.dtu.dk/~gigu/Geneland/
- Genepop on the web – an online population genetics software package: http://genepop.curtin.edu.au/
- Arlequin – software for population genetics data analysis: http://cmpg.unibe.ch/software/arlequin35/
- STRUCTURE – uses multi-locus genotype data to investigate population structure: http://pritch.bsd.uchicago.edu/software.html
- Isolation by distance web service – an online program for determining whether your data show IBD: http://ibdws.sdsu.edu/~ibdws/codominant.html
- Nature Network population genetics forum: http://network.nature.com/groups/popgen/forum/topics
- SMOGD – software for the measurement of genetic diversity: http://www.ngcrawford.com/django/jost/

Review Questions

4.1. A study of six house sparrow (*Passer domesticus*) populations in Western Canada revealed an average F_{ST} of 0.136 and an average F_{IS} of 0.085. What is the extent of non-random mating within all six populations combined?

4.2. From the data in question 4.1, calculate the average indirect measure of gene flow among the house sparrow populations.

4.3. An investigation into four populations of the damselfly (*Coenagrion puella*) revealed an average N_em of 8.8 per generation, whereas mark–recapture studies identified only four immigrants during the 8-week breeding season. How could you explain this discrepancy?

4.4. What factors determine the rate at which two populations genetically diverge from one another?

4.5. F_{ST} values between two populations of the vernal sweetgrass *Anthoxanthum odoratum* were estimated from AFLP and microsatellite data as follows:

AFLP	
Locus 1	0.124
Locus 2	0.098
Locus 3	0.103
Locus 4	0.699
Locus 5	0.111
Locus 6	0.107

Microsatellites	
Locus 1	0.327
Locus 2	0.396
Locus 3	0.421
Locus 4	0.372
Locus 5	0.345

Do these data suggest that natural selection is acting on any of the loci?

4.6. What conclusion might you reach about height QTL in ten populations of sunflowers (*Helianthus annus*) situated along a latitudinal gradient if $Q_{ST} < F_{ST}$?

Further Reading

Books

Arnold, M. L., 2006. *Evolution Through Genetic Exchange*. Oxford University Press, Oxford.

Futuyma, D. J., 2009. *Evolution*, 2nd edn. Sinauer Associates, Inc., Sunderland, Massachusetts.

Lowe, A., S. Harris and P. Ashton. 2004. *Ecological Genetics*. Blackwell Publishing, Oxford.

Review articles

Ashley, M. V., 2010. Plant parentage, pollination, and dispersal: How DNA microsatellites have altered the landscape. *Critical Reviews in Plant Sciences***29**: 148–161.

Broquet, T. and E. J. Petit, 2009. Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology Evolution and Systematics***40**: 193–216.

Hellberg, M. E., 2009. Gene flow and isolation among populations of marine animals. *Annual Review of Ecology Evolution and Systematics***40**: 291–310.

Guillot, G., R. Leblois, A. Coulon and A. C. Frantz, 2009. Statistical methods in spatial genetics. *Molecular Ecology***18**: 4732–4756.

Holderegger, R. and H. H. Wagner, 2008. Landscape genetics. *Bioscience***58**: 199–207.

Jenkins, D. G., M. Carey, J. Czerniewska, *et al*., 2010. A meta-analysis of isolation by distance: relic or reference standard for landscape genetics? *Ecography***33**: 315–320.

Leinonen, T., R. B. O'Hara, J. M. Cano and J. Merila, 2008. Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology***21**: 1–17.

Manel, S., O. E. Gaggiotti and R. S. Waples, 2005. Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution***20**: 136–142.

Manel, S., M. K. Schwartz, G. Luikart and P. Taberlet, 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution***18**: 189–197.

Morjan, C. L. and L. H. Rieseberg, 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology***13**: 1341–1356.

Nosil P., T. H. Vines and D. J. Funk, 2005. Perspective: reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution***59**: 705–719.

Pearse, D. E. and K. A. Crandall, 2004. Beyond F_{ST}: Analysis of population genetic data for conservation. *Conservation Genetics***5**: 585–602.

Pujol, B., A. J. Wilson, R. I. C. Ross and J. R. Pannell, 2008. Are Q(ST)-F-ST comparisons for natural populations meaningful? *Molecular Ecology***17**: 4782–4785.

Ronce, O., 2007. How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annual Review of Ecology Evolution and Systematics***38**: 231–253.

Segelbacher, G., S. A. Cushman, B. K. Epperson, *et al.*, 2010. Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics***11**: 375–385.

Sexton, J. P., P. J. McIntyre, A. L. Angert, K. J. Rice. 2009. Evolution and ecology of species range limits. *Annual Review of Ecology Evolution and Systematics***40**: 415–436.

Weir, B. S. and W. G. Hill, 2002. Estimating F-statistics. *Annual Review of Genetics***36**: 721–750.

Online activities

The analysis of multiple populations is the core of many molecular ecology studies; therefore, we have provided several data sets of varying complexity as part of the online exercises relating to this chapter. Each data set has been formatted for relevant software (GENPOP and GENALEX) so that students can easily start running analyses that will allow them to investigate differentiation among divided populations. The data sets that we have provided start with a simple scenario of just two populations and progress to data sets of five populations. We have also provided the data sets in STRUCTURE format so that students can examine data in this widely used programme that does not assume any *a priori* population structure. Online activities can be found at www.wiley.com/go/freeland_molecular2e.