28

Inferences from Spatial Population Genetics

F. Rousset

Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution, Montpellier, France

This chapter reviews theoretical models and statistical methods for inference from genetic data in subdivided populations. With few exceptions, these methods are based on neutral models of genetic differentiation and have been mainly concerned with estimation of dispersal rates. However, simulation-based methods allow to draw inferences under models involving additional demographic processes such as changes in dispersal rates over time. The formulation and main results of migration matrix, island, and isolation-by-distance models, are briefly described. The definition and basic properties of F-statistics are reviewed, and moment methods for their estimation are contrasted with likelihood methods. Then, the application of the different methodologies to simple biological scenarios is reviewed. Their practical performance is discussed in light of comparisons with demographic estimates, as well as of their robustness to different assumptions and of concepts of separation of timescale.

28.1 INTRODUCTION

Since the advent of molecular markers in population genetics, there have been many efforts to define methods of inference from the spatial genetic structure of populations. This chapter can only review a small selection of them including, in particular, some recent developments of simulation-based likelihood methods, and also of less sophisticated methods in so far as they provide analytical insight and proven performance in realistic conditions. With few exceptions, I will focus on allele frequency data; some methods for other types of data are described in **Chapter 29**.

The perspective taken in this review is that studies of spatial population structure are conducted in order to make inferences about parameters considered important for the evolution of natural populations, for example, for the dynamics of adaptation. Thus,

Handbook of Statistical Genetics, Third Edition. Edited by D.J. Balding, M. Bishop and C. Cannings.
© 2007 John Wiley & Sons, Ltd. ISBN: 978-0470-05830-5.

all such analyses should ultimately be based on models of evolution in subdivided populations. This would lead to the identification of important parameters in such processes and to the formulation of appropriate statistical models to estimate them (assuming it is useful to estimate them in order to test the models). In this perspective, the material reviewed below may seem imperfect not only because the statistical models are approximate but also because the important evolutionary parameters are not always clearly identified.

In all inferences, we will consider a total sample from a population structured by restricted dispersal in a number of demes (a technical term used in the analysis of the models) or subpopulations (a somewhat looser term). The population concept must be carefully distinguished from another concept of 'population' often considered in statistics, which actually refers to the probability distribution of samples under some model. In general the value of a variable in the biological population is not the expected value of this variable in this statistical 'population', in other words, this is not an expected value in a theoretical model. In practice the word parameter is used for both, but here it will be used only for the value in the theoretical model.

A statistical corollary is that by sampling only one locus, one may compute estimates which will approach the value in the biological population, rather than the parameter value, as more individuals are sampled. In other words, it will approach a value that will depend on the realized genealogy in the biological population, and this will be a random variable. The usual solution to this problem is to analyze several loci with different genealogies, assumed independent. For a nonrecombining DNA, it may not be very useful to sequence longer fragments: Since the whole DNA has the same genealogy, any estimate will depend on the single realized random genealogy in the biological population sampled (see **Chapter 25**).

28.2 NEUTRAL MODELS OF GEOGRAPHICAL VARIATION

The major models considered for statistical analysis describe the evolution of neutral genetic polymorphisms; among models of selected markers, statistical analysis will be considered only for clines.

28.2.1 Assumptions and Parameters

We consider a set of subpopulations each with N_i adults, and with dispersal rates m_{ij} from subpopulation j to subpopulation i. These dispersal rates are defined as the probability that an offspring had its parent in some subpopulation: Thus they are defined by looking backward in time (backward dispersal rates), rather than by looking where offspring go (forward dispersal rates). Forward and backward rates will differ, for example, when individuals that disperse at longer distances have a higher probability of dying before reproduction.

These models are known as *migration-matrix models*, with migration matrix (m_{ij}). Limit cases of these models when all deme sizes $\to \infty$, all backward rates $\to 0$, with their products $N_i m_{ij}$ remaining finite, have been described as 'structured coalescents' (see e.g. **Chapter 25**). With many subpopulations, the number of parameters may be large. However, some symmetric structure is usually assumed, as in the island and isolation-by-distance models developed below. Further, the migration matrix, as well as the subpopulation sizes N_i, are supposed to be invariant in time. These assumptions allow for more detailed mathematical analysis. Simulation-based methods have allowed to investigate more complex historical scenarios involving range expansions, interruptions of gene flow, and so on. A relatively well-worked case is the isolation-with-migration model (Nielsen and Wakeley, 2001), according to which an initially panmictic population differentiates at some time T in the past into two subpopulations that will keep on exchanging migrants at rate m until the time of sampling.

The island model (Wright, 1931a) with n_d subpopulations is the simplest form of migration-matrix model: For different subpopulations i, j, the dispersal rate is supposed to be independent of i, j and may be written as $m_{ij} = m/(n_\mathrm{d} - 1)$ where m is the total dispersal rate; $m_{ii} = 1 - m$. The subpopulation sizes $N_i = N$ are also supposed to be independent of i. The infinite island model is the limit process as $n_\mathrm{d} \to \infty$. This is the most often considered model, because of its ease of analysis. However, it should be noticed that most of the results of the infinite island model with $N_i = N$ can easily be extended to infinite island model with N_i different for different subpopulations and with total dispersal rate into each subpopulation i being a function of i (see the discussion of (28B.1)). Thus the main defining assumption of such island models is that immigrants may come with equal probability from any of the other subpopulations.

Dispersal is often localized in space, so that immigrants preferentially come from close populations. Two kinds of models that take this into account have been considered, one for demes on a discrete lattice, and one for ‘continuous’ populations (e.g. Malécot, 1951; 1967). In a continuous population, the local density may fluctuate in space and time, but there is no rigorous mathematical analysis of models incorporating such fluctuations. In the lattice models, different demes are arranged on a regularly spaced lattice and the dispersal rates are a function of the distance between demes. There is a fixed number of adults, N, in every generation on each node of the lattice. Thus the position of individuals is rigidly fixed and density does not fluctuate. The island model may be recovered as a specific case.

In models of isolation by distance, the parameter σ^2 often appears (e.g. Malécot, 1967; Nagylaki, 1976; Sawyer, 1977). This is an average squared distance between parent and offspring. In two dimensions this is the average square of the projection of the two-dimensional (vectorial) distance on an axis, also known as the *axial distance* (Figure 28.1). This parameter is a measure of the speed at which two lineages descending from a common ancestor depart from each other in space. The models as formulated above may be generalized to include age- or stage-structure, and it is possible to generalize some of the results for island and isolation by distance models given below in terms of concept of effective dispersal rate and effective deme size or effective population density, albeit through some approximations (Rousset, 1999a; 2004, Chapter 9). Then, effective dispersal is the asymptotic rate of increase of the second moment of distance between two independently dispersing gene lineages per unit time. The definition of population density also needs to be generalized. First, it is actually not simply a density but a rate of coalescence per surface and per unit time (Rousset, 1999a). In the basic models, it can be computed as the expected number of coalescence events per generation among all pairs of genes in the total population, divided by the total surface occupied by the population (or habitat length in linear habitats). With age structure, it can be computed as a weighted

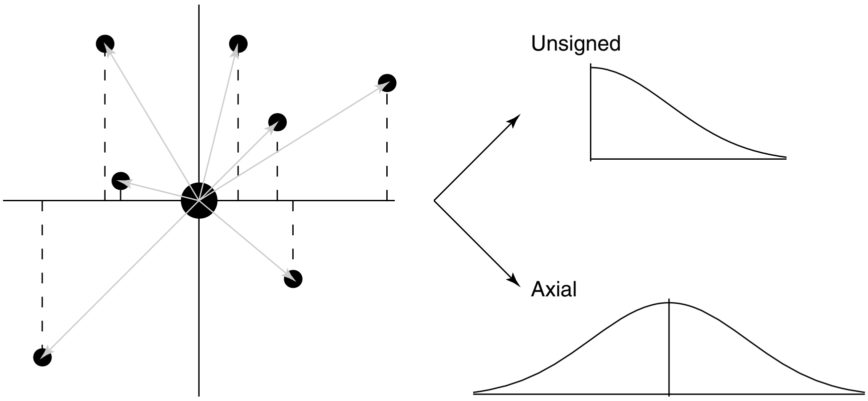

Figure 28.1 σ^2 in two dimensions. One considers the two-dimensional dispersal distances (gray arrows) between one parent (large central dot) and different offspring (or different parent–offspring pairs). The projection of these vectors on two axes yield signed axial distances on each axis. In terms of variance, σ^2 is the variance of the distribution of one such axial dispersal distance (bottom right). This is *not* the variance of the unsigned dispersal distance (top right).

average of such among all pairs, these being weighted by reproductive value weights as in the computation of the effective dispersal parameter.

28.3 METHODS OF INFERENCE

With few exceptions, explicit formulas for the likelihood of samples under the models formulated above are not available. This section therefore focuses on moment methods for which explicit analytical results are available, and on simulation methods for likelihood inference.

28.3.1 *F*-statistics

Moment methods are based on the analysis of moments of order k of allele frequencies. By far the most common of them (analysis of variance) consider only squares of allele frequencies or equivalently frequencies of identical pairs of genes. This is the basis for the theory of F-statistics in population genetics. Autocorrelation methods (e.g. Sokal and Wartenberg, 1983; Epperson and Li, 1997) are constructed from pair-wise comparisons of genes or genotypes, hence there should be essentially the same information in such statistics as in the more standard moment methods. The relationship between autocorrelation methods and some of the methods described below is discussed by Hardy and Vekemans (1999).

28.3.1.1 Probabilities of Identity and F-statistics

To define genetic identity, we consider a pair of homologous genes and ask whether they descend without mutation from their most recent common ancestor. If no mutation has occurred since the coalescence of ancestral lineages, there is *identity by descent* (IBD).

By *identity in state* (IIS) of a pair of genes we simply consider whether they have the same sequence (if the alleles are distinguished by their sequence), the same length (if the alleles are distinguished by the number of repeats of a microsatellite motif), the same electrophoretic mobility, etc. In short, we only look at the allelic state of a gene. IBD is a specific case of IIS for the infinite allele model, in which each allele produced by mutation is considered different from preexisting alleles. The generic notation Q will be used to denote expected values of IIS under any model.

If we consider a population structured in any way (age, geography, etc.), one may always define Q_{w}, the IIS probability within a class of genes (for example among individuals of some age class, in the same subpopulation, etc.), and Q_{b}, the IIS probability between two different classes of individuals. In a generic way one may then define:

$$F \equiv \frac{Q_{\mathrm{w}} - Q_{\mathrm{b}}}{1 - Q_{\mathrm{b}}}. \tag{28.1}$$

Such quantities are known as *F-statistics*, but Q_{w}, Q_{b}, and F as defined above are parameters. That is, Q_{w} and Q_{b} are expectations under independent replicates of the stochastic process considered, and F is a function of these parameters. In other words, they are functions of the parameters that define the model under study, such as subpopulation sizes, mutation rates, migration rates, etc. If (say) deme size is by itself random, then F and the Qs, being expectations in the process considered, are function of the parameters of the distribution of deme size. In models of spatial genetic structure, Q_{w} and Q_{b} are generally not 'the value in the (biological) population'. Alternative definitions of F-statistics, as values in biological populations, have been used in the literature (e.g. Nei, 1986; see Nagylaki, 1998 for further discussion), but analytical results below hold only with the present parametric definitions.

Let Q_2 be the IIS probability within subpopulations, and Q_3 be the IIS probability between subpopulations. The well-known F_{ST} parameter, originally considered by Wright, is best defined as

$$F_{\mathrm{ST}} \equiv \frac{Q_2 - Q_3}{1 - Q_3}. \tag{28.2}$$

F-statistics may be described as correlations of genes within classes with respect to genes between classes, that is as intraclass correlations (Cockerham and Weir, 1987).

28.3.1.2 Generic Methods for Estimation and Testing

The estimation of F-statistics is described at length in the literature (see e.g. **Chapter 29**, Weir, 1996) so I will confine myself to emphasizing a few easily missed points.

A simple way to estimate parameters such as F_{ST} is to estimate each of the probabilities of identity by the corresponding frequencies $\hat{Q}$ of identical pairs of genes in the sample, computed by simple counting. Thus F_{ST} may be estimated by

$$\hat{F} \equiv \frac{\hat{Q}_2 - \hat{Q}_3}{1 - \hat{Q}_3}, \tag{28.3}$$

where $\hat{Q}_2$ and $\hat{Q}_3$ are by definition the frequencies of identical pair of genes in the sample, within and between deme, respectively. This simple approach to defining estimators may be easily adapted to a number of different settings, given that the parameters to be estimated may be expressed as functions of probabilities of IIS. With balanced samples, this approach directly yields Cockerham's estimator of F_{ST} (Cockerham, 1973; Weir and Cockerham, 1984). This estimator has been developed by analogy with the methods of analysis of variance, and this analogy has proved difficult to understand. Appendix A details the nature of the analogy and its relationship with 28.3.

It is easy to test for differentiation (nonzero F_{ST}) by the usual exact tests for contingency tables either applied to gametic or genotypic data. These are standard statistical techniques and their application to genetic data has been discussed elsewhere (e.g. Weir, 1996; Goudet *et al.*, 1996; Rousset and Raymond, 1997). A general set of techniques to draw confidence intervals from the moment estimators are the bootstrap (e.g. Efron and Tibshirani, 1993) and related techniques based on the resampling of loci. However, the simplest applications of resampling techniques may be misleading. This may be apparent when they lead to symmetric confidence intervals while the variance of the estimator is expected to be sensitive to the parameter value, in which case more involved uses of the bootstrap (DiCiccio and Efron, 1996) may be required.

28.3.1.3 Why F-statistics?

Wright was the first to note that such measures of genetic structure appear in some theoretical models of adaptation, and his ideas remain among the most influential in population genetics. He used them to quantify his 'shifting balance' model Wright (1931a; 1931b), which remains controversial today (Coyne *et al.*, 1997). Nevertheless, F-statistics are useful descriptors of selection in one-locus models (Rousset, 2004). Wright also used them to estimate demographic parameters (Dobzhansky and Wright, 1941) and they have become a standard tool to 'estimate gene flow' or for merely descriptive studies of genetic population structure. Such studies are not always very convincing and may be questioned on statistical grounds. Two major objections are (1) the connection between such measures and the likelihood-based framework of statistics (e.g. Cox and Hinkley, 1974; Lehmann and Casella, 1998) is not obvious; and (2) although F_{ST} bears a simple relationship with the 'number of migrants' Nm in the infinite island model, it is not always clear how this would extend to more general models of population structure. Also, with the definition given above in terms of IIS, F_{ST} might be expected to depend on mutation processes at the loci considered, and how this affects estimation of dispersal parameters is not clear.

One of the main attractions of F-statistics may be their robustness to several factors. In an infinite island model, the ancestral lineages of two genes sampled within the same deme coalesce within this deme in a recent past with probability $\approx F_{\mathrm{ST}}$; with probability $\approx 1 - F_{\mathrm{ST}}$ the lineages separate (looking backward) in different demes as a result of immigration and will take a long time to coalesce. This implies that F_{ST} will depend mostly on recent events. Before considering the implications in detail, we will see how this argument can be generalized under isolation by distance.

We consider the probability $c_{j,t}$ that two genes coalesce at time t in the past. The j index corresponds to the type of pair of genes considered (e.g. $j = 2$ or 'w' for genes within demes and $j = 3$ or 'b' between demes as above). Identity by descent, here denoted $\dot{Q}$, has been defined as the probability that there has not been any mutation since the common

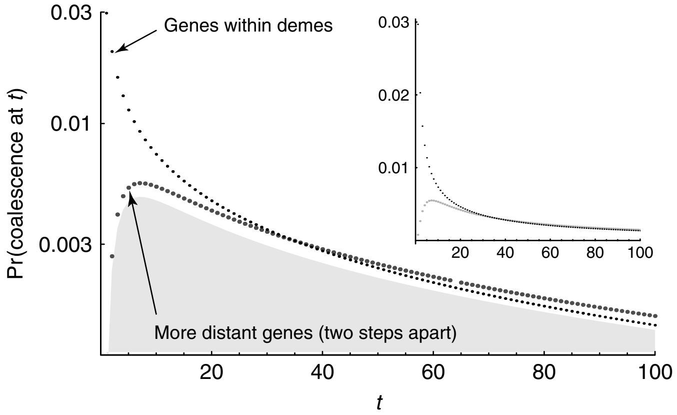

Figure 28.2 This figure compares the distributions of coalescence times in demes two steps apart (thick gray points) and within the same deme (thin black points). The inset shows the distributions on a linear y scale. The distribution for genes within demes is decomposed in two areas, the light gray one whose height is a constant times the height of the other distribution (hence it is shifted on the log scale), and the remainder (dark gray) which is the excess probability of coalescence in recent generations. The distribution were computed for 100 demes of $N = 10$ haploid individuals, with dispersal rate $m = 1/4$.

ancestor. Thus

$$\dot{Q}_j = \sum_{t=1}^{\infty} c_{j,t}(1-u)^{2t}. \tag{28.4}$$

(Malécot, 1975, (28.6); Slatkin, 1991). To understand the properties of F-statistics, we compare the distributions of coalescence times $c_{\mathrm{w},t}$ and $c_{\mathrm{b},t}$ of the pairs of genes that define these parameters. We can view the area covered by the probability distribution of coalescence time of the more related pair of genes as the sum of two 'probability areas', one part which is a smaller copy of the area covered by the probability distribution function of coalescence of less related genes, the other part being the remainder of the area for more related genes (Figure 28.2). This second part decreases faster than probabilities of coalescence (it is approximately $O(c_{\cdot,t}/t)$, Rousset, 2006), and is therefore concentrated on the recent past. As a first approximation, the value of the corresponding F-statistics is this excess probability of recent coalescence. Let us call ω the value of $1 - c_{\mathrm{w},t}/c_{\mathrm{b},t}$ for large t. This will also be the excess probability of recent coalescence (as can be deduced from the fact that both distributions must sum up to 1). Then it can be shown that

$$Q_{\mathrm{w}:k} \approx (1-\omega)Q_{\mathrm{b}:k} + \omega\pi_k \Rightarrow \frac{Q_{\mathrm{w}:k} - Q_{\mathrm{b}:k}}{\pi_k - Q_{\mathrm{b}:k}} \approx \omega \approx F, \tag{28.5}$$

where π_k is the expected frequency of allele k in the model considered. One may obtain this result by considering that with probability $1 - \omega$, the probability of identity of pairs of genes 'within' is the same as the probability of identity of genes 'between' (this corresponds to the proportional parts of the distributions of coalescence times), and with probability ω (the excess recent probability mass) the coalescence event has occurred recently in a common ancestor, of allelic type k with probability π_k.

The above result should not be overinterpreted. Expressions of the form probability of identity equals $(1 - F)p^2 + Fp$ for F independent of p are not generally valid unless p is the expectation π_k and the probability of identity is the process expectation (Rousset, 2002), although they also correctly describe the conditional probability of identity given p in the infinite island model, even in cases where p is a random variable.

The above logic will be valid as long as mutations can be neglected within the time span covered by the probability mass. This time span is shorter for higher migration rates m in the island model, or for high σ^2 relative to spatial distance in isolation by distance models, so in practice F-statistics are weakly dependent on mutation rates at small spatial scales. The same argument can also be used to show that F-statistics more quickly recover their stationary values than probabilities of identity under the same conditions after a single demographic perturbation, a fact noticed by several authors (e.g. Crow and Aoki, 1984; Slatkin, 1993; Pannell and Charlesworth, 1999). This kind of approximate independence is important for statistical applications since it makes F-statistics analyses at a small spatial scale interpretable despite the fact that past demographic history and mutation processes are generally not known. At a local scale, F_{ST} is also only weakly dependent on the total population size.

28.3.2 Likelihood Computations

With few exceptions, likelihood computation in population genetics are based on ‘coalescent’ arguments, i.e. they derive the probability of the sample from consideration of the sequence of states that relate the individuals in the sample to their common ancestor (e.g. Kingman, 1982; Hudson, 1990, **Chapter 25**, **Chapter 26**). This sequence of events may be the genealogy, G, of the sample that includes information about the coalescence time of ancestral lineages and about which lineages coalesce. In other cases it may be a ‘gene tree’ H, which takes into account the relative timing of coalescence and mutation events, as well as the nature of mutation events, i.e. the states before and after mutation, but which does not take into account the time between events, nor which lineage, among several with identical state, was involved in each event (see **Chapter 26**, Griffiths and Tavaré, 1995).

Coalescent arguments are used to estimate the likelihood by simulation using importance sampling algorithms. In one class of algorithms (see **Chapter 26**, Beerli and Felsenstein, 1999), the likelihood of the parameters $\mathcal{P}$ as a function of the data D may be written as

$$L(\mathcal{P}; D) = \sum_G \Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P}), \tag{28.6}$$

where the sum is over all possible genealogies G,

$$= \sum_G \Pr(D|G; \mathcal{P}) \frac{\Pr(G; \mathcal{P})}{f(G)} f(G), \tag{28.7}$$

for any distribution $f(G)$ such that $f(G) > 0$ when $\Pr(G; \mathcal{P}) > 0$

$$= \mathcal{E}\left[\Pr(D|G; \mathcal{P}) \frac{\Pr(G; \mathcal{P})}{f(G)}\right], \tag{28.8}$$

where $\mathcal{E}$ is an expectation over sample paths of a Markov chain with stationary distribution $f(G)$

$$\approx \frac{1}{s}\sum_{i=1}^{s} \Pr(D|G(i);\mathcal{P})\frac{\Pr(G(i);\mathcal{P})}{f(G(i))}, \tag{28.9}$$

where the sum is over the sample path of a Markov chain with stationary distribution $f(G)$.

In neutral models, given the genealogy G, the data only depend on the mutation process with parameters $\mathcal{N}$, while the genealogy itself does not depend on the mutation process but only on demographic parameters $\mathcal{D}$ (with $\mathcal{P} = (\mathcal{N}, \mathcal{D})$). Thus we may choose the importance sampling function

$$g \equiv \Pr(G|D;\mathcal{P}_0) = \frac{\Pr(G;\mathcal{D}_0)\Pr(D|G;\mathcal{N})}{L(\mathcal{P}_0;D)}, \tag{28.10}$$

for some value $\mathcal{D}_0$ of $\mathcal{D}$ and for $\mathcal{P}_0 = (\mathcal{N}, \mathcal{D}_0)$. Then from 28.8

$$L(\mathcal{P};D) = \mathcal{E}\left[L(\mathcal{P}_0;D)\frac{\Pr(G;\mathcal{D})\Pr(D|G;\mathcal{N})}{\Pr(G;\mathcal{D}_0)\Pr(D|G;\mathcal{N})}\right] = L(\mathcal{P}_0;D)\mathcal{E}\left[\frac{\Pr(G;\mathcal{D})}{\Pr(G;\mathcal{D}_0)}\right], \tag{28.11}$$

for any $\mathcal{N}$. We try to find the maximum likelihood estimate (MLE) of $\mathcal{P}$ or equivalently the maximum value of

$$\frac{L(\mathcal{P};D)}{L(\mathcal{P}_0;D)} = \mathcal{E}\left[\frac{\Pr(G;\mathcal{D})}{\Pr(G;\mathcal{D}_0)}\right], \tag{28.12}$$

which may be estimated by

$$\frac{L(\mathcal{P};D)}{L(\mathcal{P}_0;D)} \approx \frac{1}{s}\sum_{i=1}^{s}\frac{\Pr(G(i);\mathcal{D})}{\Pr(G(i);\mathcal{D}_0)}, \tag{28.13}$$

where the $G(i)$s are generated by a Markov ~~C~~hain with stationary distribution g. Thus an algorithm to find the maximum must define such a Markov ~~C~~hain (for parameters $\mathcal{P}_0$), and compute $\Pr(G(i);\mathcal{D})/\Pr(G(i);\mathcal{D}_0)$ for different $\mathcal{P}$ values and for this single Markov ~~C~~hain.

Beerli and Felsenstein (1999) have used the importance sampling function (28.10) to estimate the ratio (28.12). They define a Markov chain on genealogies G, and use the Metropolis-Hastings algorithm (Hastings, 1970) to ensure that the importance sampling function g is the stationary distribution of this chain. Their (28.14) shows that the transition probabilities of this chain are determined by the probabilities $\Pr(D|G;\mathcal{N})$, which for sequence data may be computed as previously described (e.g. Swofford and Olsen, 1990).

Griffiths and Tavaré (1994) have proposed a different class of algorithms. They derive recursions for the stationary probability $\Pr(D'|S')$ of a sample D' given sample size S' (here a vector and subsample sizes in different populations) over a time interval (typically the interval between two genealogical or mutation events). For any state D of ancestors in the previous time, $\Pr(D'|S')$ is the probability that a sample of size S' derives from a sample of size S, times the stationary probability of ancestral states $\Pr(D|S)$, times the forward probability that a sample D' derives from a sample D given the sample sizes:

$$\Pr(D'|S') = \sum_{D} \Pr(D'|D, S, S') \Pr(S|S') \Pr(D|S). \tag{28.14}$$

It is relatively straightforward to express the backward transition probabilities $\Pr(S|S')$ and the forward transition probabilities $\Pr(D'|D, S, S')$ in terms of model parameters $\mathcal{P}$ (further details are lengthy; see de Iorio and Griffiths, 2004b). An importance sampling algorithm is then derived by writing $\Pr(D'|D, S, S') \Pr(S|S')$ in the form $w(D'D)p(D|D')$ where the $p(D|D')$ define an absorbing Markov chain going backward over possible ancestral histories,

$$\Pr(D') = \sum_{D} w(D'D)p(D|D') \Pr(D). \tag{28.15}$$

Iterating this recursion until the ancestor of the whole sample shows that

$$\Pr(D) = \mathcal{E}_p\left[\prod_{D_i} w(D_i, D_{i-1})\right], \tag{28.16}$$

where the product is over successive states D_i of the ancestry of the sample. Then w is an importance sampling weight and p is a proposal distribution (compare 28.8).

Different choices of p, and of the implied w, are possible. Griffiths and Tavaré (1994) describe one such choice and show that the likelihood for different values of $\mathcal{P}$ may be obtained by running the Markov chain for only one value of $\mathcal{P}$. However, more recent works have aimed to optimize the choice of p for each parameter value analyzed so that the variance of $\prod_{D_i} w(D_i, D_{i-1})$ among runs of the Markov chain would be minimal. An optimal choice of the proposal distribution would be such that any realization of the Markov chain would give exactly the likelihood of the sample. This occurs when the proposal distribution has transition probabilities given by the reverse probabilities in the biological process considered, $\Pr(D|D') = \Pr(D'|D)\Pr(D)/\Pr(D')$ (see **Chapter 26**). In cases of interest these reverse probabilities cannot be computed from this formula since the aim is precisely to evaluate the probabilities $\Pr(D)$. Nevertheless, the variance of $\prod_{D_i} w(D_i, D_{i-1})$ should be low if good approximations for $\Pr(D)/\Pr(D')$ are used. de Iorio and Griffiths (2004a; 2004b) write such ratios as simple functions of the probabilities π that an additional gene sampled from a population is of a given allelic type, conditional on the result of previous sampling, and propose approximations $\hat{\pi}$ for them, from which approximations for $\Pr(D)/\Pr(D')$ and for the proposal distribution follow. Their approximation scheme method applies in principle for any stationary migration-matrix model and any Markov mutation model (for allele frequency data). The $\hat{\pi}$ are not given in closed form but as solutions of a system of $n_\mathrm{d} \times K$ linear equations for a model of n_d populations and K allelic types. By assuming independence between the mutation and genealogical processes, this can be reduced to a system of n_d equations holding for each of the different eigenvalues of the mutation matrix, a technique used by de Iorio *et al.* (2005) to analyze a two-demes model with stepwise mutation.

Despite substantial improvement over previous proposals, the computation times of the latter algorithm would remain prohibitively long in many practical applications. A method known as product of approximate conditional likelihoods (PAC-likelihood, Li and Stephens, 2003) has been proposed to derive heuristic approximations of the likelihood estimation. Here a sample of genotypes (g_k) is described as a sequential addition of genotypes, so that the likelihood of an ordered sample is the product of the probabilities π that an additional genotype is g_i given previously added genotypes were $g_1, \ldots, g_{i-1}$. Approximations are then considered for these conditional probabilities. This was originally applied to inference of recombination rate in a panmictic population. Using de Iorio and Griffiths's approximations $\hat{\pi}$ in one-locus models, it turns out to perform well under stepwise mutation, where the expectation of the PAC-likelihood statistic was indistinguishable from the likelihood, but computation was much faster (Cornuet and Beaumont, 2007). In a linear stepping stone model, one can find small differences between the expectation of the PAC-likelihood statistic and the likelihood, but estimation of model parameters based on PAC-likelihood is essentially equivalent to maximum likelihood (ML) estimation while again requiring far less computation than likelihood computation via the importance sampling algorithm (F.R., unpublished data).

28.4 INFERENCE UNDER THE DIFFERENT MODELS

In this section I review the implementation and application of the different methodologies in specific cases. Published genetic and demographic data from Gainj- and Kalam-speaking people of New Guinea (Wood *et al.*, 1985; Long *et al.*, 1986) will conveniently illustrate several conclusions.

28.4.1 Migration-Matrix Models

For any migration matrix at stochastic equilibrium, the distribution of frequencies p_{ki} of allele k in each deme i follows some probability distribution with (parametric) covariances which can be written as

$$\mathrm{E}[(p_{ki} - \pi_k)(p_{ki'} - \pi_k)] = Q_{ii':k} - \pi_k^2, \qquad (28.17)$$

where $Q_{ii':k}$ is the expected frequency of pairs of genes in demes i, i' that are of allelic type k. For mutation models assuming identical mutation rates between K alleles, this is also $(Q_{ii'} - \sum_{k=1}^{K} \pi_k^2)/K$, where $Q_{ii'}$ is the probability of IIS of pairs of genes in demes i, i'. Probabilities of IIS $Q_{ii':k}$ or $Q_{ii'}$ may be derived from the probabilities of IBD for various mutation models (Markov chain models, or stepwise mutation models; see e.g. Tachida, 1985; Rousset, 2004). For any migration matrix model, probabilities of IBD within and among demes can be computed as solutions of a linear system of equations (see e.g. Nagylaki, 1982 or Rousset, 1999a; 2004 for details and examples). In principle, the demographic parameters can be estimated by inverting such relationships. This approach has been taken seriously only in a few cases, in particular the island and isolation-by-distance model, as detailed below, and does not generate likelihood expressions in a straightforward way, although some heuristic likelihood formulas have been proposed (Tufto *et al.*, 1996) by using Gaussian approximations for the distribution of allele frequencies, with the covariances given above.

28.4.2 Island Model

In the island model, one has the well-known approximation $F_{\mathrm{ST}} \approx 1/(1 + 4Nm)$ (with $2N$ genes per deme, Wright, 1969). This has led to the usage of computing F_{ST}s and expressing the results in terms of ‘estimates of Nm’, i.e. in terms of $(1/\hat{F} - 1)/4$. This usage is often problematic. The worst sin is to estimate an F_{ST} between a pair of samples far apart, to translate it into a nonzero ‘Nm’, and to conclude that the populations must have exchanged migrants in the recent past. In the context of the island model, an F_{ST} between a pair of subpopulations is not a function of the number of migrants exchanged specifically between these two subpopulations. More generally, in many models of population structure, it is expected that subpopulations that never exchange migrants will have nonzero ‘Nm’ values. It may be seen from the above definition of F_{ST} that its maximum value is the probability of identity within demes Q_2 (when $Q_3 = 0$), which results in a minimum possible value of ‘Nm’ of $(1/Q_2 - 1)/4$ which may well be > 1 even for demes that never directly exchange migrants. Likewise, the practice of equating $Nm > 1$ to panmixia and $Nm < 1$ to divergence is not useful.

Likelihood functions for allele frequency data may be derived relatively easily by diffusion techniques for the infinite island model (see (28B.5)) and can in principle be recovered by coalescent arguments (Balding and Nichols, 1994). These sampling formulas allow analytical insight, and may be used to define estimators of Nm as well as to discuss efficient estimation of F_{ST} by moment methods (See Appendix B). Kitada *et al.* (2000) have implemented likelihood estimation under this model. Approximation have also been considered such as the ‘pseudo maximum likelihood estimator’ (PMLE, Rannala and Hartigan, 1996) of the number of migrants in an island model (see (28B.6)). These authors found that this estimator of Nm generally (though not always) had lower mean square error than the moment estimator $(1/\hat{F} - 1)/4$, depending on sampling scheme and Nm values. The MLE is also biased when the number of sampled populations is small and some corrections have been proposed (Kitakado *et al.*, 2006). For the New Guinea population, application of pseudo maximum likelihood (PML) estimation yields an estimate of 10.2 migrants per generation. This is one-fourth of the average value, 41.87, that can be computed from maternal and paternal dispersal rates and total subpopulations sizes (Tables 1–3 in Wood *et al.* (1985)), but it is closer to one third if we take ‘effective size’ considerations into account following Storz *et al.* (2001).

In comparison with the simulation methods for likelihood computation, it should be noted that no mutation model has to be considered here. Wright’s formula (28B.1) is an approximation for low mutation, common to the different ‘Markov chain’ models of mutation, and so is the likelihood formula (28B.5). In this respect it is analogous to the methods based on F-statistics.

28.4.3 Isolation by Distance

Here analytical insight is available only for the moment methods. The results reviewed here are not tied to a Gaussian model of dispersal. We consider

$$a(\mathbf{r}) \equiv \frac{Q_{\mathbf{0}} - Q_{\mathbf{r}}}{1 - Q_{\mathbf{0}}}, \tag{28.18}$$

which is $F_{\text{ST}}/(1 - F_{\text{ST}})$ at (vectorial) distance $\mathbf{r}$. Approximations for F_{ST} immediately follow from those for $a(\mathbf{r})$. I will use a dot on a or Q to emphasize that the results given hold strictly only for IBD, but the differences with IIS do not affect the main practical conclusions drawn below (Rousset, 1997).

Two cases are usually considered, the one-dimensional model for populations in a linear habitat, and the two-dimensional model. In one dimension, at distance r,

$$\dot{a}(r) \approx \frac{A_1}{4N\sigma} + \frac{1 - e^{\frac{-(2u)^{1/2}r}{\sigma}}}{4N\sigma(2u)^{1/2}} \overset{r\text{small}}{\approx} \frac{A_1}{4N\sigma} + \frac{r}{4N\sigma^2} \approx \frac{A_1}{4D\sigma} + \frac{r}{4D\sigma^2}, \qquad (28.19)$$

where A_1 is a constant determined by the dispersal distribution, but not by N nor u. Its definition is given by Sawyer(1977, eq. 2.4).

In two dimensions, for genes at Euclidian distance r,

$$\begin{aligned}\dot{a}(\mathbf{r}) &\approx \frac{-\ln((2u)^{1/2}) - K_0((2u)^{1/2}r/\sigma) + 2\pi A_2}{4N\pi\sigma^2} \overset{r\text{small}}{\approx} \frac{\ln(r/\sigma) - 0.116 + 2\pi A_2}{4N\pi\sigma^2} \\ &\approx \frac{\ln(r/\sigma) - 0.116 + 2\pi A'_2}{4D\pi\sigma^2}, \qquad (28.20)\end{aligned}$$

where K_0 is the modified Bessel function of second kind and zero order (e.g. Abramovitz and Stegun, 1972), and A_2 is of the same nature as A_1 above. Its definition is given by Sawyer, (1977, eq. 3.4); see also Rousset (1997 eq. A11).

In the last two equations the first expression is given for σ measured in the length unit of the model (i.e. one interdeme distance on the lattice), the second is the small distance/low mutation limit of the first, and the third is the second for any length unit. They are in terms of population density D per length or surface unit, and the A'_2 constant depends on the length unit.

In the same equations, the second approximation shows a linear relationship between $\dot{a}(r)$ and geographical distance in one dimension, and between $\dot{a}(r)$ and the logarithm of geographical distance in two dimensions. In both cases, the slope of this relationship is a function of $D\sigma^2$.

These different expressions emphasize two points. First, differentiation is a function of the A constants, which are not simple functions of σ^2 but also of other features of the dispersal distribution. In fact, when the total migration rate is low, the differentiation between adjacent subpopulations is close to that expected under an island model with the same total number of migrants. This confirms that σ^2 is not the only relevant parameter of the dispersal distribution. Second, the value of A'_2 depends on the spatial unit chosen to measure σ and D. A method of inference from A'_2 values that would not take into account the discrepancy between the length unit used and the idealized interdeme distance would therefore be internally incoherent.

The above approximations allow a relatively simple description of the expected differentiation in these models as well as relatively simple estimation of $D\sigma^2$ from genetic data. Estimates of $a(r)$ at different distances may be obtained in some cases as estimates of $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}})$ for pairs of samples, and simply regressed to spatial distance (Rousset, 1997, as implemented in GENEPOP, Raymond and Rousset, 1995). An estimate of $1/(D\sigma^2)$ may be deduced from the slope of the regression. Two early applications of this method yielded estimates about twice the demographic estimate (Rousset, 1997). For the New Guinea population, the regression equation $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}}) \hat{=} 0.0191 + 0.0047 \ln(\text{distance in km.})$ provides an estimate of $D\sigma^2$ which is about twice the demographic estimate (after application of effective density correction following Storz *et al.* (2001), and after correction of clerical errors affecting the reported σ^2 of females and males in Rousset (1997), which should be 3.1 and 0.76 km^2 respectively).

When the migration rate is low, an estimate of the number of immigrants per generation may also be computed by the '$(1/\hat{F}-1)/4$' method, taking the value of the estimated regression equation at the distance between the closest subpopulations as an estimate of $\hat{F}/(1-\hat{F})$. The estimate of number of migrants in the New Guinea population is then 11.5, close to the pseudo maximum likelihood estimate (10.2, see above). This result illustrates the approximate convergence of estimates by different methods and under different dispersal models to Wright's classic result, even though the dispersal rate in this study is not precisely low (its average value being 0.43 from demographic data).

The regression of $F_{\text{ST}}/(1-F_{\text{ST}})$ to distance is not always applicable, particularly when there are no recognizable demes of several individuals, as for 'continuous' populations. A variant based on the comparison of pairs of individuals has been designed to address this problem (Rousset, 2000). Simulations have shown that this method performs reasonably well when σ is small (a few times interindividual distance at most) and when most individuals are sampled within an area of about $20\sigma \times 20\sigma$ (Leblois *et al.*, 2003,2004). For higher dispersal, the variant considered by Vekemans and Hardy (2004) provides more accurate upper confidence bounds for $D\sigma^2$ (Watts *et al.*, 2007). Several comparisons have found agreement within a factor of two with independently derived demographic estimates (Rousset, 2000; Sumner *et al.*, 2001; Winters and Waser, 2003; Fenster *et al.*, 2003; Broquet *et al.*, 2006; Watts *et al.*, 2007). Whether this is considered an important discrepancy or not will depend on the accuracy expected from such analyses, but this is certainly much better than usually reported (see e.g. Slatkin, 1994; Koenig *et al.*, 1996). They actually go against an earlier long stream of reported discrepancies between genetic and demographic estimates, which needs explaining.

Part of the discrepancies hinge on misunderstandings of the models. For example, Wright assumed that the value of F-statistics under isolation by distance (Wright, 1946) was determined by the 'neighborhood size'. The value of this parameter would be ~~$2D$~~σ^2 under the assumption of two-dimensional Gaussian dispersal and its more general definition would be a function of 'the chance that two uniting gametes came from the same individual' (Wright, 1946). A third common 'definition' found in the literature is that the 'neighborhood size' would be the size of a subpopulation that would behave as a panmictic unit. It is not clear in which respect the subpopulation would behave as a panmictic unit nor whether there is a subpopulation that behaves as a panmictic unit in some useful sense. In any case Wright's measures do not correctly predict the value of unambiguously defined parameters in unambiguously defined models. In the analysis of Malécot's model, neither $D\sigma^2$ (because of the important A_2 term), nor the more generally defined neighborhood, determine the differentiation alone (Rousset, 1997). In one dimension, Wright proposed that $D\sigma$ was the important parameter, but the above results show that $D\sigma^2$ is important. One must give up the idea that $D\sigma^2$ equals neighborhood equals a number of individuals. In one dimension, $D\sigma^2$ scales as number of individuals times a length, not as a number of individuals, since density is a number of individuals per unit length.

The neighborhood concept was an attempt to account for different families of dispersal distributions. On the other hand, it has recurrently been assumed that differentiation is essentially a function of σ^2 and not of other features of the dispersal distribution. If so, it would be easy to seemingly improve on the regression method by considering only a family of dispersal distributions with a single parameter, completely determined by σ^2, for example a discretized Gaussian. In this case, F_{ST} or $a(r)$ values, not simply their increase with distance, would contain information about σ^2. But such improvements would not be robust to misspecification of the dispersal distribution.

To explain reported discrepancies, it has often been argued that genetic patterns are highly sensitive to long-distance dispersal, which occurrence is easily missed in demographic studies. While some genetic patterns are indeed affected by long-distance dispersal (e.g. Austerlitz *et al.*, 2000), this is much less so for the patterns considered in the regression analyses, and this contributes to their concordance with demographic estimates. If a fraction m of immigrants come from an infinite distance (so that the 'true' σ^2 is infinite and does not predict any local pattern of differentiation), such migrants will be unrelated to their neighbors, and these migration events are analogous to mutation events. Hence we can deduce the effect of such immigrants from the effect of mutation, e.g.

$$\dot{a}(\mathbf{r}) \approx \frac{-\ln(\sqrt{2m}) - K_0(\sqrt{2m}r/\hat{\sigma})}{\cancel{2}D\pi\hat{\sigma}^2} + \text{constant}, \qquad (28.21)$$

where $\hat{\sigma}$ is the parameter of the dispersal distribution for the fraction $1 - m$ of locally dispersing individuals. This result implies that there is an approximately linear increase of differentiation, determined by $D\hat{\sigma}^2$, roughly up to distance $0.56\hat{\sigma}/\sqrt{2m}$ (from Figure 3 in Rousset (1997)). For example if we ignore a 1 % (respectively, 0.1 %) tail of the distribution of dispersal distance in a demographic study which estimates $\hat{\sigma} = 10$ distance units, the prediction of increase of differentiation with distance will reach 20 % error at $0.56\hat{\sigma}/\sqrt{2m} = 39.6$ (respectively, 125) distance units. This is a wide overestimate of the error for any data set spread over such a distance, but more accurate predictors of bias will depend on the distribution of spatial distances in the sample.

Naive application of testing methodology has been another factor contributing to confusion. The absence of a pattern of isolation by distance (null slope of the regression, $D\sigma^2$ infinite) may be tested by the exact permutation procedure known as the *Mantel test* (Mantel, 1967; see Rousset and Raymond, 1997, for a simple description). In practice, nonsignificant test results have often been interpreted as evidence that dispersal is not localized. However, the Mantel test has often been applied in conditions of low power. In many populations with localized dispersal, the value of $D\sigma^2$ will be large, and thus expected patterns of isolation by distance (increase of differentiation with distance) will be weak (particularly in two-dimensional habitats), even though differentiation will be inferred by classical tests for differentiation.

Finally, variation in expected gene diversity due to spatial heterogeneity of demographic parameters may result in larger variation in expected differentiation than that due to isolation by distance. For example, if several demes with very small deme size and restricted dispersal are clustered in space, they will have low expected gene diversity and will show a larger differentiation between them than with more distant demes with higher expected diversity, and the above methods will obviously fail unless this heterogeneity is taken into account (Rousset, 1999b).

28.4.4 Likelihood Inferences

Maximum likelihood methods have not been developed and tested to a comparable level. The migration matrix models have been implemented for allele frequency and sequence data in the software MIGRATE. A more restricted set of models based on the 'isolation-with-migration' scenario has been implemented in the software IM (Nielsen and Wakeley, 2001; Hey, 2005). In the coalescent algorithms as in previous inferences based on per-locus information from samples taken at one time, it is not possible to estimate the deme sizes, mutation rates, and immigration rates separately: only products of deme size with migration probability and mutation probability, and (as a consequence) the ratio of migration and mutation probabilities, can be estimated. In the latest version of IM it is possible to analyze the divergence between two populations in terms of their deme sizes scaled relative to mutation rate ($N_1\mu$ and $N_2\mu$), of immigration rates in each of them m_1 and m_2, either scaled relative to deme size (e.g. N_1m_1) or relative to mutation rate (e.g. m_1/μ), of the scaled size of the ancestral population, and of relative fractions of this ancestral which contributed to the two descendant populations.

In a two-populations setting, one study has compared ML estimates with demographic estimates and with some other genetic estimation methods (Wilson *et al.*, 2004). Both ML and the two-locus method of Vitalis and Couvet (2001) produced estimates roughly in agreement with demographic data, while F_{ST} did not. It appears difficult to estimate all parameters of a four-demes migration-matrix model (Beerli, 2006). Likewise, it has not been possible to choose among different scenarios of colonization of the Americas using the IM software (Hey, 2005). The effect of unsampled demes on estimation of dispersal between two sampled demes has been investigated by Beerli (2004), for sequence data (100 000 bp per individual). As expected, the biases increase with immigration from the unsampled population(s), but it was found that estimates of immigration rate between the two populations were hardly affected by an equal total immigration rate from unsampled population(s) (estimates of scaled populations sizes were more affected). Abdo *et al.* (2004) argued that confidence intervals given by MIGRATE are not accurate, a fact attributed to too short run times of the algorithm (Beerli, 2006).

The above simulation scenarios are rather distinct from those considered in the previous section, and it would be hard to perform ML analyses of the New Guinea data using current software. With MIGRATE for example, estimation of a full migration matrix from the New Guinea data has been attempted (R. Leblois and F.R., unpublished results), yielding larger estimates of dispersal than inferred by the regression method described above and from demographic data. Attempts to estimate fewer parameters could be more successful. In addition, one can question the convergence of the Markov chain algorithm, a problem which remains with no easy solution (e.g. Brooks and Gelman, 1998). In this respect, the importance sampling algorithms of Griffiths and collaborators are more convenient as estimates are derived from independent runs of an absorbing Markov chain (28.15), so traditional techniques based on independent variables apply.

28.5 SEPARATION OF TIMESCALES

The properties of F-statistics illustrate the more general idea of separation of timescales, in which some events occur at a much faster rate than others. For example, in an island model, when the number of demes n_{d} increases indefinitely, the rate of coalescence of ancestral lineages of genes sampled in different demes decreases as $1/n_{\mathrm{d}}$, while for genes sampled within the same deme, the probability of coalescence of ancestral lineages in some recent generation is nonvanishing in this limit. Thus the events in the genealogy of a sample can be described as a sum of two processes, a fast process by which lineages either coalesce within the demes they are sampled or separate in distinct demes, at rate $O(1)$ as

$n_\mathrm{d} \to \infty$, and a slow process by which genes in different demes coalesce at rate $O(1/n_\mathrm{d})$ as $n_\mathrm{d} \to \infty$ (e.g. Wakeley and Aliacar, 2001). This slow process is a rescaled version of the well-studied coalescence process for an unstructured population (the n-coalescent, Kingman, 1982; **Chapter 25**).

Under a separation of timescales, the likelihood can be expressed in terms of distinct terms for fast and slow processes (see Nordborg, 1997 for a population with selfing), and in the above example, analytical results or simulation techniques for the n-coalescent can be used as soon as ancestral lineages have separated in different demes in a simulation of the ancestry of a sample. The traditional formula giving probability of identity conditional on allele frequency as $(1 - F)p^2 + Fp$ also results from such a separation of timescales, provided that allele frequency does not change in the total population until the completion of the fast process.

However, convergence to the n-coalescent may hold under sometimes restrictive conditions. No convergence to the n-coalescent has been found in one-dimensional models of isolation by distance (Cox, 1989; Wilkins, 2004). In two-dimensional models on a lattice of size $L \times L$, the genealogy of genes sampled far enough (at distance $O(L)$) from each other converges to that of an unstructured population (the n-coalescent) with coalescence events occurring at rate $O(1/[L^2 \ln(L)])$ as $L \to \infty$ (Cox, 1989; Zähle *et al.*, 2005). A separation of timescales may hold if the genealogy of genes sampled closer in space compounds a process of ‘fast’ coalescence (in less than $O[L^2 \ln(L)]$ generations) and the n-coalescent. The closest results are those of Zähle *et al.* according to which genes at distance $O(L^\beta)$ for $0 < \beta \leq 1$ either coalesce in less than $L^2/2$ generations with probability $1 - \beta$, or follow the scaled n-coalescent with probability β. This differs qualitatively from the island model where the fast process becomes negligible in finite time, so defining a finite time span for a fast process from these results seems less than straightforward. Instead, they suggest applying n-coalescent approximations in a backward simulation algorithm when ancestral lineages are distant enough relative to the total size of the lattice, although the minimal distance to consider is itself not obvious since all results stand for $\beta > 0$ only, i.e. for spatial separation of genes increasing with L. As shown in Figure 28.2, a separation of timescales holds for fixed distance and fixed $N\sigma^2$, with the fast process vanishing in finite time, but the slower process is not an n-coalescent.

Some moment methods have attempted to extract more information from the data by taking into account allele size (for microsatellites) or DNA sequence divergence. However, in an island model, most of the information about Nm is in whether genes sampled within a deme have their most recent common ancestor within this deme (in which case they are identical, unless mutation rates are higher than migration rates). Otherwise, the ancestral lineages separate in different demes, and in this case the allelic divergence may contain little useful information about dispersal rates. This may explain why moment methods based on microsatellite allele size often yield estimates of demographic parameters with higher variance and mean square error than estimates derived from allelic identity statistics (Gaggiotti *et al.*, 1999; Balloux and Goudet, 2002; Leblois *et al.*, 2003) despite their lower asymptotic bias (Slatkin, 1995), except when mutation rates are larger than migration rates (Balloux *et al.*, 2000) or when genetic correlations are not determined by a distinctly fast process (Tsitrone *et al.*, 2001). A key issue in evaluating the performance of likelihood methods will be the extent to which they specifically capture the information from fast processes and how much better they are than pair-wise identity methods in this respect.

28.6 OTHER METHODS

As likelihood computations are often difficult, simulation methods based on other summary statistics have been developed. In essence, the likelihood of the data is substituted with the probability of observing in simulations values of a summary statistic S sufficiently close to its value observed in the data. These methods are reviewed by Beaumont (see **Chapter 30**) and have been applied to some subdivided population scenarios (Estoup *et al.*, 2004; Hamilton *et al.*, 2005). There is a huge collection of other methods of analysis of spatial patterns of genetic variation in the literature. Here again a small selection is presented on the basis of their impact in the field or of their perceived practical validity.

28.6.1 Assignment and Clustering

Of particular interest are assignment methods, which aim to assign individuals to their subpopulations of origin. For example, in an ecological perspective it may be useful to know whether immigrants differ from residents in some aspects of their behavior, and therefore to individually identify immigrants; further it might be of interest to identify their habitat of origin. Sometimes there will be independent information about the potential source populations. A more challenging problem (on which this section will focus) is to estimate dispersal rates by inferring the number of subpopulations and then assigning individuals to them (see **Chapter 30** for other aspects of these methods). In this perspective, the traditional problem of clustering becomes part of the assignment task.

Assignment methods stem from the idea that individuals are more likely to originate from populations with higher frequencies of the alleles they possess. Thus, considering an haploid organism for simplicity, for each individual one can compute a statistic such as

$$\prod_{k=1}^{K} \tilde{q}_{ki}^{x_k}, \tag{28.22}$$

where $\tilde{q}_{ki}$ is the observed frequency for allele k in sample i (possibly with some correction such as excluding the focal individual itself), and $x_k = 1$ if the individual bears allele k and $x_k = 0$ otherwise. This is usually viewed as a likelihood statistic (e.g. Paetkau *et al.*, 1995). If each locus is considered independent of the others, multilocus statistics are the product of single locus statistics, and the individual is assigned to the sample i that maximizes the multilocus statistic.

It can be expected that, given some differentiation between different subpopulations, this method will preferentially assign an individual to its original subpopulation. It is also expected that such assignments will be more accurate when differentiation is higher. More generally, if the individuals are correctly assigned to their subpopulations of origin, then we could estimate from the same data and the same likelihood formulas the dispersal rates between each of them in the last round of dispersal, independently of past dispersal rates. Conversely if we cannot consistently estimate the dispersal rates in this way, this implies that there is no way to consistently assign individuals to their populations of origin. So we consider the question whether we can estimate the dispersal rate in order to address the question whether immigrants can be assigned to their population of origin.

Consider the allele frequencies in the subpopulations before the last round of dispersal (p_{ki} for allele k in subpopulation i), and the dispersal rate for this last round of dispersal

($m_{ii'}$ from subpopulation i' to i). For each locus and each individual in sample i the likelihood is approximately $\prod_{k=1}^{K} q_{ki}^{x_{ki}}$ where $q_{ki} = \sum_{i'} m_{ii'} p_{ki'}$ is the frequency of allele k in subpopulation i after the last round of dispersal. For the total sample from n_d demes the likelihood is therefore proportional to

$$\prod_{i=1}^{n_\mathrm{d}} \prod_{k=1}^{K} q_{ki}^{n_{ki}}. \tag{28.23}$$

Since one can always explain the data by some model assuming that there was no dispersal in the last round, there is not enough information in the data to separately estimate the dispersal rates and the p_{ki}s, at least when each locus is considered independent of the others as done previously. This implies that it is not possible to consistently assign individuals to their populations of origin without introducing additional assumptions or additional knowledge.

The most obvious assumption is to assume no linkage disequilibrium within populations before dispersal, which is a good approximation for many organisms (though not highly selfing ones). In principle, one- and two-locus measures of genetic association contain information which allows to estimate separately dispersal rates and subpopulation sizes. A numerical method has been developed to use such information (Vitalis and Couvet, 2001). Assignment methods may be viewed as using multilocus associations, in that their current formulation relies on assuming no within-population disequilibria. Additional assumptions have been made, for example specifying a prior probability model for the distribution of allele frequencies, often a Dirichlet distribution as per an island model (see (28B.1); Rannala and Mountain, 1997; Falush *et al.*, 2003; Wilson and Rannala, 2003; Corander *et al.*, 2003). In this respect, although an asserted aim of such methods is to infer migration rates without the many assumptions of other methods, the stationary island model is still lurking in the background.

How do these methods perform in practice? Cornuet *et al.* (1999) reported $>75\,\%$ correct assignment probabilities by Rannala and Mountain's method for populations diverged since several hundred generations. Évanno *et al.* (2005) found that the 'most likely' number of populations reported by the program STRUCTURE was a biased estimator of the actual number of populations. Waples and Gaggiotti (2006) reported that this program correctly identified the number of populations when the number of immigrants per population was less than 5, mutation rates were high, and for large sample size (20 loci genotype in 50 individuals from each population), but quickly degraded in other conditions. The latter study also reported similar problems with the methods implemented in the programs BAPS (Corander *et al.*, 2003) and IMMANC (Rannala and Mountain, 1997), and found that a method based on traditional contingency tests of spatial structure performed better than these different methods in identifying the number of populations. In a comparison with mark-recapture data, Berry *et al.* (2004) found good performance in estimating dispersal. In this study, the populations were known in advance and at most 4, the number of immigrants was known from mark-recapture data to be small, and the latter information was somehow used as prior information in STRUCTURE.

Most of the recently formulated methods have been presented as 'Bayesian', a label which in practice covers various compromises between subjective Bayesian (e.g. Lindley, 1990) and frequentist (Neyman, 1977) statistics, to the point of being uninformative. However, numerous accounts of supposed differences between Bayesian and frequentist methods have drawn attention away from the real issue, which is the criterion by which to measure the performance of statistical methods. It is always possible to generate 'better than previous' methods by changing the measure of average performance. In the context of estimation of dispersal rates, the results of Beerli (2006) illustrate this, where performance averaged over a prior distribution for model parameters was better than that of maximum likelihood ignoring the prior distribution, a predictable result from textbook statistical theory (Cox and Hinkley, 1974, Chapter 11; Lehmann and Casella, 1998, Chapter 4). Likewise, the performance of an assignment method defined in terms of priors over allele frequencies may be evaluated in terms of its performance for any given allele frequency, or of averaged performance over the distribution of allele frequencies. The context of scientific inference should determine the appropriate measure of performance, but the averaged measure can balance misleading assignment inferences in some species with more efficient inferences in some other species, depending on the spectrum of allele frequencies in different species. One may need to know when this is the case, and comparing results for two choices of the prior distribution is not enough in this respect.

As seen above, a more prosaic problem with assignment methods is that it is difficult from available information to give simple bounds on the frequency of erroneous inferences, even averaged over prior distributions.

28.6.2 Inferences from Clines

An important class of models of spatial structure for selected genes are the cline models. Clines arise in two contexts: selection for two distinct alleles or genotypes in two adjacent habitats exchanging migrants, or selection against hybrid genotypes between two taxa ('tension zones'). Theoretical models predict the shape of clines, notably the steepness in the cline center, as a function of the σ parameter defined above and of one or several selection coefficients (for tension zones: Bazykin, 1969; Barton, 1979; for selection variable in space: Nagylaki, 1975). Additional information on dispersal and selection is given by linkage disequilibria between loci. Such methods, therefore, depend on assumptions specific to each case study (e.g. external information on recombination rates and epistasis between loci).

A typical expression for the shape of a cline, i.e. for allele frequency p at distance $x - x_0$ from the center of the cline, is $p = \left(1 + \exp((x_0 - x)(2s)^{1/2}/\sigma)\right)^{-1}$ (Bazykin, 1969; Barton, 1979). In the center of the cline, it holds approximately for different models of selection (Barton and Gale, 1993) and is relatively insensitive to the shape of the dispersal distribution, as a small rate of unaccounted long-distance immigration has little effect on the shape of the center of the cline (Rousset, 2001). Inferences about s and σ separately are possible by considering multilocus clines and by taking linkage disequilibria into account. Then, the expected shapes of clines must be computed numerically. Drift is neglected relative to selection: Allele frequencies in the different subpopulations are fixed values, functions of the parameters defining the demography and the selection regime, not random variables as in the neutral models.

The likelihood function is then given by multinomial sampling in each subpopulation. Thus, the statistical model is conceptually straightforward and relatively easily tailored to the specific demography and selection regime of different organisms, at least when only a few loci subject to selection need be considered and when the expected frequencies of each genotype are directly computed using recursion equations for specific values of the parameters to be estimated (e.g. Lenormand *et al.*, 1999). Variants of the Metropolis algorithm (Metropolis *et al.*, 1953) including simulated annealing (Kirkpatrick *et al.*, 1983) have been used to find MLEs (a software ANALYSE is distributed by Barton and S. J. Baird, `http://helios.bto.ed.ac.uk/evolgen/Mac/Analyse/`). Sites *et al.* (1995) reported an agreement with demographic estimates of σ similar to that of isolation-by-distance analyses reported above.

28.7 INTEGRATING STATISTICAL TECHNIQUES INTO THE ANALYSIS OF BIOLOGICAL PROCESSES

Several of the methods reviewed in this chapter have been both widely used and widely criticized. Much of the criticisms rest on the difficulty of formulating precise quantitative models of population genetic processes. Theoretical studies of robustness are important, but may themselves overlook factors that turn out to be important in natural populations. For these reasons, this review has emphasized comparisons with demographic estimates. Independent demographic estimates may have their own problems, but it will be hard to detect misapplications of the genetic inferences if no such comparison is made. This last section discusses some of these problems and partial solutions to them.

We have seen that many discrepancies between 'models and data' inferred from empirical studies using F-statistics derive from various misunderstandings of the models. The more successful studies, in terms of comparisons with independent estimates, were conducted at a small spatial scale (between 1 and 20 σ). This is somehow unavoidable because of the need for good demographic data in these comparisons, but one may expect larger discrepancies over larger spatial scales, for many reasons: spatial variation of demographic parameters should be taken into account; the effects on genetic differentiation of some demographic events such as range expansions will be more likely to be observed (Slatkin, 1993); mutation will have measurable consequences (e.g. Estoup *et al.*, 1998); and selection variable in space may also affect the markers.

A frequently raised concern is the possible nonneutrality of the markers used. On the positive side a number of authors have realized that divergent selection would increase levels of differentiation between different subpopulations. Thus potentially selected loci may be detected in a first step by 'weak' statistical approaches such as classifying loci as showing structure or not by conventional significance tests (see Kreitman, 2000 for tests of selection at a molecular level not specifically using geographical information). Formal statistical evidence for selection may be obtained by other experiments in a second step. This approach has proven efficient (e.g. Feder *et al.*, 1997). Lewontin and Krakauer (1973) proposed a quantitative test of selection from the heterogeneity of F_{ST} estimates. This procedure was inadequate in several ways, but there have been more recent attempts to refine the detection of candidate selected loci (e.g. Beaumont and Nichols, 1996; Vitalis *et al.*, 2001; Beaumont and Balding, 2004).

Another often expressed criticism of the models and analyses reviewed above is that they assume equilibrium, while the populations are often not at demographic equilibrium, i.e. population sizes and migration rates fluctuate in time. If so, it is not clear what is estimated by such techniques: the present demography, an average over 'recent' times, a 'long-term' average, or none of them? If the fluctuations can be described as a fast process, they may be described by some effective size correction. In other cases, such as the range expansion of a species, this cannot be so, and either modeling the demographic expansion, or using methods insensitive to it, are the only coherent alternatives. To understand this, it suffices to note that spatial patterns of pairs of genes approach equilibrium faster the smaller the spatial scale considered. Therefore, if the effect of a demographic event was captured by some effective size correction, the effective size would differ at different scales. Hence this effect cannot be described by a single effective size characterizing the total population.

One way to avoid some assumptions of equilibrium is to analyze sequential samples. All the above methods assume that samples have been taken at one point in time, yet sometimes temporal information is available. See e.g. Robledo-Arnuncio *et al.* (2006) for inferences of dispersal distributions from mother−offspring data, Wang and Whitlock (2003) for estimation of immigration rate and deme size from samples over wider time spans, and Ewing and Rodrigo (2006) for implementation of Markov chain algorithms for inference of changes in demographic parameters from sequential samples.

The idea of estimating dispersal parameters such as σ is also open to difficulties. By allowing an arbitrarily small fraction of immigrants to come from far enough, it is easy to design cases where the theoretical σ value would be arbitrarily large, and where long-distance immigrants would have arbitrarily small effect on the likelihood of samples. The question that one must address prior to statistical analysis is how important such long-distance immigrants are for population processes. For example, the speed of range expansions is known to be affected by the most extreme long-distance migrants in a way generally not characterized by the σ parameter (Mollison, 1977; Clark *et al.*, 2001), so if one is interested in characterizing such processes, not only it will be difficult to estimate σ but this may be irrelevant. One the other hand, some processes of local adaptation (as may lead to allele frequency clines, for example) are not very sensitive to long-distance migrants, and then approximations ignoring them are not only adequate but required to formulate good statistical inferences.

Current methods of estimation still have low range of applications, low efficiency, or both. In principle, this can be improved by the development of likelihood methods, yet this leaves room for different methodologies, and it is unclear how far research practices will be improved. One common theme is that genealogical structure is affected by events occurring at different timescales, and that inferences based on models of the faster processes could be relatively independent to uncontrolled historical processes, and therefore perhaps more reliable. It is not yet clear how much complexity we can add in the models, for given data, nor where will be the limit between reliable and unreliable inference; further, the answer will likely differ whether sequence data or allele frequency data are considered.

Acknowledgments

I thank D. Balding, M. Beaumont, M. Lascoux, R. Leblois, R. Vitalis, and M. Nordborg for comments or discussion related to this or previous versions of this chapter.

Related Chapters

Chapter 25; **Chapter 26**; **Chapter 29**; and **Chapter 30**.

REFERENCES

Abdo, Z., Crandall, K.A. and Joyce, P. (2004). Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology* **13**, 837–851.

Abramovitz, M. and Stegun, I.A. (eds) (1972). *Handbook of Mathematical Functions*. Dover Publications, New York.

Austerlitz, F., Mariette, S., Machon, N., Gouyon, P.-H. and Godelle, B. (2000). Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics* **154**, 1309–1321.

Balding, D.J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology* **63**, 221–230.

Balding, D.J. and Nichols, R.A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.

Balloux, F., Brünner, H., Lugon-Moulin, N., Hausser, J. and Goudet, J. (2000). Microsatellites can be misleading: an empirical and simulation study. *Evolution* **54**, 1414–1422.

Balloux, F. and Goudet, J. (2002). Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology* **11**, 771–783.

Barton, N.H. (1979). The dynamics of hybrid zones. *Heredity* **43**, 341–359.

Barton, N.H. and Gale, K.S. (1993). Genetic analysis of hybrid zones. In *Hybrid Zones and the Evolutionary Process*, R.G. Harrison, ed. Oxford University Press, Oxford, pp. 13–45.

Bazykin, A.D. (1969). Hypothetical mechanism of speciation. *Evolution* **23**, 685–687.

Beaumont, M.A. and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–980.

Beaumont, M.A. and Nichols, R.A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B* **263**, 1619–1626.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology* **13**, 827–836.

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**, 341–345.

Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–773.

Berry, O., Tocher, M.D. and Sarre, S.D. (2004). Can assignment tests measure dispersal?. *Molecular Ecology* **13**, 551–561.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

Broquet, T., Johnson, C.A., Petit, E., Burel, F. and Fryxell, J.M. (2006). Dispersal kurtosis and genetic structure in the American marten, *Martes americana*. *Molecular Ecology* **15**, 1689–1697.

Chakraborty, R. (1992). Multiple alleles and estimation of genetic parameters: computational equations showing involvement of all alleles. *Genetics* **130**, 231–234.

Chuang, C. and Cox, C. (1985). Pseudo maximum likelihood estimation for the Dirichlet-multinomial distribution. *Communications in Statistics Part A: Theory and Methods* **14**, 2293–2311.

Clark, J.S., Lewis, M. and Horváth, L. (2001). Invasion by extremes: population spread with variation in dispersal and reproduction. *American Naturalist* **157**, 537–554.

Cockerham, C.C. (1973). Analyses of gene frequencies. *Genetics* **74**, 679–700.

Cockerham, C.C. and Weir, B.S. (1987). Correlations, descent measures: drift with migration and mutation. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 8512–8514.

Corander, J., Waldmann, P. and Sillanpää, M. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Cornuet, J.M. and Beaumont, M.A. (2007). A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theoretical Population Biology* **71**, 12–19.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.

Cox, J.T. (1989). Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. *Annals of Probability* **17**, 1333–1366.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Coyne, J.A., Barton, N.H. and Turelli, M. (1997). A critique of Sewall Wright's shifting balance theory of evolution. *Evolution* **51**, 643–671.

Crow, J.F. and Aoki, K. (1984). Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 6073–6077.

DiCiccio, T.J. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science* **11**, 189–228.

Dobzhansky, T. and Wright, S. (1941). Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura*. *Genetics* **26**, 23–51.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, Boca Raton, FL.

Epperson, B.K. and Li, T.-Q. (1997). Gene dispersal and spatial genetic structure. *Evolution* **51**, 672–681.

Estoup, A., Beaumont, M., Sennedot, F., Moritz, C. and Cornuet, J.-M. (2004). Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution* **58**, 2021–2036.

Estoup, A., Rousset, F., Michalakis, Y., Cornuet, J.-M., Adriamanga, M. and Guyomard, R. (1998). Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). *Molecular Ecology* **7**, 339–353.

Évanno, G., Regnaut, S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620.

Ewing, G. and Rodrigo, A. (2006). Coalescent-based estimation of population parameters when the number of demes changes over time. *Molecular Biology and Evolution* **23**, 988–996.

Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Feder, J.L., Roethele, J.B., Wlazlo, B. and Berlocher, S.H. (1997). Selective maintenance of allozyme differences between sympatric host races of the apple maggot fly. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 11417–11421.

Fenster, C.B., Vekemans, X. and Hardy, O.J. (2003). Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). *Evolution* **57**, 995–1007.

Gaggiotti, O.E., Lange, O., Rassmann, K. and Gliddon, C. (1999). A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. *Molecular Ecology* **8**, 1513–1520.

Goudet, J., Raymond, M., de Meeüs, T. and Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics* **144**, 1931–1938.

Griffiths, R.C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Griffiths, R.C. and Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical Biosciences* **127**, 77–98.

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M. and Excoffier, L. (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**, 409–417.

Hardy, O.J. and Vekemans, X. (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**, 145–154.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology* **3**, e193.

Hudson, R.R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 1–44.

de Iorio, M. and Griffiths, R.C. (2004a). Importance sampling on coalescent histories. *Advances in Applied Probability* **36**, 417–433.

de Iorio, M. and Griffiths, R.C. (2004b). Importance sampling on coalescent histories. II. Subdivided population models. *Advances in Applied Probability* **36**, 434–454.

de Iorio, M., Griffiths, R.C., Leblois, R. and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology* **68**, 41–53.

Kingman, J.F.C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.

Kirkpatrick, S., Gelatt, C. and Vecchi, M. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Kitada, S., Hayashi, T. and Kishino, H. (2000). Empirical Bayes procedure for estimating genetic distance between populations and effective population size. *Genetics* **156**, 2063–2079.

Kitakado, T., Kitada, S., Kishino, H. and Skaug, H.J. (2006). An integrated-likelihood method for estimating genetic differentiation between populations. *Genetics* **173**, 2073–2082.

Koenig, W.D., Van Vuren, D. and Hooge, P.N. (1996). Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology and Evolution* **11**, 514–517.

Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**, 539–559.

Leblois, R., Estoup, A. and Rousset, F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Molecular Biology and Evolution* **20**, 491–502.

Leblois, R., Rousset, F. and Estoup, A. (2004). Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics* **166**, 1081–1092.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York.

Lenormand, T., Bourguet, D., Guillemaud, T. and Raymond, M. (1999). Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*. *Nature* **400**, 861–864.

Lewontin, R.C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233. Correction: **167**, 1039.

Lindley, D.V. (1990). The present position in Bayesian statistics. *Statistical Science* **5**, 44–89.

Long, J.C. (1986). The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics* **112**, 629–647.

Long, J.C., Naidu, J.M., Mohrenweiser, H.W., Gershowitz, H., Johnson, P.L. and Wood, J.W. (1986). Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. *American Journal of Physical Anthropology* **70**, 75–96.

Malécot, G. (1951). Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en génétique de population. *Annales de l'Université de Lyon A* **14**, 79–117.

Malécot, G. (1967). Identical loci and relationship. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, L.M. Le Cam and J. Neyman, eds. University of California Press, Berkeley, CA, pp. 317–332.

Malécot, G. (1975). Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology* **8**, 212–241.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society, Series B* **39**, 283–326.

Nagylaki, T. (1975). Conditions for the existence of clines. *Genetics* **80**, 595–615.

Nagylaki, T. (1976). The decay of genetic variability in geographically structured populations. II. *Theoretical Population Biology* **10**, 70–82.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.

Nagylaki, T. (1998). Fixation indices in subdivided populations. *Genetics* **148**, 1325–1332.

Nei, M. (1986). Definition and estimation of fixation indices. *Evolution* **40**, 643–645.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97–131.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347–354.

Pannell, J.R. and Charlesworth, B. (1999). Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. *Evolution* **53**, 664–676.

Rannala, B. and Hartigan, J.A. (1996). Estimating gene flow in island populations. *Genetical Research (Cambridge)* **67**, 147–158.

Rannala, B. and Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197–9201.

Raufaste, N. and Bonhomme, F. (2000). Properties of bias and variance of two multiallelic estimators of F_{ST}. *Theoretical Population Biology* **57**, 285–296.

Raymond, M. and Rousset, F. (1995). GENEPOP version 1.2: population genetics software for exact tests and ecumenicism. *The Journal of Heredity* **86**, 248–249.

Robledo-Arnuncio, J.J., Austerlitz, F. and Smouse, P.E. (2006). A new method of estimating the pollen dispersal curve independently of effective density. *Genetics* **173**, 1033–1046.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–1228.

Rousset, F. (1999a). Genetic differentiation in populations with different classes of individuals. *Theoretical Population Biology* **55**, 297–308.

Rousset, F. (1999b). Genetic differentiation within and between two habitats. *Genetics* **151**, 397–407.

Rousset, F. (2000). Genetic differentiation between individuals. *Journal of Evolutionary Biology* **13**, 58–62.

Rousset, F. (2001). Genetic approaches to the estimation of dispersal rates. In *Dispersal*, J. Clobert, E. Danchin, A.A. Dhondt and J.D. Nichols, eds. Oxford University Press, Oxford, pp. 18–28.

Rousset, F. (2002). Inbreeding and relatedness coefficients: what do they measure?. *Heredity* **88**, 371–380.

Rousset, F. (2004). *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Rousset, F. (2006). Separation of time scales, fixation probabilities and convergence to evolutionarily stable states under isolation by distance. *Theoretical Population Biology* **69**, 165–179.

Rousset, F. and Raymond, M. (1997). Statistical analyses of population genetic data: old tools, new concepts. *Trends in Ecology and Evolution* **12**, 313–317.

Sawyer, S. (1977). Asymptotic properties of the equilibrium probability of identity in a geographically structured population. *Advances in Applied Probability* **9**, 268–282.

Searle, S.R. (1971). *Linear Models*. John Wiley & Sons, New York.

Sites, J.W. Jr., Barton, N.H. and Reed, K.M. (1995). The genetic structure of a hybrid zone between two chromosome races of the *Sceloporus grammicus* complex (Sauria, Phrynosomatidae) in central Mexico. *Evolution* **49**, 9–36.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research (Cambridge)* **58**, 167–175.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279.

Slatkin, M. (1994). Gene flow and population structure. In *Ecological Genetics*, L.A. Real, ed. Princeton University Press, Princeton, NJ, pp. 3–17.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.

Smouse, P. and Williams, R.C. (1982). Multivariate analysis of HLA-disease associations. *Biometrics* **38**, 757–768.

Sokal, R. and Wartenberg, D.E. (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**, 219–237.

Storz, J.F., Ramakrishnan, U. and Alberts, S.C. (2001). Determinants of effective population size for loci with different modes of inheritance. *The Journal of Heredity* **92**, 497–502.

Sumner, J., Estoup, A., Rousset, F. and Moritz, C. (2001). 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. *Molecular Ecology* **10**, 1917–1927.

Swofford, D. and Olsen, G. (1990). Phylogeny reconstruction. In *Molecular Systematics*, D. Hillis and C. Moritz, eds. Sinauer Associates, pp. 411–501.

Tachida, H. (1985). Joint frequencies of alleles determined by separate formulations for the mating and mutation systems. *Genetics* **111**, 963–974.

Tsitrone, A., Rousset, F. and David, P. (2001). Heterosis, marker mutational processes, and population inbreeding history. *Genetics* **159**, 1845–1859.

Tufto, J., Engen, S. and Hindar, K. (1996). Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* **144**, 1911–1921.

Vekemans, X. and Hardy, O.J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**, 921–934.

Vitalis, R. and Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157**, 911–925.

Vitalis, R., Dawson, K. and Boursot, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811–1823.

Wakeley, J. and Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics* **159**, 893–905. Correction in *Genetics* **160**, 1263.

Wang, J. and Whitlock, M.C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* **163**, 429–446.

Waples, R.S. and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* **15**, 1419–1439.

Watts, P.C., Rousset, F., Saccheri, I.J., Leblois, R., Kemp, S.J. and Thompson, D.J. (2007). Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using an improved estimator. *Molecular Ecology* **16**, 737–751.

Weir, B.S. (1996). *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

Weir, B.S. and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Wilkins, J.F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**, 2227–2244.

Wilson, A.J., Hutchings, J.A. and Ferguson, M.M. (2004). Dispersal in a stream dwelling salmonid: inferences from tagging and microsatellite studies. *Conservation Genetics* **5**, 25–37.

Wilson, G.A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191.

Winters, J.B. and Waser, P.M. (2003). Gene dispersal and outbreeding in a philopatric mammal. *Molecular Ecology* **12**, 2251–2259.

Wood, J.W., Smouse, P.E. and Long, J.C. (1985). Sex-specific dispersal patterns in two human populations of highland New Guinea. *American Naturalist* **125**, 747–768.

Wright, S. (1931a). Evolution in Mendelian populations. *Genetics* **16**, 97–159. Reprinted in Wright (1986), pp. 98–160.

Wright, S. (1931b). Statistical theory of evolution. *Journal of the American Statistical Society* **26**(Suppl.), 201–208. Reprinted in Wright (1986), pp. 89–96.

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics* **31**, 39–59. Reprinted in Wright (1986), pp. 444–464.

Wright, S. (1949). Adaptation and selection. In *Genetics, Paleontology and Evolution*, G.L. Jepson, G.G. Simpson and E. Mayr, eds. Princeton University Press, pp. 365–389. Reprinted in Wright (1986), pp. 546–570.

Wright, S. (1969). *Evolution and the Genetics of Populations. II. The Theory of Gene Frequencies*. University of Chicago Press, Chicago, IL.

Wright, S. (1986). *Evolution: Selected Papers*. University of Chicago Press, Chicago, IL.

Zähle, I., Cox, J.T. and Durrett, R. (2005). The stepping stone model, II: genealogies and the infinite sites model. *Annals of Applied Probability* **15**, 671–699.

APPENDIX A: ANALYSIS OF VARIANCE AND PROBABILITIES OF IDENTITY

Here we detail the relationship between classical formulas for estimators of F-statistics and expressions in terms of frequency of identical pairs of genes. In the framework considered here, negative 'components of variance' (which are actually not variances) arise naturally.

We use the following notation: the total sample is made up of samples of n_i ($i = 1, \dots, n_s$) individuals in n_s subpopulations; $X_{ij:k}$ is an indicator variable for gene j ($j = 1, \dots, n_i$) in sample i being of allelic type k ($k = 1, \dots, K$), i.e. $X_{ij:k} = 1$ if the sampled gene is of type k and $X_{ij:k} = 0$ otherwise; standard dot notation is used for sample averages: e.g. for weights w_j, $X_. \equiv (\sum_j w_j X_j)/(\sum_j w_j)$ is a weighted average of the Xs. Here the weighting for each individual will be simply 1. A discussion of optimal weighting with respect to allele frequencies or samples sizes will be given in Appendix B. The indicator variables are given a single index (X_j) if no reference is made to a specific sample. π_k is the expected frequency of allele k, the expectation of $X_{j:k}$ over independent replicates of some evolutionary process (typically a mutation–drift stationary equilibrium, but this is in no way required).

For haploid data the statistical model is generally described as

$$X_{ij:k} = \mu + \alpha_i + \varepsilon_{ij}, \tag{28A.1}$$

$\mu = \pi_k$ here, α_i is a random effect with zero mean and variance σ_a^2, and ε_{ij} is a random effect with zero mean and variance σ_e^2. It is also assumed that $\mathrm{E}[\alpha_i \alpha_{i'}] = 0$ for $i \neq i'$ and that $\mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] = 0$ for $j \neq j'$ (e.g. Searle, 1971, p. 384), but this is precisely what we will not do here, in order to obtain the most general analogy. What remains more generally valid is a basic algebraic relationship of analysis of variance,

$$\begin{aligned}\sum_j w_j (X_j - \mu)^2 &= \sum_j w_j (X_j - X_{.} + X_{.} - \mu)^2 \\ &= \sum_j w_j (X_j - X_{.})^2 + \sum_j w_j (X_{.} - \mu)^2, \end{aligned} \tag{28A.2}$$

for any variable X and weights w_j because we still have $\sum_j w_j (X_{j:k} - X_{.:k}) = 0$ by definition of $X_{.:k}$.

The method of analysis of variance is then based on the computation of weighted sums of squares related as follows:

$$\begin{aligned}\sum_{\text{samples}}^{n_s} \sum_{\text{genes}}^{n_i} w_i (X_{ij:k} - X_{..:k})^2 &= \sum^{n_s} \sum^{n_i} w_i (X_{ij:k} - X_{i.:k})^2 \\ &\quad + \sum^{n_s} \sum^{n_i} w_i (X_{i.:k} - X_{..:k})^2 \qquad (28A.3) \\ &\equiv SS_{\text{w[ithin]}} + SS_{\text{b[etween]}} \text{for allele } k. \qquad (28A.4)\end{aligned}$$

For $w_i = 1$ these sums of squares will be expressed in terms of $S_1 \equiv \sum_i n_i$ and $S_2 \equiv \sum_i n_i^2$, of the observed frequency of pairs of different genes within samples which are both of type k, $\hat{Q}_{2:k} \equiv \sum_i \sum_{j \neq j'} X_{ij:k} X_{ij':k} / (S_2 - S_1)$, and of the observed frequency of pairs of different genes between samples which are both of type k, $\hat{Q}_{3:k} \equiv \sum_{i \neq i'} \sum_{j,j'} X_{ij:k} X_{i'j':k} / (S_1^2 - S_2)$. We first express the sums of squares using (28A.2), as follows:

$$SS_{\text{w}} = \sum_i \sum_j^{n_i} (X_{ij:k} - X_{i.:k})^2 = \sum_i \sum_j^{n_i} (X_{ij:k} - \pi_k)^2 - \sum_i n_i (X_{i.:k} - \pi_k)^2, \tag{28A.5}$$

and

$$SS_{\text{b}} = \sum_i \sum_j^{n_i} (X_{i.:k} - X_{..:k})^2 = \sum_i n_i (X_{i.:k} - \pi_k)^2 - S_1 (X_{..:k} - \pi_k)^2. \tag{28A.6}$$

The values of different variables that appear in these expressions, $(X_{j:k} - \pi_k)^2$ and $(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)$, for pairs j, j' of different genes, are summarized in Table 28.1 where $Q_{:k}$ is the probability that both genes of a pair are of type k. Note that $\mathrm{E}[(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)]$ is the covariance between allele frequencies in the subpopulations from

Table 28.1 Values of variables in comparisons of pairs of genes.

Pair of gene	kk	k, not k	None k
Probability of each pair	$Q_{:k}$	$2(\pi_k - Q_{:k})$	$1 - 2\pi_k + Q_{:k}$
Frequency in total sample	$\hat{Q}_{:k}$	$2(\tilde{\pi}_k - \hat{Q}_{:k})$	$1 - 2\tilde{\pi}_k + \hat{Q}_{:k}$
Variable	Value of variable for each pair		
$(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)$	$(1 - \pi_k)^2$	$-\pi_k(1 - \pi_k)$	π_k^2
$(X_{j:k} - \pi_k)^2 + (X_{j':k} - \pi_k)^2$	$2(1 - \pi_k)^2$	$(1 - \pi_k)^2 + \pi_k^2$	$2\pi_k^2$

which j and j' are sampled. From this table we see that

$$\sigma_a^2 + \sigma_e^2 = \mathrm{E}[(X_{j:k} - \pi_k)^2] = \pi_k(1 - \pi_k), \tag{28A.7}$$

$$\mathrm{E}[(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)] = Q_{:k} - \pi_k^2. \tag{28A.8}$$

where $Q_{:k}$ is $Q_{w:k} = Q_{2:k}$ for two genes within a sample and $Q_{b:k} = Q_{3:k}$ for two genes between samples. In particular we have

$$Q_{2:k} - \pi_k^2 = \mathrm{Cov}[X_{ij} X_{ij'}] = \sigma_a^2 + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}], \tag{28A.9}$$

$$Q_{3:k} - \pi_k^2 = \mathrm{Cov}[X_{ij} X_{i'j'}] = \mathrm{E}[\alpha_i \alpha_{i'}]. \tag{28A.10}$$

The latter expression confirms that $\mathrm{E}[\alpha_i \alpha_{i'}] \neq 0$: in general two genes in different subpopulations are more likely to be identical than two independent genes, i.e. $Q_{3:k} - \pi_k^2 > 0$. In the present case one could consider a slightly different parameterization of the model, so that this positive component would appear as a variance (e.g. Cockerham and Weir, 1987), but more generally this would be confusing because we may also have to consider negative terms, as shown below for diploid data.

The table also shows that the sample averages of $(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)$ and $(X_{j:k} - \pi_k)^2$ are $\hat{Q}_{:k} + \pi_k^2 - 2\pi_k\tilde{\pi}_k$ and $\hat{\pi}_k + \pi_k^2 - 2\pi_k\tilde{\pi}_k$, respectively. Here $\tilde{\pi}_k$ is the average allele frequency among all *pairs of genes* for which the average is written. This is the observed allele frequency by gene counting (denoted $\hat{\pi}_k$ or $\hat{\pi}_{i:k}$) when all pairs in the total sample or in sample i are considered. Among all pairs of genes sampled without replacement within each sample, this is $\hat{\pi}_{\mathrm{w}:k} \equiv \sum_i n_i(n_i - 1)\hat{\pi}_{i:k}/(S_2 - S_1)$. Among all pairs of genes from two different samples, $\hat{\pi}$ has value $\hat{\pi}_{\mathrm{b}:k}$ given by $(S_1^2 - S_2)\hat{\pi}_{\mathrm{b}:k} + (S_2 - S_1)\hat{\pi}_{\mathrm{w}:k} = (S_1^2 - S_1)\hat{\pi}$.

Then

$$\sum_i \sum_j^{n_j} (X_{ij:k} - \pi_k)^2 = S_1(X_{..:k} + \pi_k^2 - 2\pi_k X_{..:k}). \tag{28A.11}$$

Next

$$\begin{aligned}(X_{..:k} - \pi_k)^2 &= \left(\frac{\sum_{i,j} X_{ij:k} - \pi_k}{S_1}\right)^2 \qquad (28A.12)\\ &= \frac{1}{S_1^2}\Bigg(\sum_{i,j}(X_{ij:k} - \pi_k)^2\end{aligned}$$

$$+ \sum_{i \neq i', j, j'} (X_{i' j':k} - \pi_k)(X_{i' j':k} - \pi_k)$$

$$+ \sum_{i, j \neq j'} (X_{ij:k} - \pi_k)(X_{ij':k} - \pi_k)\Bigg) \tag{28A.13}$$

$$= \frac{1}{S_1^2}\Bigg(S_1(\hat{\pi} + \pi_k^2 - 2\hat{\pi}\pi_k) + (S_1^2 - S_2)(\hat{Q}_{3:k} + \pi_k^2 - 2\hat{\pi}_{\mathrm{b}:k}\pi_k)$$

$$+ (S_2 - S_1)(\hat{Q}_{2:k} + \pi_k^2 - 2\hat{\pi}_{\mathrm{w}:k}\pi_k)\Bigg) \tag{28A.14}$$

$$= (\hat{\pi} + \pi_k^2 - 2\hat{\pi}\pi_k) + \frac{1}{S_1^2}\Big((S_1^2 - S_2)(\hat{Q}_{3:k} - \hat{\pi}) + (S_2 - S_1)(\hat{Q}_{2:k} - \hat{\pi})\Big), \tag{28A.15}$$

by definition of $\hat{\pi}_{\mathrm{b}:k}$. Likewise

$$(X_{i.:k} - \pi_k)^2 = \left(\frac{\sum_{j=1}^{n_i} X_{ij:k} - \pi_k}{n_i}\right)^2 \tag{28A.16}$$

$$= \frac{1}{n_i^2}\left(\sum_{j=1}^{n_i}\Big(X_{ij:k} - \pi_k\Big)^2 + \sum_{j \neq j'}\Big(X_{ij:k} - \pi_k\Big)\Big(X_{ij':k} - \pi_k\Big)\right) \tag{28A.17}$$

$$= \frac{1}{n_i^2}\Big(n_i(\hat{\pi}_{i:k} + \pi_k^2 - 2\hat{\pi}_{i:k}\pi_k) + n_i(n_i - 1)(\hat{Q}_{i2:k} + \pi_k^2 - 2\hat{\pi}_{i:k}\pi_k)\Big) \tag{28A.18}$$

$$= (\hat{\pi}_{i:k} + \pi_k^2 - 2\hat{\pi}_{i:k}\pi_k) + \frac{n_i - 1}{n_i}(\hat{Q}_{i2:k} - \hat{\pi}_{i:k}), \tag{28A.19}$$

hence

$$\sum_{i=1}^{n_\mathrm{s}} n_i(X_{i.:k} - \pi_k)^2 = S_1(\hat{\pi}_k + \pi_k^2 - 2\hat{\pi}_k\pi_k) + \sum_i (n_i - 1)(\hat{Q}_{i2:k} - \hat{\pi}_{i:k}). \tag{28A.20}$$

Then from (28A.5), (28A.11) and (28A.20)

$$SS_\mathrm{w} = (S_1 - n_\mathrm{s})(\hat{\pi}_k - \hat{Q}_{2:k}), \tag{28A.21}$$

and from (28A.6), (28A.15) and (28A.20)

$$SS_\mathrm{b} = \sum_i (\hat{\pi}_{i:k} - \hat{\pi}_k) + \sum_i (n_i - 1)\hat{Q}_{i2:k} - (S_1 - S_2/S_1)\hat{Q}_{3:k} - (S_2/S_1 - 1)\hat{Q}_{2:k}. \tag{28A.22}$$

Note that as in (28A.20), allele frequencies terms do not reduce to a function of $\hat{\pi}$ only: the term $\sum_i (\hat{\pi}_{i:k} - \hat{\pi})$ will usually be nonzero when sample sizes are unequal. Taking expectations, one has

$$\begin{aligned}\mathrm{E}\left[SS_{\mathrm{w}}\right] &= (S_1 - n_{\mathrm{s}})(\pi_k - Q_{2:k}),\\ \mathrm{E}\left[SS_{\mathrm{b}}\right] &= (S_1 - S_2/S_1)(Q_{2:k} - Q_{3:k}) + (n_{\mathrm{s}} - 1)(\pi_k - Q_{2:k})\\ &= (S_1 - S_2/S_1)(\sigma_a^2 - \mathrm{E}[\alpha_i\alpha_{i'}] + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}]) + (n_{\mathrm{s}} - 1)(\sigma_e^2 - \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}]).\end{aligned} \tag{28A.23}$$

These relationships hold whatever the model considered (fixed or random, etc.). They are formally equivalent to a standard analysis of variance (e.g. Searle, 1971) on the indicator variables $X_{ij:k}$, except that (1) $\mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}]$ and $\mathrm{E}[\alpha_i\alpha_{i'}]$ are not assumed null, and (2) the sums of squares are themselves summed over alleles. When we write these two modifications as '$\xrightarrow{1}$' and '$\xrightarrow{2}$' the equivalence of expectations in the standard formulas of analysis of variance with expressions in terms of probabilities of identity is as follows:

$$\begin{aligned}\sigma_a^2 \xrightarrow{1} \sigma_a^2 - \mathrm{E}[\alpha_i\alpha_{i'}] + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] \xrightarrow{2} Q_2 - Q_3 \equiv (1 - Q_3)F_{\mathrm{ST}}\\ \sigma_e^2 \xrightarrow{1} \sigma_e^2 - \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] \xrightarrow{2} 1 - Q_2 = (1 - Q_3)(1 - F_{\mathrm{ST}}).\end{aligned} \tag{28A.24}$$

Hence the 'intraclass covariance' $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, often taken as a definition of F_{ST}, should be interpreted as

$$F_{\mathrm{ST}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \xrightarrow{1} F_{\mathrm{ST}} = \frac{\sigma_a^2 + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] - \mathrm{E}[\alpha_i\alpha_{i'}]}{\sigma_a^2 + \sigma_e^2 - \mathrm{E}[\alpha_i\alpha_{i'}]}, \tag{28A.25}$$

where the latter expression may be considered more general.

For diploid data, the model is $X_{ijl:k} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijl}$ for gene l ($l = 1, 2$ for diploids) of individual j in population i. With $\sigma_a^2 \equiv \mathrm{E}[\alpha_i^2]$, $\sigma_b^2 \equiv \mathrm{E}[\beta_{ij}^2]$, $\sigma_e^2 \equiv \mathrm{E}[\varepsilon_{ijl}^2]$, we have

$$\begin{aligned}\sigma_a^2 \xrightarrow{1} \sigma_a^2 - \mathrm{E}[\alpha_i\alpha_{i'}] + \mathrm{E}[\beta_{ij}\beta_{ij'}] \xrightarrow{2} Q_2 - Q_3 \equiv (1 - Q_3)F_{\mathrm{ST}}\\ \sigma_b^2 \xrightarrow{1} \sigma_b^2 - \mathrm{E}[\beta_{ij}\beta_{ij'}] + \mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}] \xrightarrow{2} Q_1 - Q_2 \equiv (1 - Q_3)F_{\mathrm{IS}}(1 - F_{\mathrm{ST}})\\ \sigma_e^2 \xrightarrow{1} \sigma_e^2 - \mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}] \xrightarrow{2} 1 - Q_1 = (1 - Q_3)(1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}}),\end{aligned} \tag{28A.26}$$

where Q_1 is the probability of identity of genes within a diploid individual, Q_2 is for genes between individuals within subpopulations, and Q_3 between subpopulations. Thus in both formalisms we see that the 'components of variance' actually translate into more general expressions that can be negative. When there is an heterozygote excess within demes ($Q_1 < Q_2$), $\mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}]$ is negative, and $\sigma_b^2 - \mathrm{E}[\beta_{ij}\beta_{ij'}] + \mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}]$ is negative.

In the haploid case, (28A.23) implies that

$$\frac{(S_1 - n_{\mathrm{s}})\mathrm{E}[SS_{\mathrm{b}}] - (n_{\mathrm{s}} - 1)\mathrm{E}[SS_{\mathrm{w}}]}{(S_1 - n_{\mathrm{s}})\mathrm{E}[SS_{\mathrm{b}}] + (W_c - 1)(n_{\mathrm{s}} - 1)\mathrm{E}[SS_{\mathrm{w}}]} = \frac{Q_2 - Q_3}{1 - Q_3}, \tag{28A.27}$$

where SS_{w} and SS_{b} are now summed over alleles (e.g. $SS_{\mathrm{w}} \equiv \sum^{n_{\mathrm{s}}}\sum^{n_i}\sum_k (X_{ij:k} - X_{i.:k})^2$), and $W_c \equiv (S_1 - S_2/S_1)/(n_{\mathrm{s}} - 1)$. Although we have related the expectation of the different terms to ‘components of variance’ in a model such as (28A.1), we note again that the last equality holds independently of such a model, because it is only based on the basic relationship (28A.2). Accordingly, an estimator of F_{ST} is the ratio of unbiased estimators

$$\frac{(S_1 - n_{\mathrm{s}})SS_{\mathrm{b}} - (n_{\mathrm{s}} - 1)SS_{\mathrm{w}}}{(S_1 - n_{\mathrm{s}})SS_{\mathrm{b}} + (W_c - 1)(n_{\mathrm{s}} - 1)SS_{\mathrm{w}}}. \tag{28A.28}$$

(see also Weir, 1996, p. 182). One could hope that this estimator is directly interpretable as $(\hat{Q}_2 - \hat{Q}_3)/(1 - \hat{Q}_3)$, where $\hat{Q}_j = \sum_k \hat{Q}_{j:k}$ are the frequencies of identical pairs of genes in the sample, computed by simple counting either within (for Q_2) or between (for Q_3) samples ($\hat{Q}_2$ being is an average over the different samples, weighted according to the number of pairs in each sample). But this is not so when sample sizes are unequal, because the term $\sum_i (\hat{\pi}_{i:k} - \hat{\pi})$ from (28A.22) remains in the above expression. The expression closest to $(\hat{Q}_2 - \hat{Q}_3)/(1 - \hat{Q}_3)$ that I have found for Weir and Cockerham’s estimator is

$$\frac{\tilde{Q}_2 - \hat{Q}_3}{1 - \hat{Q}_3 + \sum_i (n_i - 1)(\hat{Q}_{i2} - \hat{Q}_2)\frac{S_2 - S_1}{(S_1^2 - S_2)(S_1 - n_{\mathrm{s}})}}, \tag{28A.29}$$

in terms of the weighted frequency

$$\tilde{Q}_2 = \frac{(S_1 - 1)\sum_i (n_i - 1)\hat{Q}_{i2} - (S_1 - n_{\mathrm{s}})(S_2/S_1 - 1)\hat{Q}_2}{(S_1 - n_{\mathrm{s}})(S_1 - S_2/S_1)}, \tag{28A.30}$$

and where $\hat{Q}_{i2}$ is the observed frequency of pairs of genes identical in state among all pairs taken without replacement within sample i. Compared to the analysis-of-variance estimator, the simple strategy of estimating any function of probabilities of identities by the equivalent function of frequencies of identical pairs of genes is equally ‘unbiased’, has no obvious drawback and is easily adaptable to different settings.

For multilocus data it is usual to compute the estimator as a sum of locus-specific numerators over a sum of locus-specific denominators; see e.g. Weir and Cockerham (1984) or Weir (1996) for details. Note that the sums are weighted differently in these two references. The numerator in Weir and Cockerham (1984), eqs. (2) and (10), is $\bar{n}/W_c$ times the one in Weir (1996), p.178–179. Parallel changes in the denominator ensure that the one-locus estimators are identical, but the multilocus estimators will be different if $\bar{n}/W_c$ varies between loci.

APPENDIX B: LIKELIHOOD ANALYSIS OF THE ISLAND MODEL

Sampling Formulas

Consider an infinite island model of haploid subpopulations where there are K alleles and n_{s} subpopulations are sampled. The following notation will be used: π_k is the frequency of allele k in the total population (which is not a random variable here) and p_{ki} is frequency of allele k in subpopulation i; n_{ki} is the number of genes of type k in the sample from subpopulation i; n_i is the size of sample i, $\overline{n}$ is the average n_i, $\tilde{\pi}_k \equiv \sum_i n_{ki}/\sum_i n_i$ and $\tilde{p}_{ki} \equiv n_{ki}/n_i$ are the observed frequencies of allele k in the total sample and in sample i, respectively.

The distribution of the p_{ki}s in population i follows a Dirichlet distribution,

$$L(p_{ki}, \dots, p_{Ki}) = \Gamma(M) \prod_k \frac{p_{ki}^{M\pi_k - 1}}{\Gamma(M\pi_k)}. \tag{28B.1}$$

This type of distribution arises as a diffusion approximation to the discrete generation Wright–Fisher model, where M is twice the number of migrant genes per generation (Wright, 1949), e.g. $2Nm$ or $4Nm$, and more generally in any scenario that can be approximated by the n-coalescent.

It should be noted that this equation is valid for each subpopulation with its own size, N_i, and its own immigration rate, m_i. Thus the likelihood of samples may be given for an infinite island model only characterized by the homogeneous dispersal of individuals to other demes: the N_is and m_is need not be identical in all subpopulations. This result implies that, with a large number of subpopulations, one can only estimate the products $N_i m_i$ for each deme.

Consider the vector of counts n_{ki} of allele k in subpopulation i, $\mathbf{n}_i \equiv (n_{1i}, \cdots, n_{Ki})$, and the corresponding multinomial coefficient $C(\mathbf{n}_i) \equiv n_i!/\prod_{k=1}^{K} n_{ki}!$. The conditional probability distribution of the ith sample, given the subpopulation frequencies $\mathbf{p}_i \equiv (p_{1i}, \dots, p_{Ki})$, is multinomial:

$$C(\mathbf{n}_i) \prod_k^K p_{ki}^{n_{ki}}. \tag{28B.2}$$

The probability distribution of a sample $\mathbf{n}_i$ in subpopulation i must be expressed as a function only of the parameters, M and of expected allele frequencies (expectations under stochastic model) $\boldsymbol{\pi} \equiv (\pi_1, \dots, \pi_K)$, by combining (28B.1) and (28B.2) and summing over the set $\mathcal{S}$ of possible values of allele frequencies $\mathbf{p}_i$:

$$L(M, \boldsymbol{\pi}; n_{1i}, \dots, n_{Ki}) = \int \underset{\mathcal{S}}{\cdots} \int \Gamma(M) \prod_k^K \frac{p_{ki}^{M\pi_k - 1}}{\Gamma(M\pi_k)} C(\mathbf{n}_i) \prod_k^K p_{ki}^{n_{ki}} \, d\mathbf{p}_i \tag{28B.3}$$

$$= \frac{\Gamma(M)}{\Gamma(M + n_i)} C(\mathbf{n}_i) \prod_k^K \frac{\Gamma(M\pi_k + n_{ki})}{\Gamma(M\pi_k)}. \tag{28B.4}$$

This distribution is the Dirichlet-multinomial distribution. In the infinite island model, subpopulation frequencies are independent from each other in each subpopulation, so that the likelihood of a total sample from n_{s} subpopulations is

$$L(M, \boldsymbol{\pi}) = \left(\frac{\Gamma(M)}{\Gamma(M + n_i)} \right)^{n_{\mathrm{s}}} \prod_{i=1}^{n_{\mathrm{s}}} C(\mathbf{n}_i) \prod_k^K \frac{\Gamma(M\pi_k + n_{ki})}{\Gamma(M\pi_k)}. \tag{28B.5}$$

The pseudo–maximum likelihood estimator $\hat{M}_A$ of M had been previously defined by Chuang and Cox (1985) as the solution of $\partial \ln L/\partial M|_{\boldsymbol{\pi}=\tilde{\boldsymbol{\pi}}, M=\hat{M}_A} = 0$. From (28B.5), this is the solution of

$$0 = \left[\sum_k^K \sum_i^{n_{\mathrm{s}}} \tilde{\pi}_k \left(\sum_{k=0}^{n_{ki}-1} \frac{1}{\tilde{\pi}_k M + k} \right) - \sum_i^{n_{\mathrm{s}}} \left(\sum_{k=0}^{n_i - 1} \frac{1}{M + k} \right) \right]. \tag{28B.6}$$

Efficiency in the Island Model

When $M \to \infty$ (high migration rates) the Dirichlet-multinomial distribution converges to a multinomial with parameter $\boldsymbol{\pi}$, so the sampling distribution for the total sample is a product of multinomials with the same parameter $\boldsymbol{\pi}$: this corresponds to the case of no population differentiation. Thus we can construct asymptotically efficient statistics for detecting weak differentiation from a study of the properties of the likelihood when $M \to \infty$. To that aim it is simpler to express it as a function of $\psi \equiv 1/M$ and compute the Taylor expansion near $\psi = 0$. From (28B.6), it may be shown that

$$\frac{\partial \ln L}{\partial \psi} = \sum_{k}^{K} \sum_{i}^{n_s} \frac{n_{ki}(n_{ki} - 1)}{2\pi_k} - \sum_{i}^{n_s} \frac{n_i(n_i - 1)}{2} + O(\psi), \qquad (28B.7)$$

and a statistic of interest (effectively a score statistic, Cox and Hinkley, 1974, Chapter 9) may be constructed as

$$\tilde{U} \equiv \lim_{\psi \to 0} \frac{\partial \ln L}{\partial \psi}\bigg|_{\boldsymbol{\pi}=\tilde{\boldsymbol{\pi}}} = \sum_{k}^{K} \sum_{i}^{n_s} \frac{n_{ki}(n_{ki} - 1)}{2\tilde{\pi}_k} - \sum_{i}^{n_s} \frac{n_i(n_i - 1)}{2}, \qquad (28B.8)$$

where the $\tilde{\pi}$s are the observed allele frequencies in the total sample, which are the MLEs of the πs in the case $\psi = 0$. This result draws a connection between the likelihood and the moment methods (see also Balding, 2003). Since the second sum in (28B.8) is fixed for given sample sizes n_i, the score statistic is essentially a sum of squares and can be considered in an analysis of variance framework. It shows that asymptotically efficient weights w_k of the sum of squares for the different alleles are proportional to $1/\tilde{\pi}_k$, and the weights of the sum of squares for the different samples are proportional to n_i^2 for the different samples, i.e. $w_i = n_i$ for each individual in (28A.3). The allele weighting is not new: it is implicit in the matrix formulations of Smouse and Williams (1982) and Long (1986) (see Weir and Cockerham, 1984; Chakraborty, 1992) and in standard test statistics such as the χ^2 or log-likelihood for multinomial models. Consistent with the above analysis assuming weak differentiation ($\psi \to 0$), it leads to estimators with efficient properties only for low differentiation (Raufaste and Bonhomme, 2000). By contrast the sample size weighting is odd and has not been previously considered in analysis of variance. But it may bring very little (F.R., unpublished data): the allele weighting is generally sufficient to turn moment statistics into efficient statistics when differentiation is low.

Weighting according to observed allele frequencies or to other measures of genetic diversity may have some drawbacks, particularly when one considers more general models than the island model. This weighting seems to imply that the ratio of expected sum of squares, conditional on genetic diversity, is independent of the value of the conditioning variable. As noted in the main Text, this is approximately so in the island model, and more generally under a separation of timescales when the slow process is an n-coalescent, but may not hold more generally. Then, the only consistent method would be to sum the $\hat{Q}$ terms directly in the numerator and denominator; any other method would introduce a bias. Further, selection of markers with specific levels of variability–as is often the case in practice–could also introduce an ascertainment bias.

Please note that the keywords will not be included in the printed book, but are required for the online presentation of this book which will be published on Wiley InterScience (www.interscience.wiley.com). If the keywords are not present below, please take this opportunity to add them now.

The keywords should be not less than 10 words.

KEYWORDS: dispersal, coalescence, likelihood, analysis of variance, F-statistics, island model, isolation-by-distance, effective population parameters