Estimating Population-Level Coancestry Coefficients by an Admixture F Model

Markku Karhunen[1] and Otso Ovaskainen
Department of Biosciences, University of Helsinki, FI-00014 Helsinki, Finland

ABSTRACT In this article, we develop an admixture F model (AFM) for the estimation of population-level coancestry coefficients from neutral molecular markers. In contrast to the previously published F model, the AFM enables disentangling small population size and lack of migration as causes of genetic differentiation behind a given level of F_{ST}. We develop a Bayesian estimation scheme for fitting the AFM to multiallelic data acquired from a number of local populations. We demonstrate the performance of the AFM, using simulated data sets and real data on ninespine sticklebacks (*Pungitius pungitius*) and common shrews (*Sorex araneus*). The results show that the parameterization of the AFM conveys more information about the evolutionary history than a simple summary parameter such as F_{ST}. The methods are implemented in the R package RAFM.

IN the fields of animal and plant breeding, coancestry coefficients are often used as measures of relatedness between individuals (Bink *et al.* 2008). For example, in a noninbred population the coancestry between full-sibs or between a parent and an offspring is $\frac{1}{4}$, and the coancestry between half-sibs is $\frac{1}{8}$ (Lynch and Walsh 1998). Coancestry is the same as probability of identity by descent (IBD) at the limit of a low mutation rate and given a noninbred ancestral population. Two genes are said to be identical by descent if and only if they have not mutated since the most recent common ancestor.

Individual-level coancestry coefficients (or probabilities of IBD) are useful in gene mapping, because they tell how much the genomes of two individuals are expected to resemble each other; *i.e.*, they summarize the expected level of genetic similarity. In analogy, population-level coancestry coefficients can be used as measures of relatedness between local populations, and they can be combined with phenotypic data to detect signals of selection in quantitative traits, as opposed to those caused by random drift (Merilä and Crnokrak 2001; Mckay and Latta 2002; Ovaskainen *et al.* 2011).

Coancestry coefficients can be calculated directly, if pedigree information is available, but their estimation for natural populations is often challenging. One approach for doing so is to use the link between coancestry coefficients and coalescence times (Rousset 2004). Coalescence time distributions can be solved, at least numerically, for a population that is in a stationary state, assuming that the demographic parameters are known (Bahlo and Griffiths 2001). However, in the context of evolutionary ecology of natural populations, this is rarely the case, as there is often limited direct information on demographic history, and it can be unrealistic to assume any kind of stationarity. Instead, a common approach is to infer the demographic history using neutral molecular markers genotyped from the present generation. One statistical framework for estimating coancestry coefficients in this way is given by the F model (Falush *et al.* 2003; Gaggiotti and Foll 2010). However, this approach suffers from the structural limitation that the subpopulations are assumed to have radiated independently from the ancestral population, so that there has been no recent gene flow. Consequently, the F model cannot account for limited gene flow and small population size as alternative sources of genetic differentiation (Gaggiotti and Foll 2010).

In animal and plant breeding, a number of alternative methods have been developed for estimating coancestry coefficients from molecular marker data for pairs of individuals. Bink *et al.* (2008) survey seven such methods, concluding that the surveyed estimators have poor statistical properties, except in the special case that the allele frequencies are known

doi: 10.1534/genetics.112.140871
Manuscript received March 31, 2012; accepted for publication June 26, 2012
Supporting information is available online at http://www.genetics.org/content/early/2012/07/10/genetics.112.140871/suppl/DC1.
[1] Corresponding author: Department of Biosciences, PO Box 65 (Viikinkaari 1), University of Helsinki, FI-00014 Helsinki, Finland. E-mail: markku.karhunen@helsinki.fi

for a hypothetical reference population. Furthermore, as Fernandez and Toro (2006) point out, many of these estimators have undesired mathematical properties; *e.g.*, they may yield logically incompatible estimates for different pairs of individuals. Software by Maenhout *et al.* (2009) removes some of these flaws by *post hoc* modification of the parameter estimates.

In this article, we focus on the case where neutral genotypic data are available for a set of subpopulations, and the problem is to infer the matrix of coancestry coefficients among these local populations. We model the demographic histories of the subpopulations by an admixture of evolutionary independent lineages, thus extending the F model in a way that relaxes the structural assumption noted above. We use an admixture of independent lineages as a phenomenological model for the evolutionary history of a metapopulation where local populations experience a limited level of gene flow. Apart from Gaggiotti and Foll (2010), our method is also a generalization of that of Fu *et al.* (2005), because we consider multiallelic loci and a more general population structure than the case of clustered subpopulations. With these extensions, our model contains both gene flow and pure random drift as factors influencing the level of differentiation. Contrary to the "pairwise methods" used in animal and plant breeding, both the original F model and our model permit writing the likelihood of individual-level data directly as a function of population-level coancestry coefficients. In the following, we first introduce the modeling approach and then its Bayesian parameterization that we have implemented in the R-package RAFM, and finally we illustrate the modeling approach with the help of simulated and real data.

The Modeling Approach

Coefficients of coancestry

Our main interest is in the estimation of population-level coefficients of coancestry, denoted by θ^{P}_{AB} for a pair of populations (A, B). We define θ^{P}_{AB} as the average coancestry between the subpopulations,

$$\theta^{P}_{AB} = \frac{1}{n_A n_B} \sum_{i \in A, i \in B} \theta_{ii'}, \tag{1}$$

where $\theta_{ii'}$ is the coancestry coefficient of individuals i and i', and n_A is the number of individuals in population A. We note that the definition of Equation 1 allows for the possibility that the level of coancestry is not identical for all pairs of individuals $\theta_{ii'}$ with $i \in A$ and $i' \in B$. *A priori*, in lack of this information, $\theta_{ii'}$ is assumed to depend only on the populations A and B, and thus it can be used interchangeably with θ^{P}_{AB} for calculating the covariance of allelic states as detailed in Supporting Information, File S1.

We follow Rousset (2004) and call two gene copies IBD if they originate from the same ancestral copy and are identical by state; *i.e.*, they have not mutated since their divergence. The coancestry coefficients and the probabilities of IBD for neutral loci are often used interchangeably, but they have a slight difference (we denote the latter by $\theta^{c}_{ii'}$ and θ^{Pc}_{AB} for the individual and subpopulation levels, respectively). The probability of IBD can be written by using the coalescence time distribution for two gene copies in populations A and B as (Rousset 2004), *e.g.*, for a model with discrete generations:

$$\theta^{Pc}_{AB} = \sum_{t=1}^{\infty} C_{AB,t} (1-\mu)^{2t}. \tag{2}$$

In this equation, $C_{AB,t}$ is the probability that the two gene copies coalesce exactly t generations before present, and μ is the per-locus per-generation probability of mutation. Bahlo and Griffiths (2001) derive formulas that allow the numerical computation of θ^{Pc}_{AB}, assuming that the migration rates between the subpopulations and their relative sizes are known. These formulas enable estimating θ^{Pc}_{AB} from demographic parameters, but this approach typically assumes that population dynamics have remained stationary over a long period of time (Bahlo and Griffiths 2001; Wilkinson-Herbots 2003; Wilkinson-Herbots and Ettridge 2004; Bhattacharya *et al.* 2007).

Sometimes the biological context is such that there has been a major perturbation, such as the last ice age, after which the subpopulations have diverged from a common ancestral pool. In this case, instead of assuming stationarity, it is more natural to consider a finite population history of T generations. In this case,

$$\theta^{Pc}_{AB} \sum_{t=1}^{T} C_{AB,t} (1-\mu)^{2t} \approx \sum_{t=1}^{T} C_{AB,t} = E\left[\theta^{P}_{AB}\right], \tag{3}$$

where the expectation is taken over the distribution of pedigrees generated by the demographic model. The approximation is justified if the mutation rate is low compared to the number of generations.

The relationship between coancestry and F_{ST}

F_{ST} is one of the most widely used statistics in population genetics, and it is routinely used as a measure of genetic differentiation (Rousset 2002, 2004; Whitlock 2011). Depending on the definition of θ, F_{ST} can be defined through coancestry, probability of IBD, or probability of identity by state as

$$F_{\mathrm{ST}} = \frac{\theta^{W} - \theta^{B}}{1 - \theta^{B}}, \tag{4}$$

where

$$\theta^{W} = \frac{1}{n_P} \sum_{A=1}^{nP} \theta^{P}_{AA}, \quad \theta^{B} = \frac{1}{n_P^2 - n_P} \sum_{B \neq A} \theta^{P}_{AB}, \tag{5}$$

and n_P is the number of populations. In this article, we define F_{ST} through population-level coancestry. In Equation 5,

θ^W is the average coancestry within subpopulations, and θ^B is the average coancestry between subpopulations. In line with the coalescent-based definition of F_{ST} (Rousset 2004), we do not weight the averages, *e.g.*, by the sizes of the local populations. We are chiefly interested in estimating the coancestry coefficients and investigating the properties of the admixture F model (AFM), but we also report F_{ST} (defined through the coancestry-based variant of Equations 4 and 5) estimates because of the centrality of F_{ST} in the literature.

The AFM

In this section, we extend the F model (Falush *et al.* 2003; Gaggiotti and Foll 2010) to an AFM that allows for gene flow among the local populations. As is the case with the original F model, we assume that the local populations are derived from a common ancestral population and consider the limit of a small mutation rate, *i.e.*, the situation that relates to Equation 3.

Denoting the frequency of allele u at locus j in the ancestral generation by q_{ju}, the expectation and variance of the allele frequency in population A can be written as

$$E\left[p_{Aju}\right] = q_{ju},$$

$$\mathrm{Var}\left[p_{Aju}\right] = \left(q_{ju} - q_{ju}^2\right)\phi, \qquad (6)$$

where ϕ is a factor that depends on the demographic model (Lynch and Walsh 1998). For an isolated population of a constant effective size,

$$\phi = 1 - \left[1 - \frac{1}{2N_{\mathrm{e}}}\right]^T \qquad (7)$$

(Lynch and Walsh 1998). A convenient distributional form that satisfies the above is

$$\boldsymbol{p}_{Aj} \sim \mathrm{Dirichlet}\left(a\boldsymbol{q}_j\right), \qquad (8)$$

where

$$a = \left(1 - \left[1 - \frac{1}{2N_{\mathrm{e}}}\right]^T\right)^{-1} - 1 \qquad (9)$$

in absence of mutation. By Equation 9, a small value of a corresponds to a small effective population size or a large number of generations T, both of which imply a high amount of random genetic drift. The Dirichlet distribution is just a convenient approximation for the distribution of allele frequencies under pure random drift, as their true distribution is difficult to implement in a statistical model (see File S2). Also the truncated normal distribution is often used to approximate this distribution (Nicholson *et al.* 2002; Balding 2003; Coop *et al.* 2010). However, the truncated normal distribution is more difficult to adapt to the multiallelic case than the Dirichlet distribution as the frequency distribution is constrained by the condition $\sum_{u=1}^{n_j} p_{ju} = 1$. For a discussion on the relative accuracy of the Dirichlet and truncated-normal approximations, see File S2 and Figure S1.

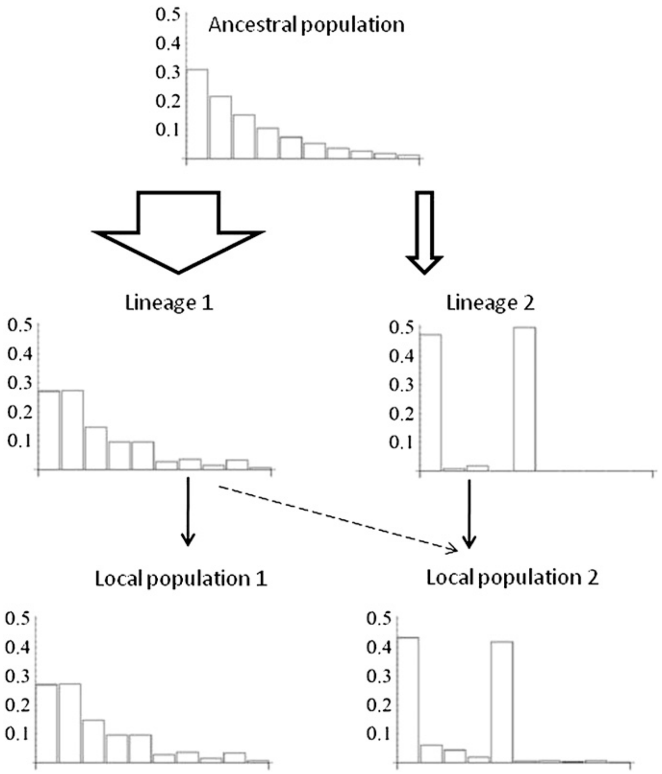


Figure 1 Schematic presentation of the admixture F model (AFM), in which subpopulations are constructed as admixtures of independent lineages. The histograms represent allele frequencies in a particular locus in the ancestral generation, in two independent lineages, and in two present subpopulations. In this example, lineage 1 has been subject to little drift (parameter value a_1 = 100). In contrast, only two alleles remain at high frequency in lineage 2 as a result of much drift (a_2 = 0.5). Population 1 is identical to lineage 1 (k_{11} = 1, k_{12} = 0). Population 2 is mainly derived from lineage 2, but has received some gene flow from lineage 1 (k_{21} = 0.1, k_{22} = 0.9). These parameter values give population-level coancestry coefficients $\theta^P_{11} = 0.010$, $\theta^P_{12} = 0.002$, and $\theta^P_{22} = 0.427$, yielding F_{ST} = 0.22.

To extend the model for n_P subpopulations that may have experienced gene flow since their divergence from a common ancestral population, we assume an admixture of n_ε evolutionary independent lineages (Figure 1). The allele frequencies in each lineage are distributed as in Equation 8; *i.e.*, we assume for locus j and lineage k,

$$\mathbf{z}_{kj} \sim \mathrm{Dirichlet}\left(a_k\mathbf{q}_j\right), \qquad (10)$$

where a_k measures the amount of drift experienced by this lineage. The allele frequencies in locus j in local population A are defined as a mixture the lineage-specific frequencies, namely

$$\mathbf{p}_{Aj} = \sum_{k=1}^{n_\varepsilon} k_{Ak}\mathbf{z}_{kj}. \qquad (11)$$

Table 1 List of main parameters and symbols

Dimensions	
No. distinct alleles in locus *j*	n_j
No. loci	n_L
No. lineages	n_ε
No. subpopulations	n_P
Coalescent theory	
Probability of IBD for two gene copies in populations *A* and *B*	θ_{AB}^{Pc}
Probability that two gene copies from populations *A* and *B* have coalesced exactly *t* generations before present	$C_{AB,t}$
Time since population divergence	T
Per-generation per-locus rate of mutation	μ
Per-capita probability of migration	m
Coancestry coefficients	
Coancestry among subpopulations	$n_P \times n_P$ matrix $\boldsymbol{\theta}^P$ with elements θ_{AB}^P
Mean within-population coancestry	$\theta^W = \frac{1}{n_P} \sum_{A \in \mathrm{P}} \theta_{AA}^P$
Mean between-population coancestry	$\theta^B = \frac{1}{n_P^2 - n_P} \sum_{B \neq AA, B \in P} \theta_{AB}^P$
Allele frequencies	
Allele frequencies in the ancestral generation	$\mathbf{q}_j = (q_{ju});\ u = 1, \ldots, n_j$ $\mathbf{q} = (\mathbf{q}_j);\ j = 1, \ldots, n_L$
Allele frequencies in lineages	$\mathbf{z}_{kj} = (z_{kju});\ u = 1, \ldots, n_j$ $\mathbf{z}_k = (\mathbf{z}_{kj});\ j = 1, \ldots, n_L$ $\mathbf{z} = (\mathbf{z}_k);\ k = 1, \ldots, n_L$
Allele frequencies in subpopulations	$\mathbf{p}_{Aj} = (p_{Aju});\ u = 1, \ldots, n_j$ $\mathbf{p}_A = (\mathbf{p}_{Aj});\ j = 1, \ldots, n_L$ $\mathbf{p} = (\mathbf{p}_A);\ A = 1, \ldots, n_P$
Parameters measuring evolutionary history	
Lineage loadings	$n_P \times n_L$ matrix $\mathbf{k}$ with element k_{Ak}
Genetic drift	$\mathbf{a} = (a_k);\ k = 1, \ldots, n_L$
Identity by state	
Indicator variable for the allele copy *k* in locus *j* of individual *i* being of the allelic type *u*	x_{ijku}
Data, *i.e.*, observed allele counts on the sample A_s of individuals originating from subpopulation *A*	$x_{Aju} = \sum_{i \in A_s} \sum_{k=1,2} x_{ijku}$ $\mathbf{x}_{Aj} = (x_{Aju});\ u = 1, \ldots, n_j$ $\mathbf{x}_A = (\mathbf{x}_{Aj});\ j = 1, \ldots, n_L$ $\mathbf{x} = (\mathbf{x}_A);\ A = 1, \ldots, n_P$

We constrain the lineage loadings k_{Ak} to sum up to unity over the lineages, $\sum_{k=1}^{n_L} k_{Ak} = 1$, implying that vector $\mathbf{p}_{Aj}$ is a proper frequency distribution. Setting the lineage-loading matrix to the identity matrix yields the special case of fully independent demes (the F model of Falush *et al.* 2003). Technically, our construction is analogous to factor analysis (Gorsuch 1983), with lineages as factors and lineage loadings k_{Ak} as factor loadings.

A convenient property of the AFM is that the subpopulation-level coancestry coefficients depend on the model parameters in a very simple way (Table 1). As shown in File S1,

$$\theta_{AB}^P = \sum_{k=1}^{n_\varepsilon} \frac{k_{Ak} k_{Bk}}{a_k + 1}. \tag{12}$$

Thus, after fitting the AFM to data it is straightforward to obtain an estimate of the matrix of population-to-population coancestry coefficients. By construction, this matrix will be always positive definite, avoiding the logical problems from which some of the earlier methods suffered (Fernandez and Toro 2006).

Assuming no genetic structure within subpopulations, *i.e.*, a random distribution of alleles among and within individuals, the genotype of each individual in subpopulation A is a multinomial random variable, $x_{ij} \sim \text{Multinomial}(2, \mathbf{p}_{Aj})$. Notably, inbreeding due to a small population size is represented by a high intrapopulation coancestry θ_{AA}^P, whereas an increased level of inbreeding due to assortative mating could be added to the model by assuming a dependency between the allelic states of the two gene copies within an individual, but we do not consider that in this article.

Parameter estimation with Bayesian inference

To parameterize the AFM with Bayesian inference, prior distributions need to be defined for the primary parameters q, α, and k. We assume the distributional forms

$$\mathbf{q}_j \sim \text{Dirichlet}\left(\boldsymbol{\beta}_j^q\right),$$

$$\log\ a_k \sim N\left(\mu_a, \sigma_a^2\right),$$

$$\mathrm{k}_A \sim \text{Dirichlet}\left(\boldsymbol{\beta}_A^x\right),$$

mainly for the sake of mathematical convenience. Indexes j, k, and A refer to loci, lineages, and subpopulations, respectively. In the case studies below, we assume the values $\boldsymbol{\beta}_j^q = \mathbf{1}_{n_j}$, $\mu_a = 2$, $\sigma_a^2 = 2$. We set the number of lineages equal to the number of subpopulations and assume that lineage A makes the dominant contribution to subpopulation A, *i.e.*, that the matrix $\mathbf{k}$ is diagonally dominant. To do so, we let

$$\beta_{AA}^k = 0.8 n_P, \quad \text{and} \quad \beta_{Ak}^k = \frac{0.2 n_P}{n_\varepsilon - 1} \quad \text{for} \quad k \neq A, \tag{13}$$

and truncate the prior by the requirement that $k_{AA}^k > k_{Ak}^k$ for all $k \neq A$. This specification links each population with a particular lineage by assuming that lineage A makes a dominating contribution to population A. It also ensures that label switching is not possible, thus improving the mixing of the Markov chain Monte Carlo (MCMC) algorithm (Gelman and Carlin 2004).

The number of alleles (n_j) in locus j in the ancestral generation is generally unknown, as some alleles may have disappeared after the lineages have diverged or are not present in the sampled individuals. Due to the aggregation property of the Dirichlet distribution, all of the unobserved alleles can be binned into a single unobserved class. Thus, we define n_j as the number of distinct alleles observed in locus j plus one.

The directed acyclic graph that illustrates the link from the primary parameters ($\mathbf{k}$, $\boldsymbol{\alpha}$, $\mathbf{q}$) through the derived parameters ($\mathbf{z}$, $\mathbf{p}$) to the data $\mathbf{x}$ is shown in Figure 2. Given the data $\mathbf{x}$, the posterior density can be decomposed as

$$\pi(\mathbf{k}, \boldsymbol{\alpha}, \mathbf{q} \mid \mathbf{x}) \propto \pi(\mathbf{x} \mid \mathbf{z}, \mathbf{k}) \pi(\mathbf{z} \mid \boldsymbol{\alpha}, \mathbf{q}) \pi(\mathbf{k}) \pi(\boldsymbol{\alpha}) \pi(\mathbf{q}), \tag{14}$$

with the distributional form of each factor being specified above. As noted above, the coancestry coefficients are not directly involved in the estimation procedure, but their posterior distribution is determined by that of ($\mathbf{k}$, $\boldsymbol{\alpha}$) (Equation 12). We use the adaptive random-walk Metropolis–Hastings algorithm of Ovaskainen *et al.* (2008) to sample the posterior density $\pi(\mathbf{k}, \boldsymbol{\alpha}, \mathbf{q} \mid \mathbf{x})$. More details of the algorithm can be found in File S3, and it is implemented in the R package RAFM.

Numerical Examples

We tested the performance of the method described above with two kinds of simulated data: data generated by the AFM itself and data generated through individual-based pedigrees that we in turn generated by a demographic model with continuous migration among subpopulations. The first type of data was used to investigate the performance of the estimation scheme in the ideal case that the data follow the structural assumptions of the model. The second type of data was used to examine whether a mixture of independent lineages can yield a good approximation of a more realistic demography in the sense of providing an accurate estimate of the matrix $\boldsymbol{\theta}^P$ and whether the parameters $\boldsymbol{\alpha}$ and $\mathbf{k}$ correlate with the demographic parameters in an intuitive way.

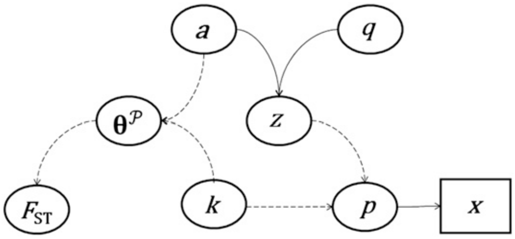


Figure 2 A directed acyclic graph (DAG) describing the dependencies among model parameters and data. Solid arrows imply probabilistic links and dashed arrows deterministic relationships. The process that is assumed to have generated the genotype data ($\mathbf{x}$) involves the ancestral allele frequencies ($\mathbf{q}$), the amount of genetic drift experienced by the lineages ($\mathbf{a}$), the allele frequencies in the lineages ($\mathbf{z}$), and the lineage loadings, *i.e.*, the contributions of lineages to the local populations ($\mathbf{k}$). Derived parameters include allele frequencies in the subpopulations ($\mathbf{p}$) and the matrix of population-level coancestry coefficients ($\boldsymbol{\theta}^P$) from which F_{ST} can be computed.

Case studies with data generated by the AFM

First, we considered $n_P = 2$ populations A and B and assumed the parameter values $\mathbf{k}$ = (0.9, 0.1; 0.1, 0.9) and $\boldsymbol{\alpha}$ = (2.7, 2.7), which leads to $\boldsymbol{\theta}^P = (0.22,\ 0.05;\ 0.05,\ 0.22)$ and consequently $F_{ST} = 0.18$. As a default case, we assumed that $n_A = n_B = 100$ individuals from each population were genotyped for $n_L = 16$ loci, each having $n_j = 4$ allelic variants that were equally common in the ancestral generation. To test the dependency of parameter estimates on sample size, we varied each of these parameters in turn, considering $n_A = n_B = 10, 100, 1000$; $n_L = 8,\ 16,\ 32$; and $n_j = 2, 4, 8$. Figure 3 shows how the accuracy of the estimated F_{ST} value increases with sample size. As expected from earlier research (Gaggiotti and Foll 2010; Wang and Hey 2010), increasing the number of loci improves the accuracy much more rapidly than increasing the number of individuals. Analogously, increasing the number of alleles per each locus, *i.e.*, increasing the level of polymorphism, brings more resolution to the data, and thus it also rapidly improves parameter estimates. Contrary to the case studies of Jost (2008), but consistent with the fact that F_{ST} is defined through coancestry, the estimates of F_{ST} do not decrease when the polymorphism of marker loci increases (Figure 3A).

To test whether local drift and lack of gene flow could be separated as alternative causes of genetic differentiation, we repeated the above (with the default sample size) with the off-diagonal value of $\mathbf{k}$ set to 0.05, 0.15, 0.25 and the value of $\boldsymbol{\alpha}$ adjusted so that $F_{ST} = 0.18$ in all cases (Figure 4). Note that gene flow sets an upper limit to population differentiation: given a value of gene flow (*i.e.*, off-diagonal of $\mathbf{k}$), there is an upper limit to F_{ST}, namely the one produced by $\boldsymbol{\alpha}$ = (0, 0). While the separation of gene flow and migration is not possible in the standard F model (Gaggiotti and Foll 2010), Figure 4A shows that the parameters $\mathbf{k}$ and $\boldsymbol{\alpha}$ are

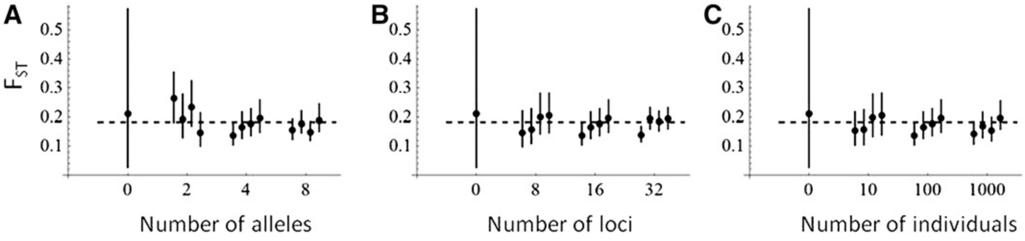


Figure 3 Accuracy of parameter estimates increases with allelic polymorphism and sample size. The solid circles with the error bars show the estimate (posterior median and 95% central credibility interval) of F_{ST} obtained by fitting the AFM to simulated data generated by the AFM. The default values of four alleles, 16 loci, and 100 individuals are assumed except for the parameter that is varied in each panel: level of polymorphism (A), number of loci (B), and number of individuals sampled from each subpopulation (C). The true value of $F_{ST} = 0.18$ is indicated by the dashed line, and the cases with sample size 0 show the prior distribution. For parameter values used in generating the data, see *Case studies with data generated by the AFM* in the main text.

identifiable in the AFM, if sufficient data are available. As a consequence, it is possible to estimate a full matrix $\boldsymbol{\theta}^P$ (Figure 4B), not only the summary parameter F_{ST}.

Case studies with an individual-based model

We constructed pedigrees for $n_P = 2$ subpopulations with nonoverlapping, constant-size generations consisting of equal numbers of males and females. For each individual in the ancestral population, we randomized the two allele copies for each locus assuming four allelic variants with equal frequency $q_{ju} = 0.25$. The two parents of each individual in the subsequent generations were randomized (independently of each other) with probability $1 - m$ among the individuals of the same subpopulation and with probability m among the individuals of the other subpopulation (thus implying a per-capita migration rate m). We modeled diploidic inheritance for 32 unlinked loci. To vary the level of gene flow and genetic drift, we considered three scenarios, in each of which the two subpopulations had diverged 50 generations ago. In the baseline scenario 1, we assumed 200 individuals per population and $m = 0.001$. In scenario 2, we increased the amount of drift (and thus also F_{ST}) by assuming 50 individuals per subpopulation. Finally, scenario 3 differed from the baseline scenario 1 by having a higher amount of gene flow, $m = 0.02$. As the purpose of this simulation study was to examine whether the AFM is able to approximate individual-based pedigrees rather than to test its statistical power (which we demonstrate in Figures 3 and 4), we assumed that large data sets were available, *i.e.*, 100 individuals per subpopulation genotyped for 32 loci (even for the smaller subpopulations), each having four allelic variants in the ancestral generation. We created four replicate data sets for each of the scenarios 1–3.

Figure 5 shows that the AFM can mimic individual-based pedigrees in the sense that the parameters that measure gene flow ($\mathbf{k}$) and genetic drift ($\boldsymbol{\alpha}$) vary in line with the individual-level parameters of the three demographic scenarios. Increasing local population size decreases $\boldsymbol{\alpha}$, and increasing gene flow increases the off-diagonal elements of $\mathbf{k}$. Figure 5B shows that our approach performs well also for estimating F_{ST} from the individual-based data, although there is a slight bias upward for scenario 2 with a high amount of drift. Here the true values of the coancestry coefficients were computed from the simulated pedigree, using first the standard recursive relationships (File S1) and then averaging the individual-level coancestries over the natural subpopulations (not the genotyped individuals), according to Equation 1. For comparison, the Weir–Cockerham estimator (Weir and Cockerham 1984), implemented in FSTAT (Goudet 1995), gives very similar results (Figure 5B). Thus, the novelty of our approach is not in estimation of F_{ST}, but in separating gene flow and genetic drift as causal factors behind the observed level of differentiation. This separation is needed to estimate the full coancestry matrix $\boldsymbol{\theta}^P$, which in turn is needed, *e.g.*, for detecting signals of natural selection in quantitative-genetic studies (Ovaskainen *et al.* 2011).

Case studies with real data

Here we illustrate our model's output with two natural data sets. Both of these data sets are included in the R package RAFM (Karhunen 2012). The first data set consists of 183 ninespine sticklebacks genotyped for 12 microsatellite markers (a subset of data used by Shikano *et al.* 2010), and it comprises four populations: Baltic Sea (60°13′N, 25° 11′E), White Sea (66°18′N, 33°25′E), pond Bynästjärnen in Sweden (64°27′N, 19°26′E), and pond Pyöreälampi in Finland (66°15′N, 29°26′E). The pond populations are likely to have experienced a very high amount of drift, and all populations are likely to have remained reproductively isolated from each other since the last ice age (Shikano *et al.* 2010). Thus, the demographic assumptions of Equation 3 and the AFM are at least approximately in line with the biological context.

For the ninespine sticklebacks, the median (95% credibility interval) of F_{ST} given by the AFM was $F_{ST} = 0.34$ (0.31–0.37). The Weir–Cockerham estimator yielded a higher estimate, the point estimate (95% confidence interval) being $F_{ST} = 0.50$ (0.44–0.55). The median estimates of the within-population coancestries θ^P_{AA} were 0.02, 0.10, 0.57, and 0.68 for the White Sea, Baltic Sea, Swedish pond, and Finnish pond populations, respectively. These numbers may be compared to population-specific F_{ST} values, *i.e.*, θ_i of Weir and Hill (2002), calculated from pairwise F_{ST} values given by FSTAT (Goudet 1995): 0.13, 0.09, 0.77, and 0.98 in the same order. Thus, as expected intuitively, the pond populations have experienced much more drift than the sea populations. In our analysis, the White Sea population is more diverse than the Baltic Sea population, which may

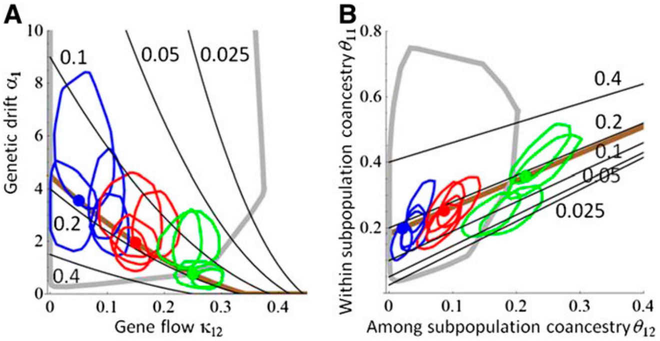


Figure 4 Separation of genetic drift and gene flow as causes of genetic differentiation. In the simulated case study consisting of two identical populations, F_{ST} and the subpopulation-level coancestry coefficients depend on the parameters $a_1 = a_2$ (measuring genetic drift) and $k_{12} = k_{21}$ (measuring gene flow). The black lines show isoclines of F_{ST} in (k_{12}, a_1) space (A) and in (θ_{12},θ_{11}) space (B). The brown line shows the isocline of F_{ST} = 0.18 corresponding to the true value in all three simulated scenarios, and the solid circles show the true parameter values for each of the scenarios. The lines show the parameter estimates of the fitted models, measured by 75% polytope quantiles of the posterior distributions. The thick gray lines show the 75% polytope quantiles for the prior distribution.

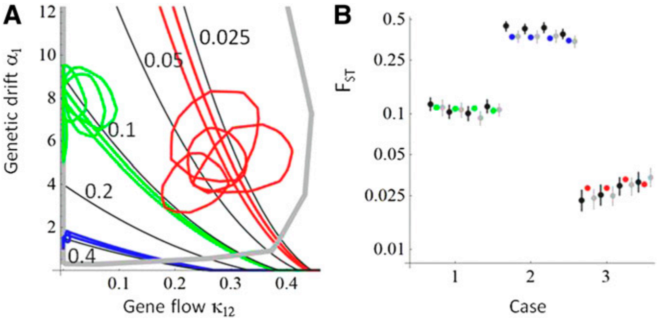


Figure 5 The AFM fitted to data generated by individual-based simulations of two identical subpopulations. The green color refers to baseline scenario 1 in which data were simulated assuming little gene flow and little random drift, blue refers to scenario 2 with a higher amount of drift, and red refers to scenario 3 with a higher amount of gene flow. (A) The 75% credible sets of the estimated parameters are plotted in (k_{12}, a_1) space. The colored lines show the isoclines of the minimal and maximal true F_{ST} values among the four replicate data sets generated for each scenario. (B) The F_{ST} values estimated by our method (black circles and error bars show the posterior median and 95% central credibility interval) are compared to the true values (colored circles) and to the Weir–Cockerham estimates (gray circles and error bars show the ML estimate and its 95% confidence interval) given by FSTAT (Goudet 1995). For parameter values used to generate the data, see *Case studies with an individual-based model* in the main text.

reflect a higher effective population size in the White Sea that is in direct contact with the Arctic Ocean. In line with the expectation of no recent gene flow due to geographic barriers, the level of between-population relatedness was very low in our analysis (median estimates of all off-diagonal terms of the matrix $\boldsymbol{\theta}^P$ were in the range $10^{-5}-10^{-3}$, attributable to numerical noise from the MCMC).

The second data set originates from a much smaller spatial setting, containing samples of the common shrew (*Sorex araneus*) on islands on the lake Sysmä (62°40′N, 31° 20′E) and the surrounding mainland in Finland (Hanski and Kuitunen 1986). Here we utilize data from the mainland, two large islands (L1 and L3, areas 3.8 and 4.4 ha), and two small islands (S5 and S10, areas 0.7 and 0.4 ha). The islands form two pairs, each consisting of a large and a small island, so that the distance between L1 and S5, as well as the distance between L3 and S10, is <500 meters, but the distance between any other pair of islands is at least 1300 meters. The diameter of the lake is ~3 km, and thus the size of the study system is comparable to the potential migration distances of shrews (Hanski and Kuitunen 1986).

The small spatial scale is reflected by the low overall degree of population differentiation, the AFM yielding the estimate $F_{ST} = 0.08$ (0.06–0.09) and the Weir–Cockerham estimator giving $F_{ST} = 0.05$ (0.04–0.07). As expected from variation in population size, the within-subpopulation relatedness (θ^P_{AA}) is lower for the mainland (median estimate 0.01) than for the islands (0.12, 0.10, 0.09, and 0.08 for L1, L3, S5, and S10, respectively). These findings are in line with the population-specific F_{ST} estimates (0.01, 0.12, 0.09, 0.09, and 0.06 in the same order). The only off-diagonal terms that are ≥0.01 in the median estimate are between the mainland and the island L1 (0.01) and between the islands L3 and S10 (0.01) that are located close to each other, but it is hard to draw conclusions on a more general pattern based on this observation. This is in line with the discriminant function analysis based on metrical traits by Hanski and Kuitunen (1986), which also revealed little indication of isolation by distance.

Discussion

The AFM can be used to infer population-level coancestry coefficients θ^P_{AB} from genotypic data. Mathematically, the AFM is a generalization of the model of Fu *et al.* (2005) for multiallelic data and a more general population structure. As discussed above, the estimates of θ^P_{AB} also relate to coalescent theory and thus to the definition of F_{ST} by Rousset (2004). Using the AFM for estimating F_{ST} is justified subject to two conditions: First, we have assumed that the subpopulations have diverged from a common ancestral population at some time in the past. Second, we have assumed that the mutation rate is low compared to the time elapsed since divergence or at least compared to the influence of potential gene flow after time since divergence. If these two conditions are met, θ^P_{AB} is close to its coalescent-based analogy (θ^{Pc}_{AB}), and thus it can be used for calculating the coalescent-based F_{ST} (Slatkin 1991, 1995; Rousset 2004). The AFM models the allele frequencies by an admixture of evolutionary independent lineages, but this assumption is less restrictive. As the simulations show, it can also be used to mimic the effects of continuous gene flow (Figure 5).

The parameters of the AFM convey information about the demographic history of the local populations, as we have demonstrated with the simulated data and the two natural data sets. Using the AFM, it is possible to analyze the level of connectivity between the subpopulations (as characterized by $\mathbf{k}$) and the relative effective population sizes of the underlying evolutionary lineages. However, it is not possible to disentangle the absolute effective population sizes and the number of generations after divergence (as they are not identifiable on basis of $\mathbf{a}$ alone), nor it is possible to deduce per-capita rates of migration.

Apart from demography, the AFM also makes a number of assumptions regarding the type of genetic data. As discussed above, the mutation rate is assumed to be low, suggesting that using microsatellite markers should be avoided. As usual in population-genetic studies, we have also assumed that the markers used are selectively neutral. Thus, markers subject to diversifying (stabilizing) selection are likely to cause an upward (downward) bias in the estimate of F_{ST}, as is the case of F_{ST} estimates obtained by other methods (Excoffier *et al.* 2009). Third, we have ignored genotyping error, which is known to increase the sampling variation of F_{ST} estimates (Bonin *et al.* 2004; Herrmann *et al.* 2010). The implementation of these features to the present framework would be an important extension that we hope to be addressed by future work. Finally, we have used the Dirichlet distribution to model random genetic drift within each of the independent lineages. This approximation should be taken with some criticism (Nicholson *et al.* 2002; Balding 2003). Some authors have used truncated normal distribution in place of Dirichlet for estimating F_{ST} (Nicholson *et al.* 2002; Weir and Hill 2002; Coop *et al.* 2010). However, both of these statistical models are approximations of the true model, and both of them have their limitations, which we discuss in File S2.

For the molecular ecologists and population geneticists, F_{ST} is probably a more familiar variable than the matrix $\boldsymbol{\theta}^P$. While most authors consider F_{ST} as a parameter, some consider it as an estimator or a point estimate of this parameter. For different types of data and different mutation models, a full "alphabet soup of related indices have been developed" (Whitlock 2011, p. 1083), which may cause part of the confusion. There has also been recent discussion concerning the aptitude of F_{ST} for measuring genetic differentiation (Jost 2008; Whitlock 2011). Some authors have reported that locus-specific values correlate with the polymorphism of the marker loci (Hedrick 2005; Carreras-Carbonell *et al.* 2006; Jost 2008). By the canonical definition (Equation 4), F_{ST} is fully determined by the coalescent, so that it is logically independent of ancestral polymorphism. On the other hand, a high rate of mutation of course shows both in F_{ST} and in the present level of polymorphism. At the limit of a low mutation rate, F_{ST} reduces into a function of expected coalescence times (Slatkin 1991, 1995; Rousset 2002, 2004; Whitlock 2011) that are independent of polymorphism. In line with this, our coancestry-based F_{ST} is a function of coancestry coefficients and the pedigree that do not depend on the ancestral polymorphism.

Jost (2008) pointed out that F_{ST} can have low values even if the subpopulations do not share any alleles. In terms of coancestry coefficients, this implies $\theta^P_{AB} = 0$ for two different populations. As illustrated by the *y*-axis of Figure 4, the value of F_{ST} can range anywhere between zero and one also in this case. However, unlike Jost (2008), we do not consider this as a problematic feature of F_{ST}. From the viewpoint of Equation 4, F_{ST} is just a summary statistic of the subpopulation-to-subpopulation coancestry matrix $\boldsymbol{\theta}^P$. A more detailed understanding of population structure can clearly be obtained by considering the entire matrix $\boldsymbol{\theta}^P$, rather than a single scalar. Like Whitlock (2011), we still consider F_{ST} to be a very useful quantity in population genetics, *e.g.*, for the reason that it is the relevant statistic for $F_{ST}-Q_{ST}$ comparisons that attempt to find signals of stabilizing and disruptive selection in quantitative traits (Merilä and Crnokrak 2001; Mckay and Latta 2002), although we note that also this analysis can be done more effectively using the full matrix of population-level coancestries θ^P_{AB} (Ovaskainen *et al.* 2011).

Acknowledgments

We thank Christopher Wheat, Juha Merilä, and Michael Whitlock for helpful comments; Takahito Shikano for providing the ninespine stickleback data; and Ilkka Hanski for providing the common shrew data. Our research was supported by the Academy of Finland (grants 124242, 129636, and 250444 to O.O. and the work by M.K. partly covered by grants 213491, 129662, 118673, and 134728 to Juha Merilä) and the European Research Council (starting grant 205905 to O.O.).

Literature Cited

Bahlo, M., and R. C. Griffiths, 2001 Coalescence time for two genes from a subdivided population. J. Math. Biol. 43: 397–410.

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. 63: 221–230.

Bhattacharya, S., A. E. Gelfand, and K. E. Holsinger, 2007 Model fitting and inference under latent equilibrium processes. Stat. Comput. 17: 193–208.

Bink, M. C. A. M., A. D. Anderson, W. E. Van De Weg, and E. A. Thompson, 2008 Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. Theor. Appl. Genet. 117: 843–855.

Bonin, A., E. Bellemain, P. B. Eidesen, F. Pompanon, C. Brochmann *et al.*, 2004 How to track and assess genotyping errors in population genetics studies. Mol. Ecol. 13: 3261–3273.

Carreras-Carbonell, J., E. Macpherson, and M. Pascual, 2006 Population structure within and between subspecies of the Mediterranean triplefin fish Tripterygion delaisi revealed by highly polymorphic microsatellite loci. Mol. Ecol. 15: 3527–3539.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411–1423.

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

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

Fernandez, J., and M. A. Toro, 2006 A new method to estimate relatedness from molecular markers. Mol. Ecol. 15: 1657–1667.

Fu, R., D. K. Dey, and K. E. Holsinger, 2005 Bayesian models for the analysis of genetic structure when populations are correlated. Bioinformatics 21: 1516–1529.

Gaggiotti, O. E., and M. Foll, 2010 Quantifying population structure using the F-model. Mol. Ecol. Res. 10: 821–830.

Gelman, A., and J. B. Carlin, 2004 *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Gorsuch, R. L., 1983 *Factor Analysis*. Lawrence Erlbaum, Hillsdale, NJ.

Goudet, J., 1995 FSTAT (Version 1.2): a computer program to calculate F-statistics. J. Hered. 86: 485–486.

Hanski, I., and J. Kuitunen, 1986 Shrews on small islands: epigenetic variation elucidates population stability. Holarct. Ecol. 9: 193–204.

Hedrick, P. W., 2005 A standardized genetic differentiation measure. Evolution 59: 1633–1638.

Herrmann, D., B. N. Poncet, S. Manel, D. Rioux, L. Gielly *et al.*, 2010 Selection criteria for scoring amplified fragment length polymorphisms (AFLPs) positively affect the reliability of population genetic parameter estimates. Genome 53: 302–310.

Jost, L., 2008 G(ST) and its relatives do not measure differentiation. Mol. Ecol. 17: 4015–4026.

Karhunen, M., 2012 RAFM: Admixture F-model. Available at: http://CRAN.R-project.org/package=RAFM.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, New York.

Maenhout, S., B. De Baets, and G. Haesaert, 2009 CoCoa: a software tool for estimating the coefficient of coancestry from multilocus genotype data. Bioinformatics 25: 2753–2754.

Mckay, J. K., and R. G. Latta, 2002 Adaptive population divergence: markers, QTL and traits. Trends Ecol. Evol. 17: 285–291.

Merilä, J., and P. Crnokrak, 2001 Comparison of genetic differentiation at marker loci and quantitative traits. J. Evol. Biol. 14: 892–903.

Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Stat. Soc. Ser. B Stat. Methodol. 64: 695–715.

Ovaskainen, O., H. Rekola, E. Meyke, and E. Arjas, 2008 Bayesian methods for analyzing movements in heterogeneous landscapes from mark-recapture data. Ecology 89: 542–554.

Ovaskainen, O., M. Karhunen, C. Zheng, J. M. C. Arias, and J. Merilä, 2011 A new method to uncover signatures of divergent and stabilizing selection in quantitative traits. Genetics 189: 621–632.

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.

Shikano, T., Y. Shimada, G. Herczeg, and J. Merila, 2010 History *vs.* habitat type: explaining the genetic structure of European nine-spined stickleback (Pungitius pungitius) populations. Mol. Ecol. 19: 1147–1161.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. 58: 167–175.

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

Wang, Y., and J. Hey, 2010 Estimating divergence parameters with small samples from a large number of loci. Genetics 184: 363–379.

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

Weir, B. S., and W. G. Hill, 2002 Estimating F-statistics. Annu. Rev. Genet. 36: 721–750.

Whitlock, M. C., 2011 G′(ST) and D not replace F(ST). Mol. Ecol. 20: 1083–1091.

Wilkinson-Herbots, H. M., 2003 Coalescence times and F-ST values in subdivided populations with symmetric structure. Adv. Appl. Probab. 35: 665–690.

Wilkinson-Herbots, H. M., and R. Ettridge, 2004 The effect of unequal migration rates on F(ST). Theor. Popul. Biol. 66: 185–197.

Communicating editor: M. A. Beaumont

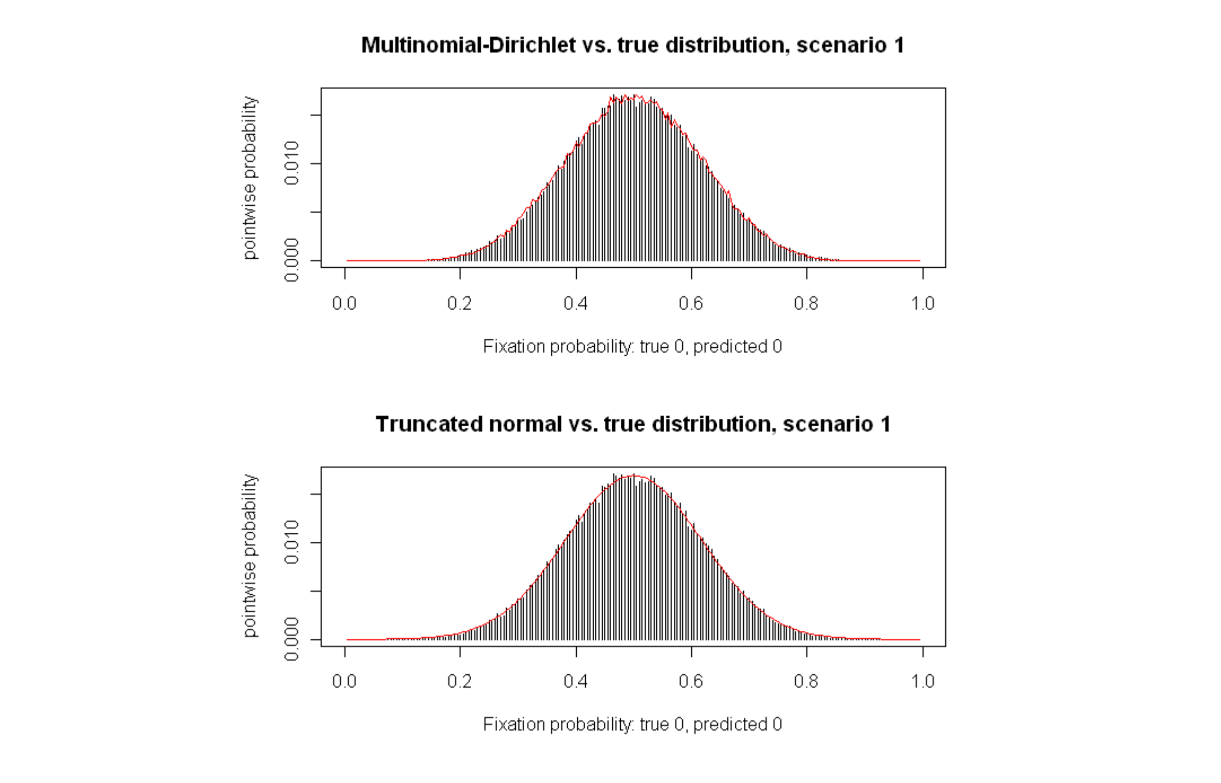
GENETICS

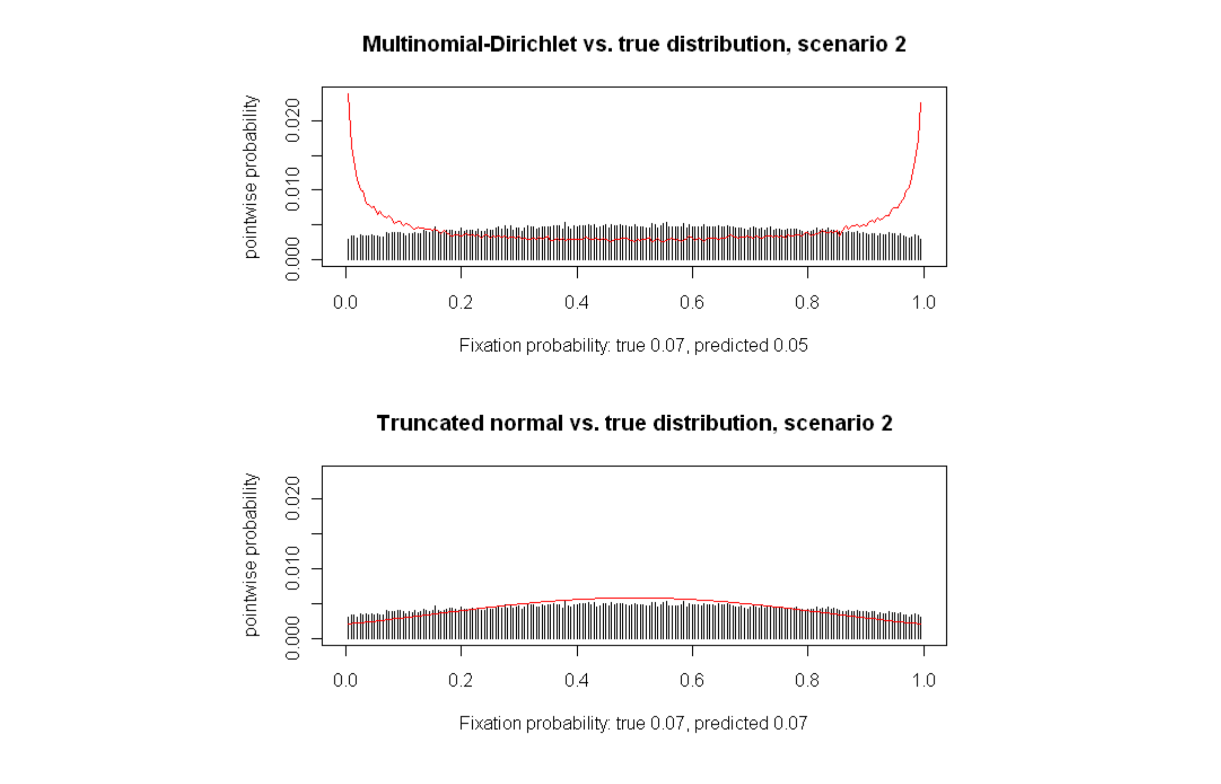
Supporting Information
http://www.genetics.org/content/early/2012/07/10/genetics.112.140871/suppl/DC1

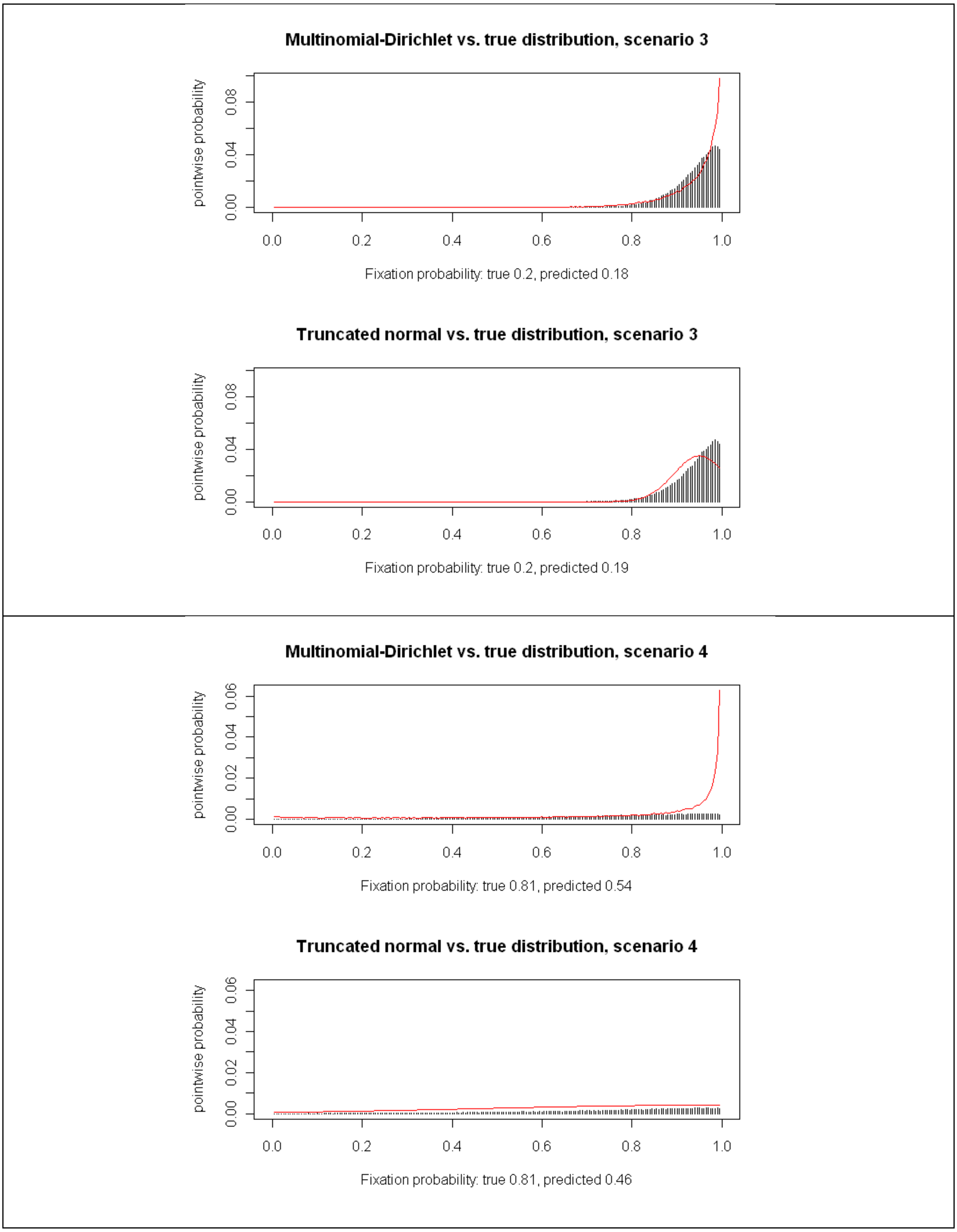
Estimating Population-Level Coancestry Coefficients by an Admixture F Model

Markku Karhunen and Otso Ovaskainen

DOI: 10.1534/genetics.112.140871


Multinomial-Dirichlet vs. true distribution, scenario 1
pointwise probability
0.000
0.010
0.0
0.2
0.4
0.6
0.8
1.0
Fixation probability: true 0, predicted 0
Truncated normal vs. true distribution, scenario 1
pointwise probability
0.000
0.010
0.0
0.2
0.4
0.6
0.8
1.0
Fixation probability: true 0, predicted 0


Multinomial-Dirichlet vs. true distribution, scenario 2
pointwise probability
0.000
0.010
0.020
0.0
0.2
0.4
0.6
0.8
1.0
Fixation probability: true 0.07, predicted 0.05
Truncated normal vs. true distribution, scenario 2
pointwise probability
0.000
0.010
0.020
0.0
0.2
0.4
0.6
0.8
1.0
Fixation probability: true 0.07, predicted 0.07



Multinomial-Dirichlet vs. true distribution, scenario 3

Truncated normal vs. true distribution, scenario 3

Multinomial-Dirichlet vs. true distribution, scenario 4

Truncated normal vs. true distribution, scenario 4

Figure S1 This figure represents an empirical sample from the true distribution of allele frequency (black discrete distribution) in four scenarios, and two approximations to it: Multinomial-Dirichlet and truncated normal. The parameter values are: Scenario 1: $T = 10$ generations, $n_A = 100$ individuals, initial frequency $q_{j1} = 0.5$; Scenario 2: $T = 100$, $n_A = 100$, $q_{j1} = 0.5$; Scenario 3: $T = 10$, $n_A = 100$, $q_{j1} = 0.95$; Scenario 4: $T = 100$, $n_A = 100$, $q_{j1} = 0.95$.

File S1

Calculating coancestry coefficients

Calculating coancestry coefficients from the admixture F-model (AFM). The frequency of allele u in subpopulation A is simply the average of the indicator variables x_{ijku},

$$p_{Aju} = \frac{1}{2n_A}\sum_{i=1}^{n_A}(x_{ij1u} + x_{ij2u}). \quad \text{(Eq. S1)}$$

The frequency p_{Aju} is a random variable, with expectation (over the flow of neutral alleles through a pedigree structure that we consider fixed) q_{ju}. The covariance among subpopulations A and B is

$$\text{Cov}(p_{Aju}, p_{Bju}) = \mathrm{E}\big[(p_{Aju} - p_{ju})(p_{Bju} - p_{ju})\big] = \mathrm{E}(p_{Aju}p_{Bju}) - q_{ju}^2 = (q_{ju} - q_{ju}^2)\theta_{AB}^{\mathcal{P}}, \quad \text{(Eq. S2)}$$

where the last equality follows (at the limit of low mutation rate) by noting that identity by state follows either from identity by descent, or by the two distinct alleles in the ancestral population being identical by state,

$$\mathrm{E}(x_{i'ju}x_{iju}) = q_{ju}^2 + \theta_{ii'}(q_{ju} - q_{ju}^2) \quad \text{(Eq. S3)}$$

and substituting the definitions of p_{Aju} (Eq. S1) and $\theta_{AB}^{\mathcal{P}}$ (Eq. 1 in the main text) into Equation S1.

In the AFM, it holds that

$$\text{Cov}(p_{Aju}, p_{Bju}) = \text{Cov}\left(\sum_{k=1}^{n_L}\kappa_{Ak}z_{kju}\,,\ \sum_{k=1}^{n_L}\kappa_{Bk}z_{kju}\right). \quad \text{(Eq. S4)}$$

Because the lineages are independent, Cov($z_{kju}, z_{k'ju}$)=0 for all $k \neq k'$. Thus, Equation S4 reduces to

$$\text{Cov}(p_{Aju}, p_{Bju}) = \sum_{k=1}^{n_L}\kappa_{Ak}\kappa_{Bk}\text{Var}\ z_{kju} = \sum_{k=1}^{n_L}\kappa_{Ak}\kappa_{Bk}\frac{q_{ju} - q_{ju}^2}{a_k + 1}. \quad \text{(Eq. S5)}$$

Combining this with Eq. S2 yields

$$(q_{ju} - q_{ju}^2)\theta_{AB}^{\mathcal{P}} = \sum_{k=1}^{n_L}\kappa_{Ak}\kappa_{Bk}\frac{q_{ju} - q_{ju}^2}{a_k + 1}, \quad \text{(Eq. S6)}$$

and hence

$$\theta_{AB}^{\mathcal{P}} = \sum_{k=1}^{n_L}\frac{\kappa_{Ak}\kappa_{Bk}}{a_k + 1}. \quad \text{(Eq. 12, main text)}$$

Calculating coancestry coefficients from a pedigree. If the complete pedigree is known, it is easy to calculate coancestry coefficients for each pair of individuals using the recursive formula (Lynch and Walsh 1998),

$$\theta_{ii'} = \frac{\theta_{is(i')} + \theta_{id(i')}}{2} = \frac{\theta_{s(i)s(i')} + \theta_{s(i)d(i')} + \theta_{d(i)s(i')} + \theta_{d(i)d(i')}}{4} \text{ for } i \neq i',$$

$$\theta_{ii} = \frac{1 + \theta_{s(i)d(i)}}{2}. \quad (\text{Eq. S7})$$

Above, $s(i)$ and $d(i)$ are the sire and dam of individual i, respectively. We used this formula for calculating the true value of $\boldsymbol{\theta}^{\mathcal{P}}$ in our simulated data sets.

References

Lynch, M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer Associates Incorporated, New York.

File S2

Models for allele frequencies

The vector of allele counts at locus $\boldsymbol{j}$ in an isolated population $\boldsymbol{A}$ for generation $\boldsymbol{t+1}$ follows the multinomial distribution

$$n_{Aj}(t+1)|n_{Aj}(t)\sim\mathrm{Mult}\big(2n_A,p_{Aj}(t)\,\big)\quad(\mathrm{Eq.\,S8})$$

where $\boldsymbol{p}_{Aj}(t)=\frac{\boldsymbol{n}_{Aj}(t)}{2n_A}$ is the allele frequency for the generation t. While $\boldsymbol{n}_{Aj}$ follows a multinomial random walk, $\boldsymbol{p}_{Aj}$ follows a corresponding process on an n_j-1 dimensional simplex. This discrete process is often approximated by a continuous-valued random process, the so-called Wright-Fisher diffusion (see e.g. Nicholson *et al.* 2002). Kimura (1955) first derived the exact solution for the distribution of allele frequencies of a biallelic locus under Wright-Fisher diffusion. This solution is not Gaussian, because the diffusion is non-isotropic. Solutions have also been obtained for multiallelic loci (Tavaré 1984; Xie 2011). However, implementing these solutions in the AFM framework would pose considerable computational challenges because of the need to iterate infinite, high-dimensional sums. In case of biallelic loci, the solution of Wright-Fisher diffusion is often approximated by a truncated normal distribution (e.g. Balding 2003; Coop *et al.* 2010; Nicholson *et al.* 2002). However, this approximation cannot be applied on multiallelic loci as such, because the distribution of $\boldsymbol{p}_{Aj}$ needs to be restricted on the simplex Δ^{n_j-1}. The alternative that we apply here is to use the Dirichlet distribution as a phenomenological, i.e. non-mechanistic, model for allele frequencies.

Application of the Dirichlet distribution as a model of pure drift may be considered questionable for two reasons. Firstly, the Dirichlet distribution is known to arise as an equilibrium distribution from the balance of random drift and mutation or migration (e.g. Nicholson *et al.* 2002; Rannala 1996), but not as a result of pure random drift in an isolated population. Secondly, the Dirichlet distribution is a continuous distribution such that each component is restricted on the open interval $]0,1[$, which gives a zero probability for the fixation of any one allele. However, with a small value of the parameter a_A, the Dirichlet distribution can have much of its probability mass very close to the boundaries. Thus, when supplemented with a sampling model for a finite population,

$$\boldsymbol{n}_{Aj}\sim\mathrm{Mult}\big(2\boldsymbol{n}_A,\boldsymbol{p}'_{Aj}\big),$$

$$\boldsymbol{p}'_{Aj}\sim\mathrm{Dirichlet}\big(a_A\boldsymbol{q}_j\big),$$

the Dirichlet model is able to predict a high probability of fixation. In the AFM, we use Dirichlet-distributed allele frequencies z_{Aj} to model the evolutionary history of the independent lineages, and the multinomial step naturally follows from the fact that the sample of genotypes is finite, even if the whole subpopulation is sampled. Below, we investigate this model by a comparison with the truncated normal distribution in a biallelic case where both distributions are easily tractable.

We consider a closed population of N individuals that mates randomly for T generations, and assume that the initial frequency of allele 1 has been q_{j1}. In this Supplement, we focus on four representative cases: symmetric allele frequencies with moderate drift (Scenario 1, Fig. S1), symmetric allele frequencies with a high amount of drift (Scenario 2, Fig. S1), uneven allele frequencies with moderate drift (Scenario 3, Fig. S1) and uneven allele frequencies with a high amount of drift (Scenario 4, Fig. S1). To sample from this model, we first generated a sample of size 10^5 from the last generation by using the true model (repeated application of Eq. S8). Then, we derived a corresponding sample from the Dirichlet approximation by randomizing $\boldsymbol{p}'_{Aj}$ for 10^5 times and sampling the allele counts for each realization from $\mathrm{Mult}(2n_A, \boldsymbol{p}'_{Aj})$. Finally, we considered the model of allele frequencies under the truncated normal approximation. As suggested by Nicholson et al. (2002), we specified the allele frequency as

$$p_{Aj1} \sim \mathrm{N}(q_{j1}, cq_{j1}(1-q_{j1})) \coloneqq \Phi$$

so that the extinction probability of allele 1 was calculated as $\Phi(0)$ and the fixation probability as $1-\Phi(1)$. We calculated the pointwise probabilities of the discrete classes as $\Phi'(p_{Aj1})/2n_A$, i.e. by dividing the Gaussian density function by the number of discrete values in]0,1]. While theoretical values exist for the drift parameters c and a_A given the demographic model, we optimized the values of these parameters in each scenario by minimizing the square distance (denoted D^2) with the true (empirical) distribution.

The results show that both the Dirichlet and truncated normal are imperfect approximations. In scenario 1, where drift is moderate and fixations do not occur, both approximations are qualitatively good, while the truncated normal distribution has a better goodness of fit (Multinomial-Dirichlet $D^2 = 9.2\times10^{-8}$; truncated normal $D^2 = 5.9\times10^{-8}$). In scenario 2, the truncated normal approximation has a better goodness of fit ($D^2 = 5.9\times10^{-7}$), than the Dirichlet approximation ($D^2 = 1.5\times10^{-5}$) which has an inconveniently convex shape in this case. In scenario 3, the truncated normal approximation has a better goodness of fit ($D^2 = 1.1\times10^{-5}$), but it is qualitatively different from the data by having a clear mode in the interior of $[0,1]$ which the Dirichlet approximation ($D^2 = 2.5\times10^{-5}$) does not have. In scenario 4, the Dirichlet approximation is better ($D^2 = 3.8\times10^{-4}$ as opposed to truncated normal $D^2 = 6.1\times10^{-4}$). In general, both distributions have problems in coping with the data when the amount of drift is high, which shows in the increase of the square distances. Finally, we note that the expectation of the truncated normal distribution is not strictly q_{j1} which would be expected under pure random drift. On the other hand, this is likely to be unimportant when the amount of drift is low.

References

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol **63:** 221-230.

Coop, G., D. Witonsky, A. Di Rienzo and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics **185:** 1411-1423.

Kimura, M., 1955 Solution of a Process of Random Genetic Drift with a Continuous Model. Proceedings of the National Academy of Sciences of the United States of America **41:** 144-150.

Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson *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.

Rannala, B., 1996 The Sampling Theory of Neutral Alleles in an Island Population of Fluctuating Size. Theor Popul Biol **50:** 91-104.

Tavare, S., 1984 Line-of-Descent and Genealogical Processes, and Their Applications in Population-Genetics Models. Theoretical Population Biology **26:** 119-164.

Xie, X. H., 2011 The Site-Frequency Spectrum of Linked Sites. Bulletin of Mathematical Biology **73:** 459-494.

File S3

The MCMC sampling scheme

We used a random-walk Metropolis-Hastings algorithm to sample the joint posterior density of $\boldsymbol{q}$, $\boldsymbol{a}$ and $\boldsymbol{\kappa}$. Below, we describe how each parameter was sampled while keeping the other parameters fixed.

- Sampling the drift parameters $\boldsymbol{a}$. We used N$(a_k, \delta^2_{a_k})$ distributions separately for each k to draw proposals for $\log a_k$. The variance parameters $\delta^2_{a_k}$ were adjusted during the burn-in as in Ovaskainen *et al*. (2008) to give an accept ratio of 0.44.
- Sampling lineage loadings $\boldsymbol{\kappa}$. We used TDD$(\delta_{\boldsymbol{\kappa}_A}\boldsymbol{\kappa}_A)$, i.e. truncated Dirichlet, distributions (Fang *et al*. 2000) separately for each A and j to draw proposals for $\boldsymbol{\kappa}_A$. The $\delta_{\boldsymbol{\kappa}_A}$'s are proposal parameters that were adjusted during the burn-in as in Ovaskainen *et al*. (2008) to give an accept ratio of 0.44.
- Sampling ancestral allele frequencies $\boldsymbol{q}$ and lineage-specific allele frequencies $\boldsymbol{z}$. We used TDD$(\delta_{q_j}\boldsymbol{q}_j)$ and TDD$(\delta_{q_j}\boldsymbol{z}_{kj})$ distributions separately for each j and k to draw proposals for the allele frequencies. The δ_{q_j}'s are proposal parameters that are adjusted during the burn-in as in Ovaskainen et al. (2008) to give an accept ratio of 0.44.

We thus used the truncated Dirichlet distribution of Fang *et al*. (2000) to perform the Metropolis-Hastings random walk for the Dirichlet-distributed variables $\boldsymbol{\kappa}$**,** $\boldsymbol{q}$ and $\boldsymbol{z}$ with a pre-set truncation threshold $\tau = 10^{-7}$. This greatly improves the mixing properties of the Markov chain, because it helps to avoid numerical problems on the boundary of the parameter space (i.e. on the edges of the simplices Δ^{n_j-1} and $\Delta^{n_\mathcal{P}-1}$). According to our observation, the method that Fang *et al*. (2000) present for sampling from TDD may produce biased samples for high truncation thresholds such as $\tau = 10^{-1}$. However, to our experience, this does not compromise the statistical power of our algorithm with $\tau = 10^{-7}$.

We have implemented the algorithm described above in the R-package RAFM (Karhunen 2012).

References

Fang, K.T., Z. Geng and G.L. Tian, 2000 Statistical inference for the truncated Dirichlet distribution and its application in misclassification. Biometrical Journal **42**: 1053-1068.

Karhunen, M., 2012 RAFM: Admixture F-model. http://CRAN.R-project.org/package=RAFM.

Ovaskainen, O., H. Rekola, E. Meyke and E. Arjas, 2008 Bayesian methods for analyzing movements in heterogeneous landscapes from mark-recapture data. Ecology **89:** 542-554.