ACADEMIC
PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Theoretical Population Biology 63 (2003) 221–230

Theoretical Population Biology

http://www.elsevier.com/locate/ytpbi

Likelihood-based inference for genetic correlation coefficients

David J. Balding*

Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK

Received 18 August 2002

Abstract

We review Wright's original definitions of the genetic correlation coefficients F_{ST}, F_{IT}, and F_{IS}, pointing out ambiguities and the difficulties that these have generated. We also briefly survey some subsequent approaches to defining and estimating the coefficients. We then propose a general framework in which the coefficients are defined, their properties established, and likelihood-based inference implemented. Likelihood methods of inference are proposed both for bi-allelic and multi-allelic loci, within a hierarchical model which allows sharing of information both across subpopulations and across loci, but without assuming constancy in either case. This framework can be used, for example, to detect environment-related diversifying selection.

Keywords: Fixation indices; Inbreeding; Selection; Bayesian statistics; Population genetics; Multinomial-Dirichlet; Forensic DNA profiles

1. Introduction

Samples of homologous sequences drawn from an interbreeding, diploid, population cannot be interpreted as independent observations. Shared inheritance generates correlations, and these can have a substantial impact on the inferences which may be drawn. Roughly speaking the intuition is that, because there are many DNA sequences possible at a locus of interest, any particular sequence is a priori expected to be very rare. But as soon as a gamete is sampled and its DNA sequence is observed, it becomes much more likely that there are further copies of the same or similar sequences in the population, carried by gametes sharing recent ancestry with the observed gamete. Even when a large sample from the population is available, so that population allele frequencies are well estimated by sample counts, if the population is structured then there can be correlations within subpopulations. Further, non-random mating can generate correlations between the two gametes forming a zygote.

Wright (1943, 1951) defined F_{ST} as:

> the correlation between random gametes, drawn from the same subpopulation, relative to the total.

F_{ST} was interpreted by Wright as a measure of the progress of the subpopulation towards the fixation of one allele at each locus (in the absence of mutation) and hence has come to be called a "fixation index". We show below that F_{ST} can be interpreted as a scaled variance of subpopulation allele frequencies. It can also be interpreted as a measure of shared ancestry within the subpopulation, relative to that in the population, and is thus sometimes called the "coancestry coefficient" or the "kinship coefficient".

Wright's definition of F_{IT} is similar to that of F_{ST} except that the two gametes form a zygote. This is also true for F_{IS}, but in this case the correlation is relative to the subpopulation. These are interpreted as measures of inbreeding; the term "inbreeding coefficient" often refers to F_{IS}, although it is sometimes also used for F_{IT}.

Estimates of F_{ST}, F_{IT}, and F_{IS} are widely used to measure population structure and deviations from random mating, and to estimate demographic parameters such as migration rates (but see Rousset, 2001). Estimates of F_{ST} have also been used to identify loci subject to geographically-varying selection (Lewontin and Krakauer, 1973; Beaumont and Nichols, 1996; Akey et al., 2002). With the rise of more sophisticated population genetics models and computational algorithms (see, for example, Stephens, 2001) it may be argued that Wright's coefficients will become redundant, being replaced by direct estimates of quantities of

*Fax: +44-20-7402-2150.
E-mail address: d.balding@imperial.ac.uk.

0040-5809/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0040-5809(03)00007-8

interest such as migration rates and selection coefficients. However, they provide convenient summaries of subpopulation variation (Weir and Hill, 2002) that are likely to remain useful to population geneticists for many years to come.

Unfortunately, Wright's definitions are ambiguous, and conflicting interpretations of them have led to disagreements over appropriate methods of estimation. In particular, the probabilities which underly each "correlation" are not specified. Wright's earliest writings interpreted the correlations as being generated by Mendelian inheritance in a fixed, but perhaps partially unknown, pedigree relative to allele proportions in the founding stock. Later, Wright interpreted the correlations as being generated by an assumed random selection of gametes in the subpopulation, relative to the current total population. One school of thought in the literature has followed this interpretation, but it leads to difficulties as we discuss below. An alternative interpretation is that the expectations are with respect to a model for the evolution of genotypes within subpopulations. This approach has substantial advantages, but raises the question: in which model?

There are further difficulties with Wright's definitions. For useful statistical inferences, it is necessary to share information over alleles at a locus, and/or over loci, and/or over subpopulations. It is easy to verify that at a bi-allelic locus F_{ST}, F_{IT}, and F_{IS} are unaffected by the choice of allele. This is not the case for multi-allelic loci. In practice, constancy of the coefficients over alleles is often assumed, but the justification for this is not clear. Similarly, the definitions are in terms of gametes, which suggests that the coefficients are constant over loci, yet this is not clear and different mutation processes at different loci means that it may not be realistic. In practice, many authors have assumed constancy over subpopulations, but this is rarely realistic because of the subpopulations' differing sizes and demographic histories.

It is inherent in the notion of subpopulation that individuals mate preferentially within it. It is possible to artificially define groupings in which this does not hold, but this property cannot be sustained under any mating system and such groupings are not subpopulations in any useful sense. If this is accepted, then gametes within subpopulations share more ancestry on average than gametes from distinct subpopulations and hence

$$F_{ST} \geqslant 0. \tag{1}$$

This property seems to be widely assumed, and is implied in the interpretation of F_{ST} as a variance of allele frequencies, but (1) does not follow from Wright's definition of F_{ST} as a correlation. Further, assuming an island model of population structure, Wright (1943) derived the relationship:

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}), \tag{2}$$

but the assumptions underpinning (2) do not seem to have been clarified.

In this paper, we propose a general framework in which Wright's coefficients can be rigorously defined, properties such as (1) and (2) established, and likelihood-based methods of inference developed. We start by briefly reviewing how some previous authors have interpreted Wright's definitions. There have been several previous reviews of genetic correlation coefficients. Chakraborty and Danker-Höpfe (1991) give a lengthy review focussing on estimation, with little discussion of definitions. Weir (1996, Chapter 5) gives a brief introductory treatment. Excoffier (2001) provides a recommended entry point for those wishing to go beyond the brief review offered here. The focus here is different, in view of our ultimate goal of likelihood-based statistical inference.

2. Previous approaches

2.1. Descriptive

Descriptive approaches define Wright's correlation coefficients in terms of the actual population allele frequencies, without reference to the genetic processes which led to them. The formulation of Nei (1973, 1977) has been widely followed, and this has been reviewed and extended by Nagylaki (1998). For Nei and Nagylaki, the probabilities underlying the definitions of F_{ST}, F_{IT}, and F_{IS} are generated by random sampling of gametes: a subpopulation is chosen with probabilities proportional to arbitrary weights, and then two gametes are chosen randomly from this subpopulation, either from the same or different zygotes. Allele-specific probabilities are then averaged to obtain

$$F_{ST} = 1 - H_S/H_T, \quad F_{IT} = 1 - H_I/H_T,$$
$$F_{IS} = 1 - H_I/H_S, \tag{3}$$

where H_I is the probability of a heterozygote, and H_S and H_T are the probabilities that two gametes chosen in, respectively, the subpopulation and the population, have different alleles. Eqs. (3) immediately satisfy (2), since

$$\frac{H_I}{H_T} = \frac{H_I}{H_S} \times \frac{H_S}{H_T}.$$

They define parameters which are functions of the actual allele frequencies and which can be computed via complete enumeration of all the subpopulations. The definitions lead naturally to estimates which are the corresponding functions of samples. Bias corrections may also be applied (Nei and Chesser, 1983), but these do not overcome the fundamental difficulties which we now briefly discuss.

Under many population genetics models of interest, F_{ST} is constant over alleles, in which case the averaging over alleles is appropriate, but there is no justification within the descriptive framework for assuming constancy over alleles at multi-allelic loci. Further, (3) involves averages over subpopulations, thus in effect excluding the possibility that F_{ST}, F_{IT}, and F_{IS} vary over subpopulations. This restriction is crucial in practice: although it is possible to calculate a distinct value of H_I and of H_S for each subpopulation, the assumed replication is then over alleles only, so that statistical inferences are typically very imprecise, particularly in the bi-allelic case in which estimates are essentially based on a single allele frequency.

Another weakness of the approach is the arbitrary choice of subpopulation weights. Two possible choices for the weights are: (i) equal and (ii) proportional to sample size, the latter possibly as a proxy for subpopulation size. If only some of the existing subpopulations are sampled, then H_T will depend on the choice of subpopulations. One approach to overcoming this weakness is to assume an "island-continent" model of population structure, in which one subpopulation, the "continent", plays the role of the total population in definitions (3). However, in this case, and for the subpopulation-specific values discussed above, (1) is not satisfied.

2.1.1. Descriptive F_{ST}: numerical illustration

Consider two subpopulations each consisting of 10 gametes, with allele frequencies shown in Table 1. The probability that two gametes chosen with replacement have different alleles is:

$$\text{Subpopulation 1:} \quad H_S = 0.1 \times 0.9 + 0.6 \times 0.4 + 0.3 \times 0.7 = 0.54,$$

$$\text{Subpopulation 2:} \quad H_S = 0.5 \times 0.5 + 0.2 \times 0.8 + 0.3 \times 0.7 = 0.62,$$

$$\text{Total population:} \quad H_T = 0.3 \times 0.7 + 0.4 \times 0.6 + 0.3 \times 0.7 = 0.66.$$

The subpopulation-specific values of F_{ST} are thus $12/66 \approx 18\%$ for Subpopulation 1 and $4/66 \approx 6\%$ for Subpopulation 2, and the average F_{ST} is $8/66 \approx 12\%$. If we had chosen to regard Subpopulation 2 as a "continent" (i.e. a reference population), then the value of F_{ST} for Subpopulation 1 would be $8/62 \approx 13\%$. If Subpopulation 1 were taken to be the continent, then $F_{ST} = -8/54 \approx -15\%$.

Rather than entire subpopulations, the allele frequencies shown in the tables could be from samples, which is the usual situation in practice. The same numerical values for F_{ST} can be calculated from the samples, but each would now be interpreted as an estimate $\hat{F}_{ST}$ of the unknown value of F_{ST}.

Table 1
Allele frequencies at a tri-allelic locus in a hypothetical haploid population consisting of two subpopulations each of size 10

	Allele 1	Allele 2	Allele 3
Subpopulation 1	1	6	3
Subpopulation 2	5	2	3
Total	6	8	6

Descriptive definitions are conceptually simple and lead to natural estimates: a property of the population is estimated by the corresponding property of a sample from the population. However, if researchers do not want merely to describe actual gene frequencies, but to draw conclusions about the biological processes which led to them, then model-based definitions are more appropriate.

2.2. Model based

2.2.1. Components of variance

Weir and Cockerham (1984) assume a model consisting of an ancestral population from which subpopulations have descended in isolation under the same evolutionary processes, and F_{ST}, F_{IT}, and F_{IS} are defined as averages over subpopulations. Briefly, the total variation in the genetic data is broken down into three components: (a) between subpopulations; (b) between individuals within subpopulations; and (c) between gametes within individuals. The coefficients F_{ST}, F_{IT}, and F_{IS} are defined as the expectations under the model of $a/(a+b+c)$, $(a+b)/(a+b+c)$ and $b/(b+c)$ and estimated by the corresponding sample values (see also Cockerham and Weir, 1986). As for the descriptive definition, there is no general justification in this framework for assuming constancy over alleles, but separate coefficients can be defined and estimated for each allele.

Weir and Cockerham's definitions satisfy (2), since

$$\frac{c}{a+b+c} = \frac{b+c}{a+b+c} \times \frac{c}{b+c},$$

but do not satisfy (1); see the discussion in Excoffier (2001). Weir and Cockerham (1984) report simulations indicating that their estimators have good sampling properties relative to descriptive estimators, and similarly Nicholson et al. (2002) find that a method-of-moments estimator analogous to the Weir and Cockerham F_{ST} estimator was substantially superior to a heterozygosity-based estimator analogous to (3).

Although, the development of Weir and Cockerham (1984) is based on a specific model which may often be unrealistic, Weir and Hill (2002) argue that the estimators are method-of-moments estimators that require second-moment assumptions only, and hence are not restricted to any specific model. However, they remain model based in our sense that the probabilities underlying the moments are generated by evolutionary processes, and not random sampling. Weir and Hill (2002) also show that for multi-allelic loci, the requirement of a common value of F_{ST} over subpopulations can be relaxed.

2.2.2. *Identity by descent*

Crow and Kimura (1970) define F_{ST} to be the probability that two genes drawn from the subpopulation are identical by descent (ibd) through shared inheritance from a common ancestor within the subpopulation, excluding known, recent ancestors, such as a grandparent or great-grandparent. Similarly F_{IS} is the probability of a homozygote due to ibd from a common ancestor within the subpopulation. Note that these definitions imply (1) but also that $F_{IS} > 0$, and so outbreeding cannot be accommodated.

Constancy of F_{ST} over alleles follows from the definition, provided that the probability that two genes are ibd is not affected by knowing the allelic type of one of them, and that two genes which are not ibd are, in effect, independent draws from an evolutionary process whose parameters are known. The first of these assumptions requires mate selection to be independent of genotype, while the second holds, for example, if mutation is negligible and the genotypes of migrants are mutually independent.

The Crow and Kimura definition is conceptually helpful, since it highlights modelling assumptions under which, for example, constancy over alleles holds. It is, however, not directly useful for estimation since it is not possible to distinguish genes which are ibd from genes which are identical for other reasons.

2.2.3. *Ratios of coalescence times*

Slatkin (1991) uses definition (3) but interprets the underlying probabilities in terms of a population genetics model in which mutations occur independently and at constant rate. In the limit as the mutation rate becomes small, he shows that

$$F_{ST} = \frac{\bar{t} - t_0}{\bar{t}}, \tag{4}$$

in which t_0 and $\bar{t}$ denote the expected time to the most recent common ancestor in, respectively, the subpopulation and the population. This ratio of coalescence times approach is useful in interpretation of F_{ST} values in different settings, but does not lead directly to estimation methods.

3. A general framework

Satisfactory definitions of F_{ST}, F_{IT}, and F_{IS} should be model-based, to allow inferences about parameters of interest, but should not be too closely tied to a specific model. They should lead to properties (1) and (2), and should provide a framework for likelihood-based statistical inferences. We now set out a framework in which these criteria can be realised.

Consider gametes sampled in a particular subpopulation and typed at a bi-allelic locus. Write $I^k_{ij} = 1$ if the jth gamete, $j = 1, 2$, of the ith zygote bears the kth allele, otherwise $I^k_{ij} = 0$. Let $\tilde{p}_k$ denote the subpopulation relative frequency of allele k under our adopted population genetics model, so that

$$E[I^k_{ij} | \tilde{p}_k] = \tilde{p}_k. \tag{5}$$

Note that $\tilde{p}_k$ is a property of the model, and is logically distinct from the realised relative frequency of k in the actual, finite subpopulation, although the two quantities will be close if the subpopulation is large. Define

$$F^k_{IS} = \frac{E[I^k_{i1} I^k_{i2} | \tilde{p}_k] - \tilde{p}^2_k}{\tilde{p}_k(1 - \tilde{p}_k)}. \tag{6}$$

We will treat $\tilde{p}_k$ as an unobserved "random effect", whose expectation under the model,

$$E[\tilde{p}_k] = p_k, \tag{7}$$

can here be regarded as a known constant. Again, p_k is not the realised average value of $\tilde{p}_k$ in the subpopulations of an actual population, but the two quantities will be close if the population contains many subpopulations. From (5) and (7) we have $E[I^k_{ij}] = p_k$, and we can define

$$F^k_{IT} = \frac{E[I^k_{i1} I^k_{i2}] - p^2_k}{p_k(1 - p_k)}. \tag{8}$$

Finally, for $i \neq i'$, define

$$F^k_{ST} = \frac{E[I^k_{ij} I^k_{i'j'}] - p^2_k}{p_k(1 - p_k)}. \tag{9}$$

Although it suffices here to treat p_k as a constant, it is also possible for it be random under the model, in which case the expectations in (7)–(9) should be regarded as conditional on the value of p_k.

Our key assumption is that within a subpopulation of size N, the genotype sequence (I^k_{i1}, I^k_{i2}), $i = 1, \ldots, N$, can be regarded as the first N terms of an infinite, *exchangeable*, sequence. Exchangeable means that if we reorder the sequence its joint probability distribution is unaffected. Thus, the assumption requires that ordering within the subpopulation is unimportant, which implies for example that family relationships cannot be accommodated in our framework. A generalisation to allow additional levels of population stratification is straightforward, but not pursued here. The assumption also implies that the model does not impose an upper bound on subpopulation size, which rules out some pathological cases of little interest.

Under this exchangeability assumption, and assuming also exchangeability within each pair of the sequence, we may invoke the de Finetti representation theorem (see for example Bernardo and Smith, 1994) to conclude that, conditional on the value of $\tilde{p}_k$, the $I^k_{ij(i)}$, for any sequence $j(i) \in \{1,2\}$, $i = 1,2,\ldots,N$, are mutually independent. Consequently, the genotypes (I^k_{i1}, I^k_{i2}), $i = 1,2,\ldots,N$, are also mutually independent, given $\tilde{p}_k$. Since (I^k_{i1}, I^k_{i2}) can take only four possible values, and (7) specifies two of these, the joint distribution of the genotype sequence, given $\tilde{p}_k$, is fully specified by assigning a value to F^k_{IS}. The conditional independence of I^k_{ij} and $I^k_{i'j'}$, given $\tilde{p}_k$, implies that

$$F^k_{ST} = \frac{E[E[I^k_{ij} I^k_{i'j'} | \tilde{p}_k]] - p_k^2}{p_k(1-p_k)} = \frac{E[\tilde{p}_k^2] - p_k^2}{p_k(1-p_k)} = \frac{\mathrm{Var}[\tilde{p}_k]}{p_k(1-p_k)} \geqslant 0,$$

and hence F^k_{ST} can be obtained from the first two moments of the distribution of $\tilde{p}_k$. Further, conditioning on $\tilde{p}_k$ we have

$$\begin{aligned} p_k(1-p_k)F^k_{IT} &= E[(I^k_{i1} - p_k)(I^k_{i2} - p_k)] \\ &= E[(I^k_{i1} - \tilde{p}_k + \tilde{p}_k - p_k) \\ &\quad \times (I^k_{i2} - \tilde{p}_k + \tilde{p}_k - p_k)] \\ &= E[(I^k_{i1} - \tilde{p}_k)(I^k_{i2} - \tilde{p}_k)] + (\tilde{p}_k - p_k)^2 \\ &= \tilde{p}_k(1-\tilde{p}_k)F^k_{IS} + (\tilde{p}_k - p_k)^2. \end{aligned}$$

Eliminating $\tilde{p}_k$ by taking expectations gives

$$p_k(1-p_k)F^k_{IT} = (p_k - E[\tilde{p}_k^2])F^k_{IS} + p_k(1-p_k)F^k_{ST}.$$

Finally, replacing $E[\tilde{p}_k^2]$ with $p_k^2 + p_k(1-p_k)F^k_{ST}$ and rearranging we obtain (2) from which the value of F^k_{IT} can be computed.

We have thus defined F^k_{ST}, F^k_{IT}, and F^k_{IS}, and verified properties (1) and (2) in any population genetics setting in which the exchangeability assumption holds. This would seem to be the case for all models of interest and hence the definitions are fully general for all practical purposes, but there is an explicit criterion to be verified to confirm this in any particular case. If the exchangeability assumption is accepted, the role of a population genetics model is to assign both a distribution to $\tilde{p}_k$, and a value to F^k_{IS}, for the alleles k at a locus. There is no requirement for F^k_{ST} to be constant over k, but we have an explicit criterion: the distribution chosen for the $\tilde{p}_k$ should imply that $\mathrm{Var}[\tilde{p}_k]/p_k(1-p_k)$ is constant over k. In the bi-allelic case we have $\tilde{p}_{\bar{A}} = 1 - \tilde{p}_k$ and hence constancy over alleles of all three coefficients is automatically satisfied. A general multi-allelic setting in which this constancy holds is introduced below.

Note that our definitions apply to a particular subpopulation and locus. We do not assume constancy over either loci or subpopulations in general, but a hierarchical model in which information can be shared over loci and subpopulations is introduced below.

4. Likelihood-based inference for a bi-allelic locus

We consider here only bi-allelic loci, and hence we drop the subscript and superscript k. We continue to condition on the population allele proportion p. The multi-allelic case, and uncertainty about p, are discussed below. Within our general framework, the joint probability distribution of a sample is specified by (i) a probability distribution for $\tilde{p}$, taking values in the interval [0,1], and (ii) a value for F_{IS}. We briefly discuss here two parametric families that have been proposed for $\tilde{p}$, the truncated Gaussian and the beta.

Nicholson et al. (2002) assume a normal (Gaussian) distribution for $\tilde{p}$ with mean p and variance $cp(1-p)$, except that densities in the intervals $(-\infty, 0]$ and $[1, \infty)$ are replaced with atoms at 0 and 1, respectively. Because of this truncation, c is not the same as F_{ST}, but they are the same in the limit as $c \downarrow 0$. The truncated normal distribution is justified in terms of subpopulations which evolve in isolation after splitting from an ancestral population in which the allele proportion is p. The atoms at 0 and 1 correspond to probabilities of fixation. Based on this model, Nicholson et al. (2002) develop likelihood-based inference for c, implemented via Markov chain Monte Carlo. They assume constancy of the c over loci, and also develop a model for the ascertainment of single nucleotide polymorphism (SNP) markers: loci with p near $1/2$ are usually preferred to loci with p close to 0 or 1. Their simulation results indicate an important gain in precision of estimates over those based on non-likelihood methods.

Earlier, Balding and Nichols (1995, 1997) had proposed (in the bi-allelic case) a beta distribution for $\tilde{p}$, with expectation p and variance $p(1-p)/(1+\theta)$, so that $F_{ST} = 1/(1+\theta)$. The beta distribution can be justified as the equilibrium distribution for several population genetics models of interest, discussed further in the multi-allelic setting below.

The rationale for the truncated Gaussian assumption is in terms of a transient model, whereas that for the beta is in terms of an equilibrium model. The two distributions are similar when $p \approx 0.5$ and F_{ST} is not too large. Otherwise the models are qualitatively different since, for example, there is an atom of probability mass at zero for the truncated normal, compared with, for $F_{ST} > p/(1+p)$, a spike of infinite density at the origin under the beta. These differences have been exploited to try to distinguish populations in equilibrium from those in a transient phase, see for example Beaumont (2001). Marchini and Cardon (2002) compared the fit of the truncated Gaussian and beta models to two human datasets, but did not adjust for the distinction between c and F_{ST}.

The beta assumption conveys a key advantage for inference: $\tilde{p}$ can be integrated out exactly to obtain the beta-binomial likelihood in terms of F_{ST} and p only. The beta-binomial has a convenient recursive formulation as follows. Assume here that $F_{IS}=0$ and hence that, given $\tilde{p}$, gametes can be regarded as mutually independent. Suppose that n gametes have been sampled in the subpopulation, of which m have allele 1. Then the probability $P(m+1,n+1)$ that the next gamete sampled is allele 1 is

$$P(m+1,n+1)=\left(\frac{mF_{ST}+(1-F_{ST})p}{1+(n-1)F_{ST}}\right)P(m,n). \qquad (10)$$

The non-recursive form of (10) is

$$P(m,n)=\binom{n}{m}\frac{\Gamma(\theta)}{\Gamma(n+\theta)}\frac{\Gamma(m+\theta p)}{\Gamma(\theta p)} \times\frac{\Gamma(n-m+\theta(1-p))}{\Gamma(\theta(1-p))}, \qquad (11)$$

in which $\Gamma(x+1)=x\Gamma(x)$. Replacing θ with $1/F_{ST}-1$ we obtain a likelihood formula for F_{ST}. For $F_{IS}>0$, see the appendix of Ayres and Overall (1999).

In addition to permitting likelihood-based inference for F_{ST}, formulae (10) and (11) have applications in the setting of forensic DNA profiling. For example,

$$P(4,4) =\frac{(3F_{ST}+(1-F_{ST})p)(2F_{ST}+(1-F_{ST})p)(F_{ST}+(1-F_{ST})p)p}{(1+F_{ST})(1+2F_{ST})}, \qquad (12)$$

which on division by $P(2,2)=p(F_{ST}+(1-F_{ST})p)$ gives a match probability in the homozygote case for two unrelated individuals from the subpopulation when $\tilde{p}$ is unknown. For further details, and the heterozygote case, see Balding and Nichols (1995). In forensic cases involving paternity or mixed profiles, the joint probabilities of samples of size five, six, or more may be needed, and these can be obtained from (10) or (11).

5. Multi-allelic loci

Suppose now that there are K alleles, and write $\mathbf{p}$ for the vector $(p_1,p_2,\ldots,p_K)$. Weir and Hill (2002) develop estimators based on the multivariate normal distribution in this setting. An alternative approach which constrains the subpopulation allele proportions to be non-negative and sum to unity is to apply the normal assumption to log-ratios of allele frequencies (see Aitchison, 1986).

The multi-allelic analogue of the beta-binomial is the multinomial-Dirichlet, whose distribution is also given by the recursive formula (10) if we replace m with m_k, the sample count of allele k. The non-recursive form is

$$P(\mathbf{m},n)=\frac{n!\Gamma(\theta)}{\Gamma(n+\theta)}\prod_{k=1}^{K}\frac{\Gamma(m_k+\theta p_k)}{m_k!\Gamma(\theta p_k)}, \qquad (13)$$

where $\mathbf{m}=(m_1,m_2,\ldots,m_K)$ denotes the sample count vector so that $n=\sum_{k=1}^{K}m_k$. The multinomial-Dirichlet follows from our general framework if we assume that the vector $\tilde{\mathbf{p}}$ of subpopulation allele frequencies has the Dirichlet distribution with parameter $\theta\mathbf{p}$. One consequence of this assumption is that $\mathrm{Var}(\tilde{p}_k)=p_k(1-p_k)/(1+\theta)$ for all k, and so $F_{ST}=1/(1+\theta)$ is constant over alleles.

We conducted a small simulation study to investigate the gains for estimation of F_{ST} using the maximum-likelihood estimator (MLE) under sampling formula (13), when the multinomial-Dirichlet model is correct. To avoid complications unnecessary here, we assume that $F_{IS}=0$ and that p is known, hence the estimators used in the simulation study differ from those appropriate in practical situations when p is unknown. For comparison, we consider the method-of-moments (MoM) estimator

$$\frac{1}{S}\sum_{s=1}^{S}\frac{\sum_{k=1}^{K}(\hat{p}_{sk}-p_k)^2}{\sum_{k=1}^{K}p_k(1-p_k)}, \qquad (14)$$

where S denotes the number of subpopulations, and $\hat{p}_{sk}$ is the sample proportion of allele k in subpopulation s.

Both estimators are approximately unbiased and so only standard deviations are shown in Table 2. It is unsurprising that the MLE outperforms the MoM estimator when the assumed likelihood is correct, and interest focusses instead on the magnitude of the improvement, which is substantial in many cases. For example, when $F_{ST}=4\%$ or 8%, the MoM estimator based on a sample of size $n=400$ is inferior to the MLE with $n=100$.

Table 2
Standard deviations ($\times 10^4$) of MLE and MoM estimators of F_{ST} when the multinomial-Dirichlet assumption is valid

	True F_{ST}	0.2%	0.5%	1%	2%	4%	8%
$n=100$	MoM	52	66	87	132	224	404
	MLE	36	53	74	114	184	317
$n=200$	MoM	31	43	67	110	198	377
	MLE	23	37	57	91	159	288
$n=400$	MoM	19	33	56	101	192	376
	MLE	16	28	46	83	150	276

There were 10,000 simulations of samples of size n from each of five subpopulations, typed at a locus with $K=4$ alleles. The estimators of F_{ST} are (i) method-of-moments (MoM) estimator (14); and (ii) the maximum-likelihood estimator (MLE) obtained by maximising (13). The population proportion vector $\mathbf{p}$ was sampled uniformly randomly, independently for each simulation, and was regarded as known for the estimation of F_{ST}.

The multinomial-Dirichlet formula (13) is not generally valid, however there is a simple and rather general assumption which guarantees its validity: the probability that the $(n+1)$th gene sampled is allele k depends on m_k and n, but not on the counts for alleles other than k. Together with some minor technical assumptions (see Zabell, 1982), this assumption, known as "Johnson's sufficientness postulate", implies (13) and hence that F_{ST} is constant across alleles.

Formula (13) can also be derived directly in a simple migration-drift model. We briefly outline an intuitive explanation, see Balding and Nichols (1995) for further details, and Rannala (1996) for a derivation in a more general, non-equilibrium, island model. Consider a population in which migrant gametes arrive at rate $1 - F_{ST}$ with allele proportions specified by $\mathbf{p}$. Tracing back the ancestry of the sample, the $(n+1)$th gamete coalesces with a gamete of allele k at rate $m_k F_{ST}$ and is a migrant of allele k at rate $(1-F_{ST})p_k$. The sum of these terms gives the numerator of (10), whereas the denominator is the total rate of coalescence and migration events over all k alleles. Under this model, the definition of F_{ST} is identical to the ibd definition of Crow and Kimura (1970).

The multinomial-Dirichlet sampling formula (10) and (13) leads to the Ewens sampling formula (Ewens, 1972) in the limit as $k \uparrow \infty$ and $\max\{p_k\} \downarrow 0$, provided that we combine all unobserved alleles into a single class. The Ewens formula applies under the infinitely many alleles mutation model, in which the mutation rate (scaled with the population size) is θ, and every mutation generates a new allele. A recursive form of the Ewens formula in the case $m_k > 0$ can be obtained from (10) by setting the p_k to zero and substituting $F_{ST} = 1/(1+\theta)$, giving

$$P(m_k+1, n+1) = \left(\frac{m_k}{\theta+n}\right) P(m_k, n).$$

By summing over k such that $m_k > 0$ and taking the complement from one, the probability that the next allele sampled is one currently unobserved is found to be $\theta/(\theta+n)$.

Formula (13) may be invalid for microsatellite loci, at which alleles are distinguished by their lengths and the mutation process favours small changes in allele length. In this case, there are likely to be correlations within subpopulations between alleles of similar lengths and this cannot be accommodated by the Dirichlet distribution. However, since permutation of the allele labels does not affect F_{ST}, it seems plausible that the MLE based on (13) will often retain good properties as an estimator of F_{ST} even when (13) is not valid. This suggestion is supported by the results of a small simulation study shown in Table 3, at least when $F_{ST} < 5\%$, in which case the MLE gives a useful reduction in standard deviation over the MoM estimator, while its bias remains small. For $F_{ST} > 5\%$ the bias of the MLE may be substantial.

Table 3
Means and standard deviations ($\times 10^4$) of MLE and MoM estimators of F_{ST} when the multinomial-Dirichlet assumption is false

	Simulation	1	2	3	4	5	6
Mean	MoM	23	62	134	306	536	787
	MLE	24	62	134	314	572	898
SD	MoM	23	34	55	103	163	226
	MLE	17	26	44	86	147	227

Simulations were as for Table 2 except that a structured coalescent model with stepwise mutations was used instead of the multinomial-Dirichlet distribution; see text for further details. The population proportion vector $\mathbf{p}$ was set equal to the mean proportions in 100 simulated populations, each with $n = 200$, and was regarded as known for the estimation of F_{ST}.

The simulations employ the coalescent model (Nordborg, 2001) and the stepwise mutation model, for which mutations change the allele length by one unit, and increases and decreases are equally likely. The demographic model consists of a random-mating ancestral population which is of constant size until it splits into isolated, random-mating subpopulations. The split time is varied in the simulations to generate different F_{ST}-values. The absence of migration in this model, and the strict one-step mutation are both unrealistic, and weakening either of these might be expected to make Johnson's sufficientness postulate more plausible, and hence favour the performance of the MLE.

6. Combining information over loci and subpopulations

6.1. *The effects of sample size and number of alleles*

Fig. 1 shows likelihood curves for F_{ST} obtained using the multinomial-Dirichlet sampling formula (13) on data simulated under this model at a single locus, in a single subpopulation with $F_{ST} = 5\%$. We can see that F_{ST} is difficult to estimate, in the sense that the likelihood curves are in most cases not sharply peaked. There is a noticeable benefit from increasing the number of alleles at a locus: estimation is poor when $K = 2$. The problem is not fully overcome by increasing the sample size—samples of size 400 do not lead to much more precise estimation than do samples of size 100.

The latter observation can be verified directly by inspecting (10): as m and n both become large, the first term depends only weakly on F_{ST}, so that additional observations convey little information about its value. Intuitively, it is the $\tilde{p}_k$ which are informative about F_{ST}, and increasing the sample size does not increase the number of $\tilde{p}_k$ values: it increases the precision with which they are measured, but this brings diminishing

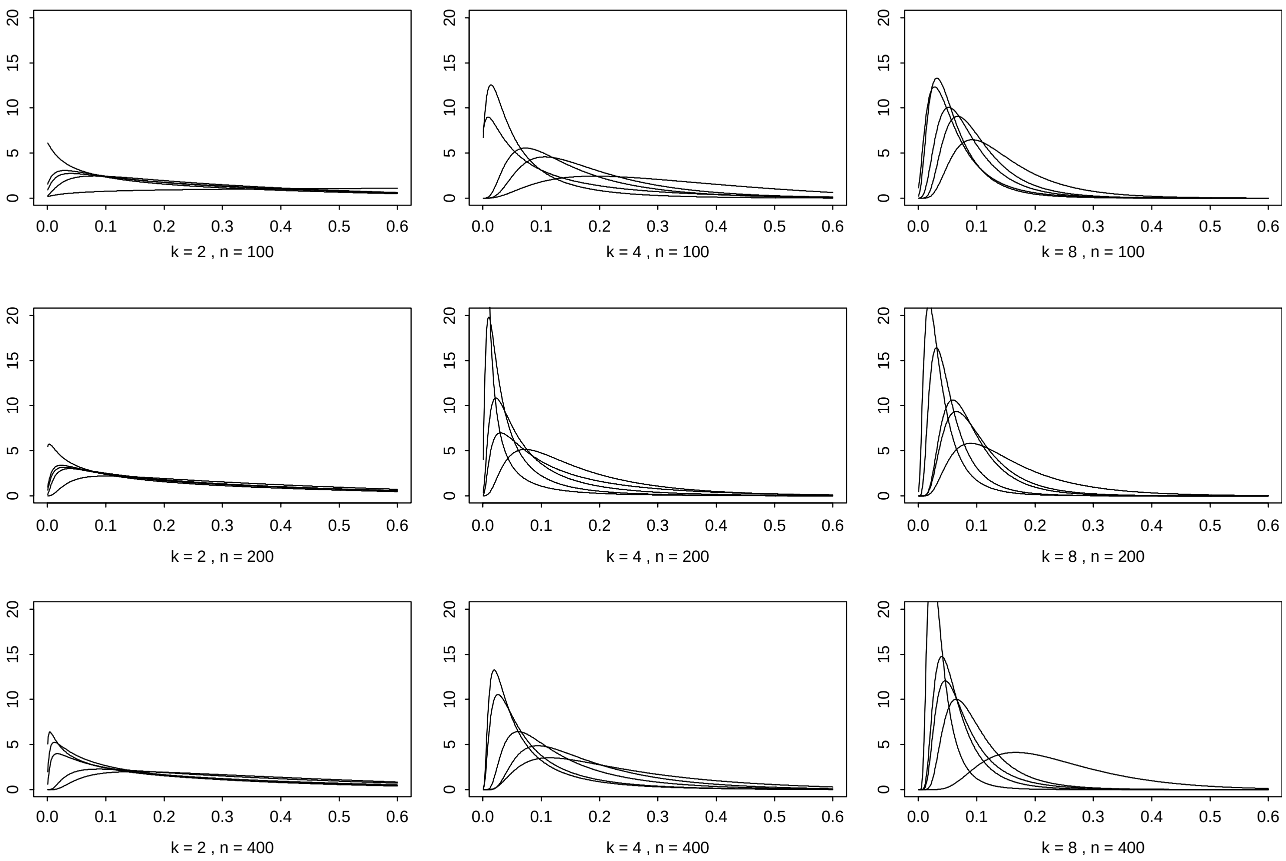

Fig. 1. Likelihood curves for F_{ST} based on the multinomial-Dirichlet sampling formula (13). Each panel shows the likelihoods for five datasets simulated under the model at one locus and in one subpopulation with $F_{ST} = 5\%$. The number of alleles (k) and sample size (n) are shown under each panel. The population allele proportions ($\mathbf{p}$) were simulated from the uniform distribution, independently for each curve, and were regarded as known in calculating the likelihoods. The y-axis is scaled so that the area under each curve is one, and hence the likelihood curves are also posterior density curves given a uniform prior.

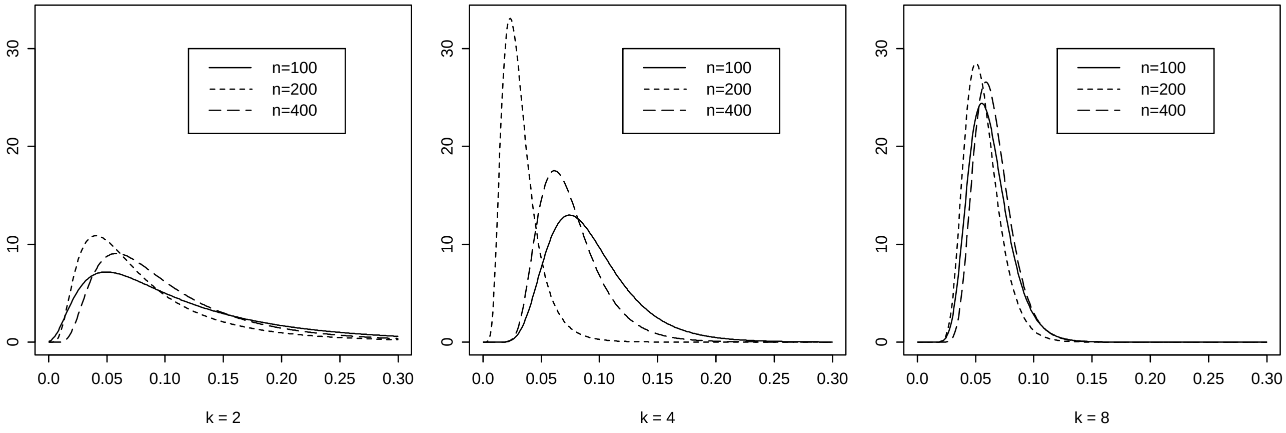

Fig. 2. The height of each curve is proportional to the product of the heights of the five curves from a panel of Fig. 1. Each curve is a likelihood curve for F_{ST}, and also a posterior density for a uniform prior, under the assumptions of a common value of F_{ST} for the five curves, and independence of the observations given the common value.

returns for inference about F_{ST}. The experimental effort involved in typing large samples is thus not always worthwhile for estimating F_{ST}.

Fig. 2 indicates the benefit of combining inferences over loci and/or over subpopulations, in the ideal case in which F_{ST} can be assumed constant over loci and/or subpopulations. Here, the five independent replications from each panel of Fig. 1 are combined. Again we see that there is little systematic benefit from increasing the sample size, but increasing the number of alleles at a locus does bring substantial benefit. Inferences remain poor when $K = 2$: many more than five replicates are needed in this case; but reasonably sharp inferences are possible when $K = 8$ for any of the sample sizes.

The assumption of constancy of F_{ST} underlying Fig. 2 is rarely valid in practice. Different subpopulations typically have different sizes and different patterns of migration and reproduction. Consequently, values of F_{ST} vary over subpopulations. Similarly, mutation rates vary over sites. This is most obviously true when different types of markers are being combined, but even for SNP markers variation in mutation rates seems likely, and hence F_{ST} may vary from one locus to another.

6.2. A hierarchical model

If the factors influencing the value of F_{ST} can be classified into those which are common to all loci within a subpopulation, such as migration and subpopulation size, and those which are common to all subpopulations at a locus, such as mutation and some forms of selection, then information about F_{ST} can be shared both across subpopulations and across loci, while still allowing a realistic degree of flexibility. This can be achieved via a hierarchical model in which the $S \times L$ individual F_{ST} values, for a system of S subpopulations and L loci, are expressed in terms of $S + L$ hyperparameters, one for each subpopulation and one for each locus. That is, we describe F_{ST}^{sl}, the value of F_{ST} at subpopulation s and locus l, in terms of hyperparameters α_s and β_l, which incorporate the effects of factors specific to, respectively, the subpopulation and the locus.

Balding et al. (1996) implemented the following model linking α_s and β_l with F_{ST}^{sl}:

$$F_{ST}^{sl} = \frac{1}{1 + \alpha_s + \beta_l}, \tag{15}$$

which was also derived by Wright (1943) in the "island" model of a subdivided population. If the migration rates between subpopulation s and the other subpopulations are all high, then α_s will be large and hence F_{ST}^{sl} will be small in this subpopulation for every locus l. Similarly, if the mutation rate at locus l, and the mutation process is stationary, then β_l will be large and hence F_{ST}^{sl} will be small at this locus for every subpopulation s. Conversely, if the β_l are small at a locus then F_{ST} is relatively large at that locus in all subpopulations, which may be interpreted as evidence of environment-specific diversifying selection.

Model (15) is therefore plausible on theoretical grounds. Nevertheless, it may not be appropriate for the actual data and this can be investigated by embedding it in the more general model:

$$F_{ST}^{sl} = \frac{1}{1 + \alpha_s + \beta_l + \gamma_{sl}}, \tag{16}$$

in which each γ_{sl} is a parameter measuring an effect specific to subpopulation s and locus l, which may be due to locus and subpopulation-specific selection. See Balding et al. (1996) and Balding and Nichols (1997) for further details and examples of application to microsatellite data from structured populations.

6.3. Implementation via Metropolis algorithm

The model based on (10) and (16) has been implemented via a Metropolis algorithm in a program FSTMET, for which ANSI C code and documentation are available at:

www.rdg.ac.uk/statistics/genetics

Because of the structure of the model, in which changes in α_s, β_l, and γ_{sl} have decreasing effects on F_{ST} as they increase in value, FSTMET works with the logarithms of these hyperparameters, and assumes Gaussian prior distributions whose means and variances may be input by the user. Independent mean-zero Gaussian distributions are used for the proposal distributions; standard deviations are input by the user to control acceptance rates.

FSTMET assumes that a large sample is available from a "continent", which allows direct estimation of the p_k. This situation often arises in the forensic setting when large forensic DNA databases are available, and interest focusses on match probabilities in a subpopulation of the population from which the database was sampled. Since the database estimates are used in place of the p_k in evaluating match probabilities, the genetic correlations with respect to these values are required in these settings.

6.4. Eliminating the population allele proportions

Throughout this paper we have regarded the p_k as known constants, which is implausible in practice. Unknown p_k can cause substantial difficulties for non-likelihood methods of estimating F_{ST}, because estimates of p_k are often affected by the number of subpopulations sampled, and the sample sizes, and a rational for adjusting for these effects is often lacking. The difficulties are greatly diminished within a likelihood based, Bayesian, framework: the p_k can be assigned a prior distribution, often a uniform prior is chosen, and inference about F_{ST} proceeds by integration over the p_k, which correctly accounts for the uncertainty arising from the finite number of populations sampled, and their finite sample sizes. Within the Bayesian framework estimates of F_{ST} within a population are affected by the number of populations sampled and their sample sizes, which is appropriate as these are informative about the p_k.

Integration over the p_k cannot usually be performed analytically, but can readily be performed within an MCMC algorithm (Holsinger, 1999). In some simple settings, the integration can safely be avoided by plugging in appropriate point estimators. This applies in the forensic database setting described above, and also when the number of subpopulations sampled is large, and all the sample sizes are large.

Acknowledgments

Karen Ayres developed the current C version of the FSTMET software described above, from an earlier Pascal program written by the author. I thank Peter Donnelly for helpful discussions about the ideas in the paper over almost a decade.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London, UK.

Akey, J.M., Zhang, G., Zhang, K., Jin, L., Shriver, M., 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12, 805–814.

Ayres, K.L., Overall, A.D.J., 1999. Allowing for within-subpopulation inbreeding in forensic match probabilities. Forensic Sci. Internat. 103, 207–216.

Balding, D.J., Greenhalgh, M., Nichols, R.A., 1996. Population genetics of STR loci in Caucasians. Int. J. Leg. Med. 108, 300–305.

Balding, D.J., Nichols, R.A., 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, D.J., Nichols, R.A., 1997. Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 78, 583–589.

Beaumont, M.A., 2001. Conservation genetics. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 779–812.

Beaumont, M.A., Nichols, R.A., 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. Roy. Soc. London B 263, 1619–1626.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, Chichester, UK.

Chakraborty, R., Danker-Höpfe, H., 1991. Analysis of population structure: a comparative study of different estimators of Wright's fixation indices. In: Rao, C.R., Chakraborty, R. (Eds.), Handbook of Statistics, Vol. 8. Elsevier, Amsterdam.

Cockerham, C.C., Weir, B.S., 1986. Estimation of inbreeding parameters in stratified populations. Ann. Hum. Genet. 50, 271–281.

Crow, J.F., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper and Row, New York.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Pop. Biol. 3, 87–112.

Excoffier, L., 2001. Analysis of population subdivision. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 271–307.

Holsinger, K.E., 1999. Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. Hereditas 130, 245–255.

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

Marchini, J.L., Cardon, L., 2002. Discussion on statistical modelling and analysis of genetic data. J. Roy. Statist. Soc. B 64, 740–742.

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

Nei, M., 1973. Analysis of gene diversity in subdivided populations. Proc. Natl Acad. Sci. USA 70, 3321–3323.

Nei, M., 1977. *F*-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41, 225–233.

Nei, M., Chesser, R.K., 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47, 253–259.

Nicholson, G., Donnelly, P., Smith, A., Jónsson, F., Gústaffson, Ó., Steánsson, K., 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. Roy. Statist. Soc. B 64, 645–716.

Nordborg, M., 2001. Coalescent theory. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 179–212.

Rannala, B., 1996. The sampling theory of neutral alleles in an island population of fluctuating size. Theor. Pop. Biol. 50, 91–104.

Rousset, F., 2001. Inferences from spatial population genetics. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 239–269.

Slatkin, M., 1991. In breeding coefficients and coalescence times. Genet. Res. Camb. 58, 167–175.

Stephens, M., 2001. Inference under the coalescent. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK.

Weir, B.S., 1996. Genetic Data Analysis II. Sinauer, Sunderland MA.

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

Weir, B.S., Hill, W.G., 2002. Estimating *F*-statistics. Annu. Rev. Genet. 36, 721–50.

Wright, S., 1943. Isolation by distance. Genetics 28, 114–138.

Wright, S., 1951. The genetical structure of populations. Ann. Eugen. 15, 323–354.

Zabell, S.L., 1982. W.E. Johnson's "sufficientness" postulate. Ann. Statist. 10, 1091–1099.