MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 821–830 doi: 10.1111/j.1755-0998.2010.02873.x

METHODOLOGICAL ADVANCES - INFERENCE OF SPATIAL STRUCTURE

Quantifying population structure using the *F*-model

OSCAR E. GAGGIOTTI* and MATTHIEU FOLL†
Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Université Joseph Fourier, BP 53, 38041 GRENOBLE, France, †CMPG, Institute of Ecology and Evolution, University of Berne, 3012 Berne, Switzerland

Abstract

We review a model-based approach to estimate local population F_{ST}'s that is based on the multinomial-Dirichlet distribution, the so-called *F*-model. As opposed to the standard method of estimating a single F_{ST} value, this approach takes into account the fact that in most if not all realistic situations, local populations differ in their effective sizes and migration rates. Therefore, the use of this approach can help better describe the genetic structure of populations. Despite this obvious advantage, this method has remained largely underutilized by molecular ecologists. Thus, the objective of this review is to foster its use for studying the genetic structure of metapopulations. We present the derivation of the Bayesian formulation for the estimation of population-specific F_{ST}'s based on the multinomial-Dirichlet distribution. We describe several recent applications of the F-model and present the results of a small simulation study that explains how the *F*-model can help better describe the genetic structure of populations.

Keywords: Bayesian statistics, *F*-statistics, genetic structure, metapopulation, multinomial-Dirichlet

Received 17 January 2010; revision received 9 April 2010; accepted 15 April 2010

Introduction

Many if not most species exhibit discontinuities in their spatial distribution. Such spatial structuring has important effects on their evolutionary potential, a fact that was recognized early on by Sewall Wright who introduced the so-called island model of population structure (Wright 1931) to explain how migration changed allele frequencies in a semi-isolated population. Under this model, genetic diversity is structured into within and among population components that can be quantified using Wright's (1951) *F*-statistics: F_{IS}, $F_{ST,}$ and F_{IT}, which measure respectively the shared ancestry between alleles of an individual relative to the population, the shared ancestry within the population relative to the metapopulation, and the shared ancestry between alleles of an individual relative to the metapopulation.

Of the three *F*-statistics, F_{ST} is the one that attracted most attention because of the important insights that it can bring into the understanding of how different evolutionary forces influence the genetic structure of populations (Holsinger & Weir 2009). In a strict island model where all populations have the same effective size and immigration rates, a single 'global' F_{ST} value suffices to characterize the genetic structure. Moreover, under this model, one of the simplest interpretations of F_{ST} is the heterozygote deficit because of population subdivision (e.g. Excoffier 2007). This has become the prevailing interpretation among molecular ecologists and has fostered the adoption of a single estimate of F_{ST} as the standard approach to measuring genetic differentiation. This approach, however, ignores the fact that in most if not all realistic situations, local populations differ in their effective sizes and migration rates making it necessary to estimate population-specific F_{ST}'s. This alternative strategy has been completely ignored in the empirical studies. Instead, the study of more complex scenarios of spatial subdivision is carried out using pair-wise F_{ST}'s. This is particularly the case when the main focus is on the detection of isolation by distance patterns (Slatkin 1993).

Several interpretations of Wright's definitions of F_{ST} exist in the literature (reviewed by Balding 2003). Here, we adopt a model-based approach that allows us to define the F_{ST} of a given population j, noted, F_{ST}^{j} as the probability that two genes chosen randomly from the population share a common ancestor within that population without immigration or colonization (Balding 2003).

Correspondence: Oscar E. Gaggiotti, Fax: +33 476514279;
E-mail: Oscar.Gaggiotti@ujf-grenoble.fr

Such a definition allows for differences in local population sizes and migration rates. Moreover, using the properties of the Dirichlet distribution, it can be shown that it leads to the well-known expression

$$F_{ST}^{ij} = \frac{\mathrm{Var}(\tilde{p}_{ijk})}{p_{ik}(1-p_{ik})},$$

where, $\tilde{p}_{ijk}$ is the frequency of allele k at locus i and population j, and p_{ik} is the frequency of this same allele and locus but for the metapopulation as a whole. This expression is true for all alleles at a locus, and therefore shows that F_{ST} is constant over alleles.

The idea of estimating population-specific F_{ST}'s is somewhat old, having been introduced by Balding & Nichols (1995). It has been revisited many times in the recent statistical genetics literature (Balding 2003; Falush *et al.* 2003; Beaumont & Balding 2004; Beaumont 2005; Foll & Gaggiotti 2006, 2008; Faubet & Gaggiotti 2008) but mostly in the context of methods aimed at identifying outlier loci or estimating migration rates.

The purpose of this review is to foster the adoption of this approach as a standard practice in molecular ecology. We will focus on a particular model-based approach that has received a lot of recent attention, the *F*-model (Falush *et al.* 2003), which is based on the multinomial-Dirichlet distribution. We note, however, that local F_{ST}'s can be calculated using other approaches, including the method of moment estimator of Weir & Hill (2002) as well as descriptive ones such as that proposed by Nei. This and many other important details are explained by Balding (2003), who presents a very rigorous albeit rather technical review of all the different approaches that can be used to estimate F_{ST}. Several other more general reviews exist (e.g. Chakrabory & Danker-Hopfe 1991; Weir & Hill 2002; Holsinger & Weir 2009) but they only address the estimation of the global F_{ST}.

We first present a review of the theoretical developments that gave rise to the *F*-model and then provide a derivation of the Bayesian formulation for the estimation of population-specific F_{ST}'s based on this model. We describe several recent applications of the *F*-model and present the results of a small simulation study that explains how it can help better describe the genetic structure of populations.

Theoretical developments leading to the *F*-model

The estimation of local F_{ST}'s is most often carried out using likelihood approaches (either Maximum Likelihood or Bayesian) based on the multinomial-Dirichlet (or multivariate-Polya) distribution, a multidimensional generalization of the Beta-binomial (Johnson *et al.* 1997). This distribution arises naturally when modelling the sampling of alleles from one of the local populations of a metapopulation (Rannala & Hartigan 1996). The basic idea is that the sample of allele counts at locus i and population j, $\{a_{ij1}, a_{ij2}, \ldots, a_{ijK}\}$, is obtained from a multinomial distribution with probability vector $\tilde{\mathbf{p}}_{ij} = \{\tilde{p}_{ij1}, \tilde{p}_{ij2}, \ldots, \tilde{p}_{ijKi}\}$, corresponding to the unknown allele frequency distribution. Furthermore, as Wright (1949) showed, under an island model at equilibrium, this allele frequency distribution follows a finite Dirichlet distribution with parameter vector $\{\theta p_1, \theta p_2, \ldots, \theta p_k\}$, where $\theta = 1/F_{ST} - 1$. As noted by Rannala & Hartigan (1996), these two distributions give rise to the multinomial-Dirichlet, which can be used as the likelihood to be employed for making inferences about population structure under both a discrete-generation and a continuous-generation island model. It is worth noting that sometime before them, Balding & Nichols (1995) provided a recursive formula for sampling alleles at a multiallelic locus in an island model but they did not explicitly mention that it was a recursive formulation of the multinomial-Dirichlet distribution. In the next section, we provide a step-by-step derivation of this useful distribution.

Rannala & Hartigan (1996) used the multinomial-Dirichlet to develop maximum likelihood and pseudo-maximum likelihood methods to estimate the rate of gene flow into island populations. Holsinger (1999) proposed the use of a Bayesian framework and showed that it can be used to obtain estimates of F_{ST} under the fixed-effect model of population sampling corresponding to Nei's (1973) G_{ST} and also under the random-effect model of population sampling corresponding to Weir & Cockerham's (1984) θ. He also showed that it is possible to extend it to consider scenarios with hierarchical population structure.

Balding *et al.* (1996) were the first to note that the use of a hierarchical Bayesian approach based on the multinomial-Dirichlet likelihood allows taking into account the fact that F_{ST} is likely to vary across demes because of differences in population size, and migration and reproduction patterns. This idea was further developed by Balding (2003). However, neither provides a full derivation of the Bayesian model, something that we do in the following section to better explain the details of the estimation method.

A Bayesian Formulation for the estimation of local population F_{ST}'s

For the sake of simplicity, we will derive the Bayesian formulation for the case of a discrete island model at migration-drift equilibrium but we note that it can also be used under more complex demographic scenarios provided that it is possible to assume that sampled populations exchange genes through a unique and common migrant pool (Balding 2003; Beaumont 2005).

The objective of the statistical model (presented in Fig. 1a) is to estimate population-specific F_{ST}'s from a sample of alleles drawn from J populations that have unknown allele frequency distributions at a set of I loci. If we let K_i represent the number of distinct alleles at locus i, then the data set consists of a matrix $\mathbf{A} = \{\mathbf{a}_{ij}\}$, where the vector $\mathbf{a}_{ij} = \{a_{ij1}, a_{ij2}, \ldots, a_{ijKi}\}$ contains the number of copies of each one of the K_i distinct alleles observed at locus i in the sample from population j. Equivalently, we will use the matrix $\mathbf{P} = \{\tilde{\mathbf{p}}_{ij}\}$ to represent the unknown allele frequency distributions at locus i in population j, $\tilde{\mathbf{p}}_{ij} = \{\tilde{p}_{ij1}, \tilde{p}_{ij2}, \ldots, \tilde{p}_{ijKi}\}$. Finally, the degree of genetic differentiation between population j and the migrant pool is noted F_{ST}^{j}.

The use of a Bayesian framework requires the definition of prior distributions for the unknown parameters, in our case, the allele frequency distributions and the F_{ST}^{j}s, and the definition of a likelihood function that relates the observed allele frequencies with these unknown parameters.

As noted by Rannala & Hartigan (1995), the population genetics theory developed by Wright (1949) shows that the appropriate prior for the allele frequency distributions $\tilde{\mathbf{p}}_{ij}$ is a Dirichlet with parameters $\theta_j \mathbf{p}_i$, where $\mathbf{p}_i = \{p_{i1}, p_{i2}, \ldots, p_{iKi}\}$ is the allele frequency distribution at locus i in the migrant pool, and $\theta_j = 1/F_{ST}^{j} - 1 = 4N_j m_j$ is the effective number of migrants in population j as N_j and m_j are respectively the effective size and the migration rate for population j. The prior for $\boldsymbol{p}_i$ is then,

$$\pi\left(\tilde{\mathbf{p}}_{ij}|\mathbf{p}_i, F_{ST}^{j}\right) = \Gamma(\theta_j)\prod_{k=1}^{K_i}\frac{\tilde{p}_{ijk}^{\,\theta_j p_{ik}-1}}{\Gamma(\theta_j p_{ik})}. \quad (1)$$

The likelihood function for the allele counts $\mathbf{a}_{ij}$ is easily obtained by noting that the observed allele counts can be viewed as sampled from the true allele frequencies $\tilde{\mathbf{p}}_{ij}$ (Rannala & Hartigan 1996; Holsinger 1999). Thus, the appropriate likelihood function is the multinomial distribution:

$$P\left(\mathbf{a}_{ij}|\tilde{\mathbf{p}}_{ij}\right) = \frac{n_{ij}!}{a_{ij1}!a_{ij2}!\cdots a_{ij1Ki}!}p_{ij1}^{a_{ij1}}p_{ij1}^{a_{ij2}}\cdots p_{ij1}^{a_{ijKi}}, \quad (2)$$

where $n_{ij} = \sum_k a_{ijk}$.

In principle, we could formulate the Bayesian model using the prior distributions and likelihood function described previously (see Fig. 1a). However, it is possible to calculate the marginal distribution of $\mathbf{a}_{ij}$ so as to obtain a likelihood function that directly links the observed allele frequency counts with the unknown parameters θ_js and the allele frequency distributions in the migrant pool, $\mathbf{p}_i$. It suffices to note that the Dirichlet distribution is the conjugate prior of the multinomial, which allows us to eliminate the nuisance parameters $\tilde{\mathbf{p}}_{ij}$ by integrating over them:

$$P\left(\mathbf{a}_{ij}|\mathbf{p}_i, F_{ST}^{j}\right) = \int\cdots\int P\left(\mathbf{a}_{ij}|\tilde{\mathbf{p}}_{ij}\right)\pi(\tilde{\mathbf{p}}_{ij}|\mathbf{p}_i, \theta_j)\mathrm{d}\tilde{p}_{ij1}\cdots\mathrm{d}\tilde{p}_{ijKi}.$$

The solution to this integral is the multinomial-Dirichlet distribution:

$$P\left(\mathbf{a}_{ij}|\mathbf{p}_i, F_{ST}^{j}\right) = \frac{n_{ij}!\Gamma(\theta_j)}{\Gamma(n_{ij}+\theta_j)}\prod_{k=1}^{K_i}\frac{\Gamma(a_{ijk}+\theta_j p_{ik})}{a_{ijk}!\Gamma(\theta_j p_{ik})}. \quad (3)$$

We can then obtain a likelihood function to carry out statistical inference by multiplying across all loci and subpopulations,

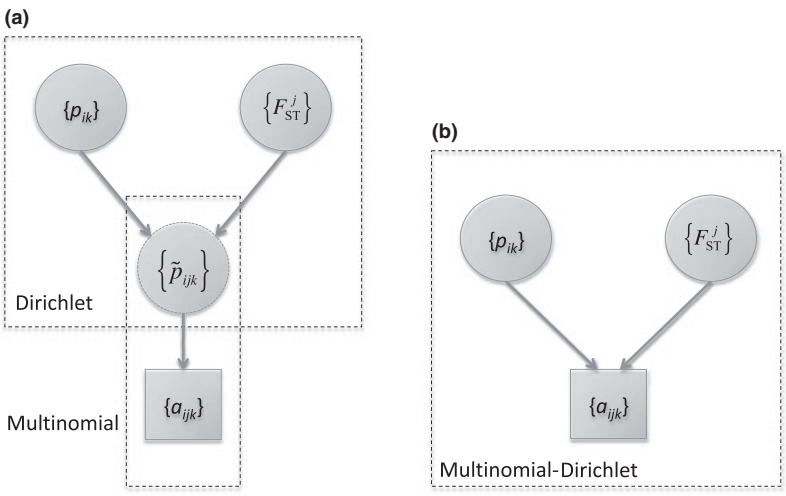

Fig. 1 Directed acyclic graph describing the Bayesian formulation of the F-model. Square nodes represent known quantities and circles represent parameters to be estimated. Dashed circles represent quantities that are directly determined by other parameters of the model and therefore are not actual parameters. Lines between the nodes represent direct stochastic relationships within the model. (a) the Bayesian model without integrating the Dirichlet likelihood and the multinomial prior, (b) the simpler Bayesian model obtained after integrating the Dirichlet prior and the multinomial likelihood to obtain the multinomial-Dirichlet likelihood.

$$L(\mathbf{p}, \mathbf{F}_{ST}) = \prod_{j}^{J} \prod_{i}^{I} P\left(\mathbf{a}_{ij} | \mathbf{p}_i, F_{ST}^{j}\right), \quad (4)$$

where $\mathbf{F}_{ST} = \left\{F_{ST}^{1}, F_{ST}^{2}, \ldots, F_{ST}^{J}\right\}$.
Therefore, we can use a simpler Bayesian formulation that uses this equation as the likelihood and a uniform prior for the F_{ST}^{j}s. The directed acyclic graph for this model is shown in Fig. 1b and its full posterior distribution is given by

$$\pi(\mathbf{p}, \mathbf{F}_{ST} | \mathbf{A}) \propto L(\mathbf{p}, \mathbf{F}_{ST}) \pi(\mathbf{p}) \pi(\mathbf{F}_{ST}), \quad (5)$$

where $\pi(\mathbf{p})$ = Dir (1,1,…,1) and $\pi(\mathbf{F}_{ST})$ = U(0,1).
This Bayesian formulation corresponds to that implemented in GESTE (Foll & Gaggiotti 2006) when the user chooses not to incorporate environmental data into the analysis.

One likelihood function with many uses

The use of a Bayesian formulation that incorporates the multinomial-Dirichlet distribution allows us to address many different questions concerning the genetic structure of populations. All that is needed is to change the prior distribution used for F_{ST} by adding hyper-parameters that model the effect of evolutionary forces on the degree of genetic differentiation. This strategy in combination with the Reversible Jump Markov Chain Monte Carlo (RJMCMC) method introduced by Gaggiotti *et al.* (2004) allows the testing of specific hypothesis concerning the factors responsible for genetic differentiation. Here, we provide some examples that illustrate how the Bayesian formulation (5) can be extended to achieve this goal.

Beaumont & Balding (2004) present a genome-scan method to identify outlier loci based on the fact that F_{ST} is controlled by locus-specific effects, such as selection and mutation, and genome-wide effects, such as demographic history and migration rates. Thus, they consider locus and population-specific F_{ST}s using indices i for locus and j for population. They describe this relationship using a hierarchical formulation (see Fig. 2a) based on the logistic regression model:

$$\log\left(\frac{F_{ST}^{ij}}{1 - F_{ST}^{ij}}\right) = \alpha_i + \beta_j + \gamma_{ij},$$

where α_i is a locus effect, β_j is a population effect and γ_{ij} is a locus-by-population effect. The Bayesian formulation (4) can now be rewritten in terms of the logistic regression parameters,

$$\pi(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{A}) \propto L(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \pi(\mathbf{p}) \pi(\boldsymbol{\alpha}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\gamma}), \quad (6)$$

where the priors $\pi(\boldsymbol{\alpha})$, $\pi(\boldsymbol{\beta})$, and $\pi(\boldsymbol{\gamma})$ are all Gaussian. The criterion for deciding if a locus is an outlier or not is based on the posterior distribution of α_i. Riebler *et al.* (2008) propose an alternative criterion based on the use of an auxiliary variable. Foll & Gaggiotti (2008) employ the RJMCMC approach that allows a rigorous and direct estimation of the probability that a locus is an outlier. They also extended the method so as to make it applicable to dominant markers such as amplified fragment-length polymorphisms, AFLPs.

Foll & Gaggiotti (2006) present a second example of how the Bayesian formulation (4) can be modified to address interesting evolutionary questions. Their method addresses the important problem of identifying the environmental factors responsible for the structuring of

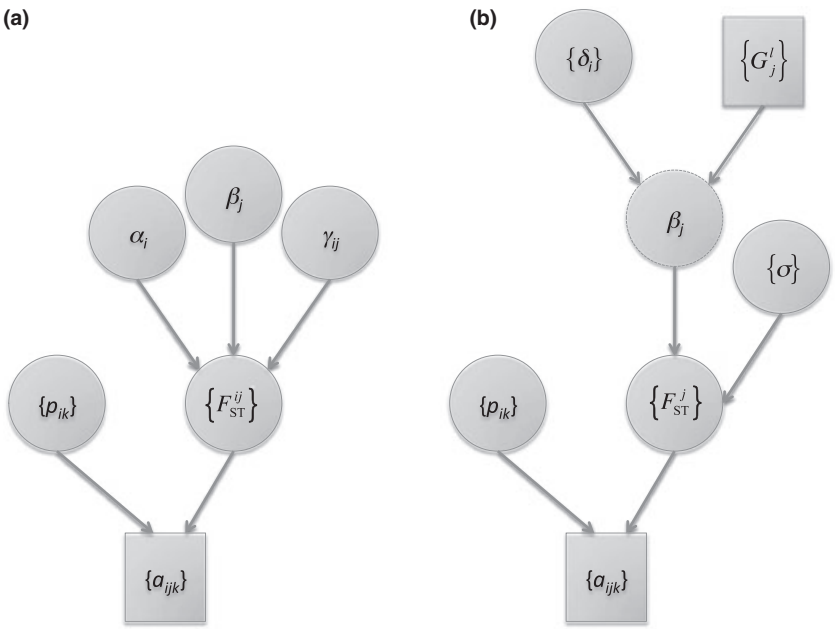

Fig. 2 Directed acyclic graph describing the Bayesian formulations of (a) the genome-scan method of Beaumont & Balding (2004) and (b) the method of Foll & Gaggiotti (2006) to identify environmental factors with an important effect on the population genetic structure. Both formulations are obtained from the Bayesian model presented in Fig. 1b by further modelling F_{ST} using hyperpriors.

neutral genetic diversity. Thus, they propose a hierarchical formulation that uses environmental data to obtain priors for F_{ST}. More specifically, they focus on the population-specific effect, β_j, described previously and decompose it into several effects because of various environmental factors using a logistic function:

$$\log\left(\frac{1}{F_{ST}^{j}} - 1\right) = \log(4N_j m_j) \sim N\left(\beta_j, \sigma^2\right),$$

with

$$\beta_j = \delta_0 + \delta_1 G_j^1 + \delta_2 G_j^2 + \cdots + \delta_l G_j^l + \cdots,$$

where, G_j^l is the observed value of the l-th environmental factor for population j, and δ_l measures the effect of environmental factor l on the genetic structure of the population. The Bayesian model (4) can now be formulated using the regression parameters and environmental factors (Fig. 2b):

$$\pi\left(\mathbf{p}, \mathbf{F}_{ST}, \boldsymbol{\delta}, \sigma^2 | \mathbf{A}\right) \propto L(\mathbf{p}, \mathbf{F}_{ST})\pi(\mathbf{p})\pi\left(\mathbf{F}_{ST} | \boldsymbol{\delta}, \sigma^2\right)\pi(\boldsymbol{\delta})\pi\left(\sigma^2\right), \quad (7)$$

where priors $\pi(\boldsymbol{\delta} = \{\delta_l\})$ are Gaussian and $\pi(\sigma^2)$ is inverse Gamma. Using this method, it is possible to investigate the effects of geographic distance, insularity, local population size, etc. and then identify the most important factors using the RJMCMC approach.

These two applications of the multinomial-Dirichlet likelihood illustrate the great flexibility of the Bayesian approach as applied to population genetics problems. Furthermore, these two methods have been used by Gaggiotti *et al.* (2009) to develop a framework that allows the identification of genome regions that may be influenced by selection and simultaneously infer the environmental factors that may be responsible for the selective pressure. It consists of first using the genome-scan method corresponding to the Bayesian formulation (6) and then carrying out analyses using the Bayesian approach described by equation (7). This second step involves an analysis that only considers those markers that have been identified as neutral and then analyses using all neutral markers and only one outlier locus. If a given environmental factor does not show an effect in the analysis with neutral markers but it does so when an outlier locus is included, we can conclude that it may be the selective force responsible for its outlier behaviour. They applied this framework to the herring population that inhabits the North and Baltic seas and concluded that salinity may be exerting a selective pressure on a gene linked to one of the microsatellite loci they studied.

Other applications of the multinomial-Dirichlet model

In a metapopulation, allele frequencies in the different populations cannot be considered as independent because either there is gene flow among them and/or they are all descended from a common ancestral population. However, some widely used methods such as assignment tests (e.g. Paetkau *et al.* 2004) make this simplifying assumption.

The multinomial-Dirichlet distribution represents an intuitive and easy to implement model for introducing a dependency among population allele frequencies. For example, Falush *et al.* (2003) extended the model-based clustering method of Pritchard *et al.* (2000) by assuming that population allele frequencies followed the multinomial-Dirichlet distribution. Following Nicholson *et al.* (2002), they described the rationale underlying their approach using an instantaneous fission model where all populations are descended from the same ancestral population but then evolve in isolation after the split. In this case, the vector $\mathbf{p}_i$ in equation (1) represents the allele frequency distribution at locus i in the ancestral population, and measures the degree of differentiation between descendant population j and the ancestral population. We note that the multinomial-Dirichlet model represents an approximation to this scenario and there are several alternative approaches that may be more appropriate to describe it at the cost of an increased complexity (see Balding 2003). Nevertheless, the use of this approach improved the performance of the genetic clustering method, in particular, when genetic differentiation among populations is weak.

Faubet & Gaggiotti (2008) also used the multinomial-Dirichlet model to introduce a dependency among allele frequency distributions in their method for estimating migration rates. In their application, the underlying model corresponds exactly to the island model described previously and local populations are related in the sense that they all receive migrants from the same migrant pool. This approach allowed them not only to better approach reality but also improve the convergence properties of the MCMC approach they used to estimate migration rates.

Simulation study

To illustrate some of the advantages of the use of population-specific F_{ST}'s, we carried out a simulation study that considers various demographic scenarios. The main objective is to show that this approach allows us to better describe the genetic structuring of the metapopulation and, in particular, easily identify populations that are somewhat distinct in terms of their genetic composition. Additionally, we explored the effect of departures from the assumption of a single migrant pool model by considering scenarios where the metapopulation was hierarchically structured into two regions each containing three populations.

We carried out the simulations using SIMCOAL (Excoffier *et al.* 2000). For all the scenarios considered, we used sample sizes of 25 diploid individuals per population and assumed that they were genotyped for 50 microsatellite loci that followed a strict stepwise mutation model with mutation rate $\mu = 10^{-5}$. Population sizes depended on the scenario being considered: the proportion of migrants under the island model was $m = 0.01$ for all populations, while for the stepping-stone scenarios was $m = 0.01$ for the four central populations and 0.005 for the two marginal populations. In the case of the hierarchically structured scenarios, we assumed that migration rates between any two populations within the same region was twice as high as that between any two populations belonging to different regions but that the overall proportion of migrants in a given population was $m = 0.01$.

The results can be more easily understood if we keep in mind that under the assumption of an island model, the F_{ST} of a population can be interpreted as the degree of differentiation between its allele frequency and that of the whole metapopulation. Thus, a small F_{ST} indicates that the genetic composition of the population resembles that of the metapopulation as a whole. A large value, on the other hand, indicates that it is different.

Figure 3 shows the results of a simulation of an island model where $N = 100$ for all populations. As expected, local F_{ST}'s are almost identical across all populations and very close to the theoretical expectation of 0.2 when $Nm = 1$. This result indicates that they all contribute equally to the genetic structure.

Figure 4 shows the results of a simulation with four small populations ($N_{1–4} = 100$) and two large populations ($N_5 = 500$ and $N_6 = 1000$) but all with the same proportion of migrants ($m = 0.01$). In this case, the local F_{ST} analysis (Fig. 4a) clearly shows that there are four populations that are genetically differentiated from the metapopulation and two other that differ little. Although it is not possible to say if these differences are because of differences in population size or in migration rates, the analysis clearly indicates that there are three types of populations that differ in the strength of genetic drift. The results of the pair-wise analysis (Fig. 4b) are much harder to interpret. High pair-wise F_{ST} values are observed between small populations, a much lower value is observed between two large populations, while intermediate values are observed for pairs including one large and one small population. Note that this analysis does not allow us to clearly identify differences between demes 5 and 6 as there is extensive overlap in the pair-wise F_{ST}'s between each one of them and the smaller populations.

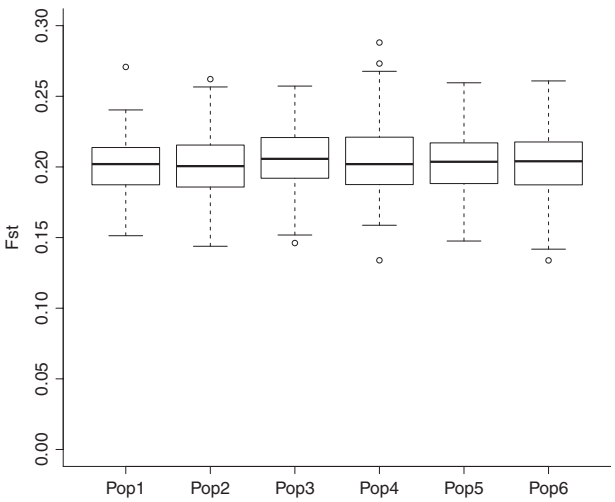

Fig. 3 Results of the simulations of an island scenario consisting in six populations with equal sizes and migration rates. $N_j = 100$ and $m_j = 0.01$ for all j.

Figure 5 shows the results of a stepping-stone model where all populations have the same size, $N_{1–6} = 100$. Under this scenario, the proportion of immigrants in the two populations at the extremes of the linear habitat (populations 1 and 6) is one half that of the other populations. The local F_{ST} analysis easily uncovers these two isolated populations but it also leads to biases in the local F_{ST} estimates of the four central populations, which should all be equal. These biases are because of the strong violation to the assumption of a unique migrant pool. The pair-wise analysis uncovers the isolation by distance pattern but it does not reveal the increased effect of genetic drift in the two marginal populations. This example highlights the complementarities of the two approaches; by looking at the results of both analyses, it is possible to obtain a comprehensive understanding of the genetic structure of the population. It should be mentioned, however, that the effect of distance could also be revealed by the population-specific F_{ST} approach by incorporating the connectivity of each local population into the analysis (e.g. Foll & Gaggiotti 2006; Kittlein & Gaggiotti 2008).

We also carried out simulations of a hierarchically structured metapopulation to determine whether violations to the single migrant pool model could introduce biases in the estimation of the population genetic structure. Figure 6 shows the result of simulations of a scenario where all populations had the same size, $N_{1–6} = 100$. In this case, the use of local F_{ST}'s correctly shows that the strength of genetic drift is the same in all populations but it fails to show any evidence for a lower migration rate between populations from different regions than between populations from the same region. Instead, it overestimates the F_{ST} of all populations by 2% to 5% (compare with the results obtained for the standard

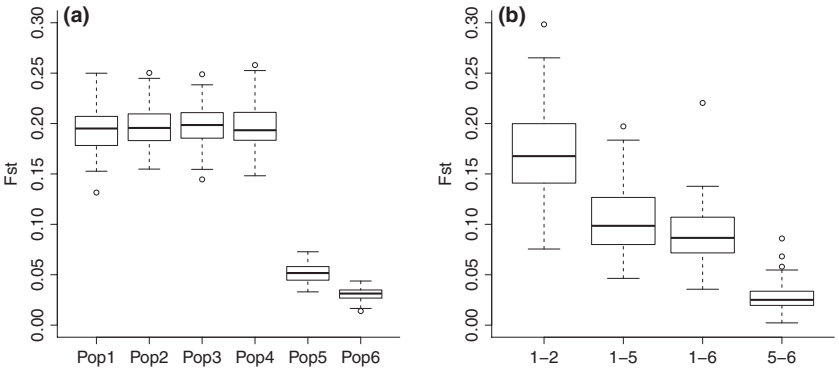

Fig. 4 Results of the simulations of an island model with six populations that differ in size: $N_{1–4} = 100$, $N_5 = 500$, $N_6 = 1000$. Migration rate is constant across populations: $m_j = 0.01$ for all j. (a) Local F_{ST} estimates obtained with GESTE, (b) pair-wise F_{ST}'s.

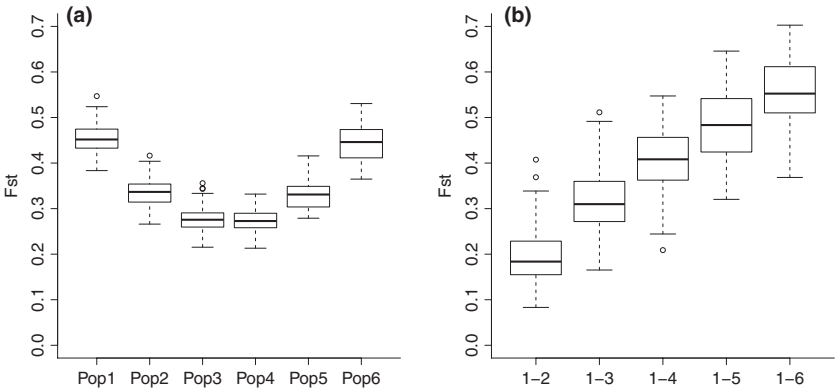

Fig. 5 Results of simulations of a stepping-stone model with populations of equal sizes: $N_j = 100$ for all j. The migration rate of the two marginal populations is lower because they only receive migrants from one neighbouring population: $m_1 = m_6 = 0.005$, $m_{2–5} = 0.01$. (a) Local F_{ST} estimates obtained with GESTE, (b) pair-wise F_{ST}'s.

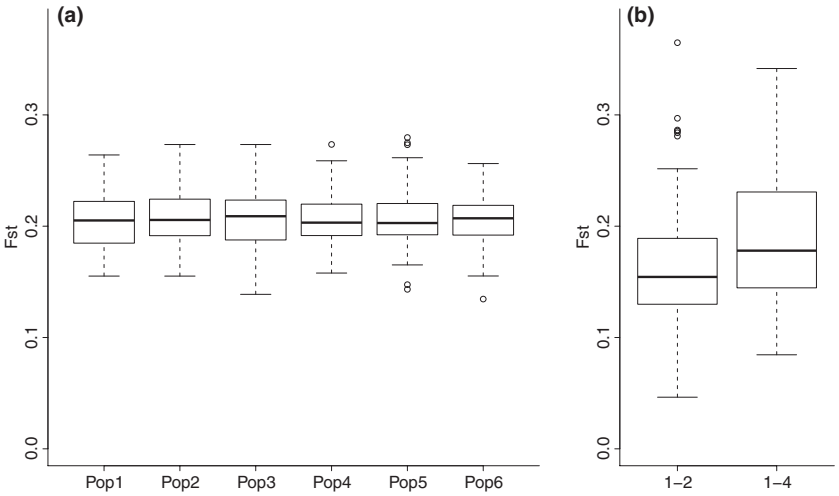

Fig. 6 Results of simulations of a hierarchical island model with populations of equal sizes ($N_j = 100$ for all j). There are two regions, each containing three local populations. Although the proportion of migrants is the same for all populations ($m_j = 0.01$ for all j), the migration rates between any two populations within the same region is twice that between any two populations belonging to different regions. (a) Local F_{ST} estimates obtained with GESTE, (b) pair-wise F_{ST}'s.

island model, Fig. 3). On the other hand, pair-wise F_{ST}'s indicate that migration rates are lower between populations from different regions.

Figure 7 shows the result of a hierarchically structured metapopulation where there is one population in each region that is larger ($N_1 = N_4 = 500$) than the others ($N_2 = N_3 = N_5 = N_6 = 100$). The use of population-specific F_{ST}'s allows us to correctly infer that genetic drift is weaker in the two large populations but it cannot identify differences in migration rates. The pair-wise analyses, on the other hand, can mislead us to believe that there is more migration between populations that belong to different regions (e.g. 1 and 4) than between populations from the same region (e.g. 1 and 2 or 2 and 3).

This limited simulation study shows that the estimation of population-specific F_{ST}'s can provide useful information about the strength of genetic drift in each population but it does not represent an alternative to the

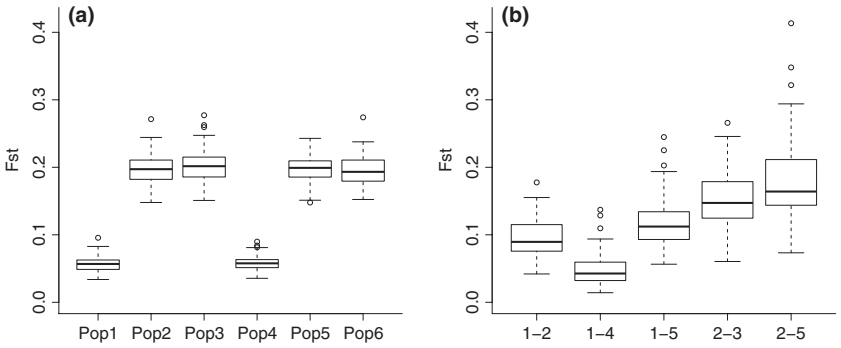

Fig. 7 Results of simulations of a hierarchical island model with populations that differ in size. There are two regions with three populations each. One of the populations in each region is larger than the other two: $N_1 = N_2 = 100$ and $N_3 = 500$ for region 1 and $N_4 = N_5 = 100$ and $N_6 = 500$ for region 2. The proportion of migrants is the same for all populations ($m_j = 0.01$ for all j), but the migration rates between any two populations within the same region is twice that between any two populations belonging to different regions. (a) Local F_{ST} estimates obtained with GESTE, (b) pair-wise F_{ST}'s.

use of pair-wise F_{ST}'s. Instead, the two approaches are complementary.

Discussion

The main objective of this review is to foster the use of a model-based method for the estimation of local F_{ST}'s. Although this approach is well known to statistical geneticists, we believe that it has remained largely underutilized by molecular ecologists. Furthermore, the applications to conservation genetics seem obvious as the identification of populations that are genetically distinct or that differ little from the migrant pool of the metapopulation constitutes important information that can help devise better management plans for endangered species. In combination with measures of allelic richness, such as those proposed by Petit *et al.* (1998), it can help evaluate the conservation value of individual populations.

Our simulation study is very limited but it does illustrate the advantages of using the *F*-model to study population genetic structure. The effect of sample size and number of loci used on the quality of the estimates was explored in some detail by Foll & Gaggiotti (2006), who show that moderate values of sample sizes (20 individuals per population) and number of loci (10) provide very reliable estimates of population-specific F_{ST}'s.

The multinomial-Dirichlet-based approach is not the only one that allows for the estimation of population-specific F_{ST}'s. Indeed, Weir & Hill (2002) extend Weir & Cockerham (1984) method of moments approach to estimate coancestry coefficients by allowing different levels of coancestry for the different populations and by allowing non-zero coancestries between populations. This approach is particularly suited for multiallelic markers, but can also be applied to Single Nucleotide Polymorphism (SNPs) when multiple loci are available. To obtain estimators that are independent of sample sizes, they propose a normal theory approach whereby the multinomial distribution describing the sampling of alleles from a single population is approximated by the multivariate normal distribution. The underlying model assumes an ancestral population from which subpopulations have descended in isolation under the same evolutionary processes. The multinomial-Dirichlet has also been used to model this situation (e.g. Falush *et al.* 2003) but it should be noted that it only represents an approximation to this scenario. Using Weir & Hill (2002) method in this case may be more appropriate. In the case of bi-allelic markers such as SNPs, it may also be more appropriate to use the approach of Nicholson *et al.* (2002), which uses a truncated normal as the prior of their Bayesian formulation and a Binomial distribution for the likelihood. To the extent of our knowledge, there is no simulation study that compares the performance of these three alternative methods of estimating population-specific F_{ST}'s under different scenarios of population structure and demographic history.

It is important to note that the *F*-model is more realistic than the standard island model but it is still a very simplistic approximation to reality. Although it allows for differences in population sizes and immigration rates, it assumes that there is a single migrant pool from which each subpopulation receives migrants each generation. This assumption is clearly violated if populations are hierarchically structured into regions or if there is isolation by distance, in which case, local F_{ST}'s are overestimated (see Figs 5 and 6). Note, however, that in this latter case, the upward bias is small, varying between 2% and 5%. Another simplifying assumption is that the Dirichlet prior used to obtain the multinomial-Dirichlet likelihood assumes equilibrium between migration and genetic drift. This assumption, however, is less critical as the multinomial-Dirichlet formula provides most of the information needed to model the gene frequencies under

any neutral-structured population model provided that the so-called 'separation-of-timescales' approximation is valid (Beaumont 2005). Under this approximation (Nordborg 1997; Wakeley & Aliacar 2001; Wakeley 2004; Wilkins 2005), sample genealogies exhibit a recent burst of coalescent events among samples taken from the same locality –the so-called 'scattering phase', followed by a more ancient historical process for the remaining ancestral lineages – the 'collecting phase'. This later phase behaves as the standard coalescent as the number of demes goes to infinity. If this approximation holds, we can use samples from independent populations to estimate the collecting-phase gene frequency without the need to model its demographic history, and thereby estimate F_{ST} from the population samples. Under the F-model, independence among local populations is obtained by conditioning on the migrant pool allele frequencies and the local F_{ST}'s. Furthermore, Wakeley (2004) shows that the 'separation-of-timescales' is valid for a broad range of specific metapopulation structures that have a large number of demes in common. This includes scenarios where local populations are grouped into regions and where there are extinction and recolonization events (Wakeley 2004).

The multinomial-Dirichlet model seems to be robust to deviations from the island model when used to identify the environmental factors responsible for the genetic structure (Foll & Gaggiotti 2006; Kittlein & Gaggiotti 2008). However, ignoring hierarchical structure leads to strong biases in the estimation of F_{ST} (Song *et al.* 2006) and to a high false positive rate in genome scans aimed at identifying outlier loci (Excoffier *et al.* 2009). These biases can be avoided by using alternative modelling strategies. For example, Fu *et al.* (2005) and Song *et al.* (2006) present Bayesian methods to estimate genetic differentiation that can take into account hierarchical structure but they are limited to biallelic loci. Also, they are less amenable to applications such as the identification of the causes of genetic structuring or the identification of outlier loci. Excoffier *et al.* (2009) present a simulation-based approach to carry out genome scans under scenarios where subpopulations are clustered into regions. The simulation study they carried out shows that this method is able to greatly reduce the false positive rate. We note, however, that these biases may become negligible if the number of local populations is very large; in which case, the 'separation-of-timescales' approximation is valid for a wide range of metapopulation scenarios. It may be useful to carry out a detailed simulation study to determine how large this number must be before the biases become negligible.

There is still much work to be performed to introduce more reality into statistical genetics methods in general. In the particular case of the multinomial-Dirichlet approach, we note that we are currently working on an extension of the *F*-model that will take into account hierarchical structuring and will overcome the limitations mentioned previously (Foll *et al.*, in preparation). Ultimately, however, the consideration of very complex situations can only be achieved using approximate methods such as approximate Bayesian computation approach as performed by Estoup *et al.* (2010) in an article published in this same issue.

We hope that this review and the availability of GESTE, a user-friendly program to carry out genetic structure inferences, will help popularize the use of local F_{ST}s within the molecular ecology community.

Acknowledgement

The comments of three anonymous reviewers have allowed us to greatly improve the original version of the manuscript.

References

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

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Balding DJ, Greenhalgh M, Nichols RA (1996) Population genetics of STR loci in Caucasians. *International Journal of Legal Medicine*, **108**, 300–305.

Balding DJ, Carothers AD, Marchini JL *et al.* (2002) Discussion on the meeting on 'Statistical modelling and analysis of genetic data'. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 737–775.

Beaumont MA (2005) Adaptation and speciation: what can F-st tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Chakrabory R, Danker-Hopfe H (1991) Analysis of population structure: A comparative study of different estimators fo Wright's fixation indices. in: *Handbook of Statistics* (eds Rao CR & Chakraborty R), pp. 203–254. Elsevier Science Publishers, Amsterdam, The Netherlands.

Estoup A, Baird SJE, Ray N, Currat M, Cornuet J-M, Santos F, Beaumont MA, Excoffier L (2010) Combining genetic, historical and geographic data to reconstruct the dynamics of bioinvasions: application to the cane toad Bufo marinus. *Molecular Ecology Resources*, **10**, 886–901.

Excoffier L (2007) Analysis of population subdivision. In: *Handbook of Statistical Genetics* (eds Balding DJ, Bishop MJ & Cannings C), pp. 980–1020. John WIley & Sons, Chichester, England, Hoboken, NJ.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in

interconnected populations with arbitrary demography. *Journal of Heredity*, **91**, 506–510.
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
Faubet P, Gaggiotti OE (2008) A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics*, **178**, 1491–1504.
Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of Populations. *Genetics*, **174**, 875–891.
Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.
Fu RW, Dey DK, Holsinger KE (2005) Bayesian models for the analysis of genetic structure when populations are correlated. *Bioinformatics*, **21**, 1516–1529.
Gaggiotti OE, Brooks SP, Amos W, Harwood J (2004) Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Molecular Ecology*, **13**, 811–825.
Gaggiotti OE, Bekkevold D, Jorgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, **63**, 2939–2951.
Holsinger KE (1999) Analysis of genetic diversity in geographically structured populations: A Bayesian perspective. *Hereditas*, **130**, 245–255.
Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F-ST. *Nature Reviews Genetics*, **10**, 639–650.
Johnson NL, Kotz S, Balakrishnan N (1997) *Discrete Multivariate Distributions*. John Wiley & Sons, New York.
Kittlein MJ, Gaggiotti OE (2008) Interactions between environmental factors can hide isolation by distance patterns: a case study of Ctenomys rionegrensis in Uruguay. *Proceedings of the Royal Society B-Biological Sciences*, **275**, 2633–2638.
Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceeding of the National Academy of Sciences of the United States of America*, **70**, 3321–3323.
Nicholson G, Smith AV, Jonsson F *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 695–715.
Nordborg M (1997) Structured coalescent processes on different time scales. *Genetics*, **146**, 1501–1514.
Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology*, **13**, 55–65.
Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, **12**, 844–855.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
Rannala B, Hartigan JA (1995) Identity by descent in island-mainland populations. *Genetics*, **139**, 429–437.
Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research*, **67**, 147–158.
Riebler A, Held L, Stephan W (2008) Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, **178**, 1817–1829.
Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.
Song S, Dey DK, Holsinger KE (2006) Differentiation among populations with migration, mutation, and drift: Implications for genetic inference. *Evolution*, **60**, 1–12.
Wakeley J (2004) Metapopulation and coalescent theory. In: *Ecology, Genetics and Evolution of Metapopulations* (eds Hanski I & Gaggiotti OE), pp. 175–198. Elsevier-Academic Press, San Diego, California.
Wakeley J, Aliacar N (2001) Gene genealories in a metapopulation. *Genetics*, **159**, 893–905.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics*, **36**, 721–750.
Wilkins JF (2005) A separation-of-timescales approach to the coalescent in a continuous population. *Genetics*, **168**, 2227–2244.
Wright S (1931) Evolution in mendelian populations. *Genetics*, **16**, 97–159.
Wright S (1949) Adaptation and selection. In: *Genetics, Paleontology and Evolution* (eds Jepsen GL, Simpson GG & Mayr E). Princeton University Press, Princeton.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.