Annu. Rev. Genet. 2002. 36:721–50
doi: 10.1146/annurev.genet.36 050802.093940

First published online as a Review in Advance on September 12, 2002

ESTIMATING F-STATISTICS

B. S. Weir
Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-7566

W. G. Hill
Institute for Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Key Words population structure, forensic profiles, inbreeding, relatedness

■ **Abstract** A moment estimator of θ, the coancestry coefficient for alleles within a population, was described by Weir & Cockerham in 1984 (100) and is still widely cited. The estimate is used by population geneticists to characterize population structure, by ecologists to estimate migration rates, by animal breeders to describe genetic variation, and by forensic scientists to quantify the strength of matching DNA profiles. This review extends the work of Weir & Cockerham by allowing different levels of coancestry for different populations, and by allowing non-zero coancestries between pairs of populations. All estimates are relative to the average value of θ between pairs of populations. Moment estimates for within- and between-population θ values are likely to have large sampling variances, although these may be reduced by combining information over loci. Variances also decrease with the numbers of alleles at a locus, and with the numbers of populations sampled. This review also extends the work of Weir & Cockerham by employing maximum likelihood methods under the assumption that allele frequencies follow the normal distribution over populations. For the case of equal θ values within populations and zero θ values between populations, the maximum likelihood estimate is the same as that given by Robertson & Hill in 1984 (70). The review concludes by relating functions of θ values to times of population divergence under a pure drift model.

CONTENTS

INTRODUCTION

In 1984, Weir & Cockerham (100) published a set of equations for estimating the parameter F_{ST} or θ that describes the genetic structure of populations. The paper is still widely cited; in the first three months of 2002 the methods it described were applied to data on ash trees (59), Barbus (86), barley (42), barnacle (22), butterfly (18), cherry (54), cod (44), cord grass (85), Drosophila (32), eelgrass (64), frog (84), housefly (23), insects (58, 103), ladybird beetle (92), mackerel (11), moose (41), mountain lion (24), pig (45), pine (66, 68), quelea (19), red drum (33), redfish (72), river otter (9), rodent (14), salmon (37), scallops (67), sea trout (94), seaweed (88), shrimp (28), snail (13), stonefly (76), sugar beet (89), trout (38, 48), tsetse fly (47), wombat (7), zooplankton (34), and humans (1, 36, 53) among other species. Population biologists, ecologists and human geneticists have a substantial interest in being able to quantify the genetic relationships among their populations; it is therefore timely to re-visit the 1984 paper they cite. It may be especially useful to allow for different values of θ in different populations.

This discussion regards population structure, or the genetic differentiation of populations within the same species, as allelic frequency variation over populations. The restriction to allele frequencies, as opposed to genotypic frequencies, carries an implicit assumption of Hardy-Weinberg equilibrium at the loci under consideration. Even if two populations are maintained under the same evolutionary conditions they will have different allele frequencies because of the stochastic nature of these forces. Different evolutionary conditions for a set of populations will increase the differentiation among them, and θ can be defined in terms of variances and covariances of allele frequencies. The magnitude of these coefficients therefore reflects the evolutionary history of the populations being studied, although the observed allele frequencies also reflect the sampling processes within each population. The various approaches to estimating θ can differ according to whether they use only expected variances and covariances of allele frequencies or the entire frequency distributions. Use of the whole distribution may appear to be better, but there is an implicit constraint on the class of evolutionary scenarios if second-moment parameters are assumed to completely characterize a distribution.

The emphasis on within-species variation, and the usual use of unlinked loci means that coalescent approaches for non-recombining DNA sequences and deep evolutionary divergences [e.g., (61, 93)] are not considered.

LITERATURE REVIEW

Estimation Strategies

The search for the best estimators of θ, and the evaluation of existing estimators, continues. One way of distinguishing estimators is to consider how much of the distribution of allele frequencies across populations is used. It is shown below that the variances and covariances of allele frequencies across populations depend on θ as well as on the mean frequencies. This suggests that θ can be estimated from just the first and second moments of the allele frequency distribution, and this is the essence of the method of moments used by Weir & Cockerham (100). No particular evolutionary model leading to specific values for θ is assumed. Other methods assume the form of the whole distribution, which constrains applicability to certain evolutionary scenarios. The Dirichlet distribution used by Balding & Nichols (4) and Lange (49) assumes an evolutionary equilibrium, and is appropriate under the infinite alleles mutation model. Strictly, it is the Multinomial-Dirichlet distribution that is needed. The Dirichlet distribution is not appropriate for the stepwise mutation model (35). It is not clear that there is an evolutionary model for which the normal distribution used by Smouse & Williams (81), Long (51), and Nicholson et al. (60) and employed below in this review is appropriate, but it is justified by convenience and an appeal to large sample theory.

More statistical issues were addressed by Weicker et al. (95). The estimator of θ described by Weir & Cockerham (100) used the actual sample sizes in each sample in order to reduce bias, and Weicker et al. showed that good approximations to that estimator can be found that use the average sample size. These authors also presented confidence intervals found by bootstrapping over loci, with an implicit assumption that the number of loci is not small. Questions of both bias and variance were covered by Raufaste & Bonhomme (62) for loci with multiple alleles. The simplest models assume that allele frequency distributions have the same variances and covariances for all alleles, so that θ could be estimated separately for each allele. Raufaste & Bonhomme confirmed the prediction of Weir & Cockerham (100) that their weighting was satisfactory for larger values of θ, whereas an alternative weighting of Robertson & Hill (70) was better for small θ. The Robertson & Hill approach is equivalent to the multivariate approaches (51) described below.

This review is concerned with the relationships of pairs of alleles within and between populations, but a further hierarchy of relationships when there are subpopulations nested within populations, sub-subpopulations nested within subpopulations, and so on (97, 105). The nested analysis of variance structure is a natural framework for the analysis of that situation, and a generic definition of population-structure parameters for a hierarchy of populations was given by Rousset (75).

The growing use of Bayesian methods to population genetics is reflected by several papers that use such methods to characterize population structure (30, 39, 40, 71). Allele frequencies are assumed to follow a Dirichlet distribution across populations, or a beta distribution in the case of loci with two alleles.

Non-Frequency Measures

Although θ is defined in terms of variances of allele frequencies, there are parallel measures that use other parameters. The fact that mutation at microsatellite markers is generally between pairs of alleles with similar numbers of repeat units suggests that allele size (i.e., number of repeats) can be used in place of allele frequency (79). Balloux & Goudet (5) and Balloux & Lugin-Moulin (6) were concerned with the case where the stepwise mutation model holds for microsatellite markers. They compared two estimators of the form $\sum_{\text{loci}} V_a / \sum_{\text{loci}} V_t$ where the variance components (V_a among populations and V_t total) were for allele frequencies (100) or allele sizes (57). They compared the estimators for data simulated under a finite island model and concluded that neither estimator was best overall, although the Weir-Cockerham estimator was better for higher levels of gene flow. Weir & Cockerham (100) pointed out that the performance of their estimator reflects the method they used for combining information over multiple alleles at a locus, and they predicted better behavior for higher values of θ. It is the magnitudes of the parameter, rather than the forces leading to those values, that should affect the quality of the estimator in the multiple-alleles case.

Merilä & Crnoka (56) compared estimates of θ from various genetic markers with an analogous quantity, Q_{ST}, defined for quantitative traits (83). The estimate is based on the genetic variances of an additive quantitative trait, V_a among populations and V_w within populations, and is given by $V_a/(V_a + 2V_w)$. If allele frequencies are available for the same loci that affect the quantitative trait, values of θ and Q_{ST} should be equal.

Estimation of Migration Rates

Molecular ecologists, in particular, have been interested in inferring migration rates from estimates of θ, usually by employing the equilibrium result for the infinite-island migration model: $\theta = 1/(1 + 4Nm)$. Here N is the effective population size of each island and m is the migration rate between each pair of islands. Because this is a monotonic transformation of θ, it is not clear that much is gained over simply presenting θ estimates, especially as real populations are unlikely to conform to the many assumptions that lead to this result (101). Cockerham & Weir (15, 16) discussed more general relationships between θ and m. Kinnison et al. (46) fitted Nm to estimated θ values without assuming equilibrium. A recent review is given by Rousset (74), and a multivariate normal approach was adopted by Tufto et al. (87). Analogous work uses estimates of θ to estimate effective population size (8, 90).

Allocation of Individuals to Populations

Even though the genetic variation within human populations tends to be much greater than that among populations, there is often sufficient genetic differentiation among populations, as described by θ, to allow individuals to be allocated

to populations. The problem was discussed for blood-type markers by Spielman & Smouse (82) and Smouse & Spielman (80). More recent studies, primarily by forensic scientists, have used microsatellite markers (10, 25, 52, 77, 78). Cornuet et al. (17) evaluated several methods for allocating individuals by assessing their behavior as functions of θ. Dawson & Belkhir (20) assessed the quality of their Bayesian method for assigning individuals to groups within a population by estimating θ from the resulting grouped data.

Forensic Applications

Genetic profiles are now widely used for human identification in a forensic setting, and also for inferring relationships in cases of disputed parentage or the identification of remains. The key question generally involves determining the probability of a set of profiles under alternative hypotheses about the sources of those profiles. In the simplest forensic situation where the profile of a suspect matches that of a stain found at the scene of a crime, this reduces to determining the probability that an unknown person in a population has the profile given that a suspect is known to have the profile (26). When allele frequencies are assumed to have a Dirichlet distribution over populations, this probability is a function of θ (3, 4), and forensic scientists routinely estimate θ for the populations with which they work (4, 30, 102).

ESTIMATION OF θ

The parameter θ provides a description of the relationship between pairs of alleles in a population. It could be defined as the probability that the two alleles are identical by descent, but this is restrictive in that its values are then constrained to lie in the range [0,1]. A more general definition is in terms of correlation coefficients, and can be expressed in terms of indicator variables x_{ju} for the jth allele in a sample:

$$x_{ju} = \begin{cases} 1 & \text{allele is of type } A_u \\ 0 & \text{otherwise.} \end{cases}$$

Then θ is the correlation between x_{ju} and $x_{j'u}$ for different alleles ($j \neq j'$), where the underlying expectation process is over replicates of the population. This correlation should be written as θ_u to allow for selection or mutation differences for different allelic types, but these differences generally are assumed not to exist. Although θ is designed to capture evolutionary variation, values of its estimates also reflect the sampling process leading to the data employed. Weir (97) made the distinction between genetic and statistical sampling for these two sources of variation. Another way of expressing this concept is to say that θ measures relatedness of pairs of alleles within a population relative to the total (i.e., the expected)

population and this is why Wright (104) used the notation F_{ST}, where S denotes subpopulation and T denotes the total population.

Under the random mating assumption, expectations of the indicator variables do not depend on the particular values of j, and

$$\mathcal{E}(x_{ju}) = p_u$$
$$\mathcal{E}\left(x_{ju}^2\right) = p_u$$
$$\mathcal{E}(x_{ju}x_{j'u}) = p_u^2 + p_u(1-p_u)\theta, \quad j \neq j',$$

where p_u is the population frequency of allele A_u, an expected value over replicates of the population. The expression for $\mathcal{E}(x_{ju}x_{j'u})$ can be taken as a definition of θ, and clearly $\mathrm{Var}(x_{ju}) = p_u(1-p_u)$, $\mathrm{Cov}(x_{ju}, x_{j'u}) = p_u(1-p_u)\theta$ so that θ is indeed a correlation coefficient over replicate populations.

It may be convenient to write the expected value of $x_{ju}x_{j'u}$ as $P_{u,u}$, the probability with which the two alleles are both of type A_u. However, for a population mating by random union of gametes, this quantity is the same as the homozygote frequency P_{uu}. For nonrandom mating populations, it is necessary to distinguish the cases where the alleles are in the same or different individuals and the indicator variables need to be defined as x_{jku} for the kth allele in the jth individual. Expectations are then

$$\mathcal{E}(x_{jku}) = p_u$$
$$\mathcal{E}\left(x_{jku}^2\right) = p_u$$
$$\mathcal{E}(x_{jku}x_{j'k'u}) = \begin{cases} p_u^2 + p_u(1-p_u)F, & j = j', k \neq k' \\ p_u^2 + p_u(1-p_u)\theta & j \neq j' \end{cases},$$

where F is the total inbreeding coefficient (sometimes written as F_{IT}). Then $P_{uu} = p_u^2 + p_u(1-p_u)F$ differs from $P_{u,u} = p_u^2 + p_u(1-p_u)\theta$.

Because θ refers to variation over the evolutionary process, it cannot be estimated from a sample from a single population. Inferences made from a single sample are for within-population parameters such as the within-population inbreeding coefficient f, or F_{IS}. This quantity satisfies $f = (F-\theta)/(1-\theta)$, and it describes the relationship of pairs of alleles within individuals relative to that between individuals within the same population. There is generally little interest in within-population analogs of θ, as the point of estimating θ is to make inferences about evolutionary processes.

MOMENT ESTIMATES

With the assumption of no local inbreeding, $F_{IS} = 0$, $F_{IT} = F_{ST} = \theta$, estimation of θ makes use only of sample allele frequencies, although these need to be inferred from sample genotype frequencies. Second moments of allele frequencies can be

expressed in terms of θ, suggesting that estimators can be constructed from sample second moments.

Overall Estimates

The variation described by θ is estimated in practice from allele frequency variation among different populations, and it has been customary to regard extant populations as providing the replicates inherent in its definition. This carries the assumption that each sampled population has the same θ value, and this will now be relaxed. To distinguish the populations sampled, an index i is added to the indicator variables for the ith sample. A general set of expectations for the jth allele in the ith sample are

$$\mathcal{E}(x_{iju}) = p_u$$

$$\mathcal{E}\left(x_{iju}^2\right) = p_u$$

$$\mathcal{E}(x_{iju}x_{i'j'u}) = \begin{cases} p_u^2 + p_u(1-p_u)\theta_i & i = i',\ j \neq j' \\ p_u^2 + p_u(1-p_u)\theta_{ii'} & i \neq i' \end{cases}.$$

Each population is assumed to have the same (expected) allele frequency. Weir & Cockerham (100) assumed that $\theta_{ii'} = 0$ for all $i' \neq i$. Later they relaxed those assumptions (15, 98).

Sample allele frequencies are denoted by tildes, and the average frequency over samples is denoted by a bar. If there are n_i alleles sampled from the ith of r populations:

$$\tilde{p}_{iu} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{iju}$$

$$\bar{p}_u = \frac{1}{\sum_i n_i}\sum_{i=1}^{r} n_i \tilde{p}_{iu},$$

so that

$$\mathcal{E}(\tilde{p}_{iu}) = p_u$$

$$\mathcal{E}(\bar{p}_u) = p_u$$

$$\mathrm{Var}(\tilde{p}_{iu}) = \frac{1}{n_i}p_u(1-p_u)[1+(n_i-1)\theta_i] \qquad 1.$$

$$\mathrm{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u}) = p_u(1-p_u)\theta_{ii'}. \qquad 2.$$

Subsequent developments are simplified with additional notation:

$$\pi_u = p_u(1-p_u)$$

$$\phi_i = \frac{1}{n_i}[1+(n_i-1)]\theta_i]$$

Equations 1, 2 can be taken as defining the θ parameters and therefore can serve as a starting point. They could be derived by considering two sets of expectations, one within (W) and one among (A) populations. If p_{iu} is the frequency of allele A_u in the ith population, the usual multinomial distribution gives:

$$\left.\begin{aligned}\mathcal{E}_W(\tilde{p}_{iu}) &= p_{iu}\\ \mathrm{Var}_W(\tilde{p}_{iu}) &= \frac{1}{n_i}p_{iu}(1-p_{iu}).\end{aligned}\right\} \quad 3.$$

Among populations, the moments are

$$\left.\begin{aligned}\mathcal{E}_A(p_{iu}) &= p_u\\ \mathrm{Var}_A(p_{iu}) &= p_u(1-p_u)\theta_i\end{aligned}\right\} \quad 4.$$

to introduce the θ's. The method of moments for estimating θ makes no more statements concerning the distribution of the p_{iu}'s about p_u. Balding & Nichols (3, 4) assumed a Dirichlet distribution with parameters $(1-\theta_i)p_u/\theta_i$ for A_u which also gives Equations 4, as does the normal distribution $N(p_u, \pi_u\theta_i)$ assumed by Nicholson et al. (60). Combining Equations 3 and 4 leads to Equations 1 and 2, emphasizing that expectations in such equations are total (within and among populations). Foulley & Hill (31) contrasted the use of the normal and Dirichlet distributions.

When it is assumed that $\theta_i = \theta$ for all i and $\theta_{ii'} = 0$ for all $i \neq i'$, Weir & Cockerham (100) note that there are two unknown quantities, π_u and θ, and define two mean squares. In the notation of Weir (97):

$$\mathrm{MSP}_u = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\tilde{p}_{iu} - \bar{p}_u)^2$$

$$\mathrm{MSG}_u = \frac{1}{\sum_{i=1}^{r}(n_i-1)}\sum_{i=1}^{r} n_i\tilde{p}_{iu}(1-\tilde{p}_{iu}).$$

The average allele frequency $\bar{p}_u$ includes sample size weights. An alternative is to use an unweighted average $\bar{p}_u^* = \sum_{i=1}^{r}\tilde{p}_{iu}/r$. Estimates based on $\bar{p}_u$ or $\bar{p}_u^*$ will be better when θ or $(1-\theta)/n_i$, respectively, are larger. Following Robertson (69), a weighted estimate could be obtained from the two.

Under the general model, the mean squares have expected values

$$\mathcal{E}(\mathrm{MSP}_u) = \frac{\pi_u}{r-1}\left[\sum_{i=1}^{r} n_{ic}\phi_i - \frac{1}{\sum_{i=1}^{r} n_i}\sum_{\substack{i,i'=1\\ i\neq i'}}^{r} n_i n_{i'}\theta_{ii'}\right]$$

$$\mathcal{E}(\mathrm{MSG}_u) = \frac{\pi_u}{\sum_{i=1}^{r}(n_i=1)}\left(\sum_{i=1}^{r} n_i - \sum_{i=1}^{r} n_i\phi_i\right),$$

where $n_{ic} = n_i - n_i^2 / \sum_{i=1}^{r} n_i$. There are two special cases that lead to simplification.

In the special case that $\theta_i = \theta$ for all i and $\theta_{ii'} = 0$ for all $i \neq i'$,

$$\mathcal{E}(\mathrm{MSP}_u) = \pi_u[(1-\theta) + n_c\theta]$$
$$\mathcal{E}(\mathrm{MSG}_u) = \pi_u(1-\theta),$$

where

$$n_c = \frac{1}{r-1}\left(\sum_{i=1}^{r} n_i - \frac{\sum_{i=1}^{r} n_i^2}{\sum_{i=1}^{r} n_i}\right) = \frac{1}{r-1}\sum_{i=1}^{r} n_{ic}.$$

This led Weir & Cockerham (100) to their moment estimator of θ:

$$\hat{\theta}_{Mu} = \frac{\mathrm{MSP}_u - \mathrm{MSG}_u}{\mathrm{MSP}_u + (n_c - 1)\mathrm{MSG}_u}.$$

To the extent that the expected value of this quantity is the ratio of expectations of its numerator and denominator, it is unbiased for θ.

In the special case of balanced data, $n_i = n$ for all i,

$$\mathcal{E}(\mathrm{MSP}_u) = \pi_u[(1-\theta_w) + n(\theta_w - \theta_a)]$$
$$\mathcal{E}(\mathrm{MSG}_u) = \pi_u(1-\theta_w),$$

where

$$\theta_w = \frac{1}{r}\sum_{i=1}^{r}\theta_i$$
$$\theta_a = \frac{1}{r(r-1)}\sum_{\substack{i,i'=1 \\ i\neq i'}}^{r}\theta_{ii'},$$

so that the moment estimate, now written as $\hat{\beta}$, is providing an estimate of $(\theta_w - \theta_a)/(1-\theta_a)$. This result should also hold if all of the sample sizes are large and approximately equal. In general, however, the usual moment estimate is of a complex function of the θ_i's and $\theta_{ii'}$'s. Alternative statistics lead to estimates of weighted averages of θ_i's and $\theta_{ii'}$'s, as shown below.

Under the assumption that the same value of θ applies to each allele at a locus, Weir & Cockerham (100) combined information over alleles by summing numerator and denominator separately

$$\hat{\theta}_M = \frac{\sum_{u=1}^{m}(\mathrm{MSP}_u - \mathrm{MSG}_u)}{\sum_{u=1}^{m}[\mathrm{MSP}_u + (n_c - 1)\mathrm{MSG}_u]}, \qquad 5.$$

and they found by simulation that this method of weighting over alleles generally provides low bias and variance. No explicit account is taken of the correlation among frequencies of different alleles. If data are collected from a series of L loci, and if θ is assumed to apply equally to each locus, then an obvious extension is to add mean squares over loci:

$$\hat{\theta}_M = \frac{\sum_{l=1}^{L}\sum_{u=1}^{m_l}(\mathrm{MSP}_{lu} - \mathrm{MSG}_{lu})}{\sum_{l=1}^{L}\sum_{u=1}^{m_l}[\mathrm{MSP}_{lu} + (n_c - 1)\mathrm{MSG}_{lu}]}.$$

Properties of Moment Estimate

Because of the difficulty in describing the properties of ratio estimates, Dodds (21) and Weir (97) suggested numerical resampling for obtaining the sampling distribution of $\hat{\theta}_M$. Resampling over populations would change the structure of the data, but resampling over loci would exploit the assumption that (unlinked) loci provide independent replicates of the evolutionary process. Resampling was also used by Raymond & Rousset (63). Jiang (43) used a Taylor series expansion and approximate higher-order moments of sample allele frequencies to obtain the mean and variance of $\hat{\theta}_M$. Li (50) appealed to asymptotic theory to show that the mean square MSP_u has a chi-square distribution in the two-allele case,

$$\mathrm{MSP}_u \sim \pi_u[1 + (n_c - 1)\theta]\chi^2_{(r-1)},$$

and that the mean square MSG_u tends to a constant value of $\pi_u(1 - \theta)$. This assumes that the θ_i's are equal and that the $\theta_{ii'}$'s are zero. These results allowed her to derive expressions for the mean and variance of $\hat{\theta}$:

$$\mathcal{E}(\hat{\theta}_M) = \theta - \frac{2(1-\theta)}{r-1}\left(\frac{1 + (n_c - 1)\theta}{n_c}\right)^2$$

$$\mathrm{Var}(\hat{\theta}_M) = \frac{2(1-\theta)^2}{r-1}\left(\frac{1 + (n_c - 1)\theta}{n_c}\right)^2.$$

The variance formula differs slightly from the variance of the intraclass correlation given by Fisher (29), but is equal to that result for large sample sizes.

Population-Specific Estimates

If independent populations have different values of θ_i, maybe reflecting the differences in population size or differences in environmental influences, there is the danger of having an over-parameterized model. There are r independent sample allele frequencies $\tilde{p}_{iu}$ for allele A_u. In the two-allele case, this means r observations but $(r + 1)$ parameters: the frequency p_u and the r values of θ_i. It is possible to construct estimates, but they will not be unique. For $m > 2$ alleles at a locus, however, there are more $[r(m - 1)]$ independent sample allele frequencies than

there are parameters: $(m-1)$ parameters p_u plus r parameters θ_i. Similarly, for $L>1$ diallelic loci, there are more observations (rL allele frequencies) than there are parameters (L allele frequencies and r θ's). The following discussion assumes that there are at least as many allele frequencies in the data as there are parameters to be estimated.

If the terms in the mean square within populations are weighted by n_{ic} instead of n_i, the sums of squares corresponding to MSP and MSG have expectations

$$\mathcal{E}\left[\sum_{i=1}^{r} n_i(\tilde{p}_{iu}-\bar{p}_u)^2\right] = \pi_u\left[\sum_{i=1}^{r} n_{ic}\phi_i - \frac{1}{\sum_{i=1}^{r} n_i}\sum_{\substack{i,i'=1\\ i\neq i'}}^{r} n_i n_{i'}\theta_{ii'}\right]$$

$$\mathcal{E}\left[\sum_{i=1}^{r} n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})\right] = \pi_u\left[\sum_{i=1}^{r} n_{ic} - \sum_{i=1}^{r} n_{ic}\phi_i\right],$$

suggesting that, for independent populations ($\theta_{ii'}=0$), π_u can be estimated as

$$\hat{\pi}_u = \frac{\sum_{i=1}^{r} n_i(\tilde{p}_{iu}-\bar{p}_u)^2 + \sum_{i=1}^{r} n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_{i=1}^{r} n_{ic}}.$$

Therefore, from the relationship

$$\mathcal{E}\left[\sum_{u=1}^{m} \tilde{p}_{iu}(1-\tilde{p}_{iu})\right] = \left(\sum_{u=1}^{m}\pi_u\right)(1-\phi_i),$$

a moment estimate of ϕ_i for independent populations is

$$\hat{\phi}_i = 1 - \frac{\left(\sum_{i=1}^{r} n_{ic}\right)\sum_{u=1}^{m}\tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_{u=1}^{m}\sum_{i=1}^{r}\left[n_i(\tilde{p}_{iu}-\bar{p}_u)^2 + n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})\right]}. \qquad 6.$$

The estimate of the mean of the ϕ_i's is

$$\hat{\bar{\phi}} = 1 - \frac{\left(\sum_{i=1}^{r} n_{ic}\right)\sum_{u=1}^{m}\sum_{i=1}^{r}\tilde{p}_{iu}(1-\tilde{p}_{iu})}{r\sum_{u=1}^{m}\sum_{i=1}^{r}\left[n_i(\tilde{p}_{iu}-\bar{p}_u)^2 + n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})\right]}.$$

When the sample sizes are equal, $n_i = n$ for all i,

$$\hat{\phi}_i = 1 - \frac{\sum_{u=1}^{m} \tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_{u=1}^{m}\left[\frac{1}{r-1}\sum_{i=1}^{r}(\tilde{p}_{iu}-\bar{p}_u)^2 + \frac{1}{r}\sum_{i=1}^{r}\tilde{p}_{iu}(1-\tilde{p}_{iu})\right]}.$$

Further, when the number r of samples is large

$$\hat{\phi}_i \approx 1 - \frac{\sum_{u=1}^{m} \tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_{u=1}^{m} \bar{p}_u(1-\bar{p}_u)}$$

$$\hat{\bar{\phi}} \approx \frac{\sum_{u=1}^{m}\sum_{i=1}^{r}(\tilde{p}_{iu}-\bar{p}_u)^2}{r\sum_{u=1}^{m} \bar{p}_u(1-\bar{p}_u)}.$$

For each independent locus indexed by $l = 1, 2, \ldots L$, the estimate of ϕ_i may be written as $1 - x_{li}/y_l$ where

$$x_{li} = \sum_{u=1}^{m} \tilde{p}_{liu}(1-\tilde{p}_{liu})$$

$$y_l = \frac{1}{\sum_{i=1}^{r} n_{lic}} \sum_{u=1}^{m}\sum_{i=1}^{r}\left[n_{li}(\tilde{p}_{liu}-\bar{p}_{lu})^2 + n_{lic}\tilde{p}_{liu}(1-\tilde{p}_{liu})\right],$$

showing the addition of locus subscripts on sample sizes and allele frequencies. These terms have expectations

$$\mathcal{E}(x_{li}) = (1-\phi_i)\sum_{u=1}^{m_l} \pi_{lu}$$

$$\mathcal{E}(y_l) = \sum_{u=1}^{m_l} \pi_{lu}.$$

Information from loci with the same values of ϕ_i can be combined as for the earlier Weir & Cockerham estimator (100): $\hat{\phi}_i = 1 - (\sum_l x_{li})/(\sum_l y_l)$. The sampling distribution of this combined estimate may be found by bootstrapping over loci if L is not small.

Nicholson et al. (60) were especially interested in SNP loci, which generally have only two alleles. In that case, the two summands in the sums over alleles u

are the same and only one needs to be used. If $\tilde{p}_i$ is the frequency of one of the alleles at a locus, the equal sample size estimate is

$$\hat{\phi}_i = 1 - \frac{\tilde{p}_i(1-\tilde{p}_i)}{\frac{1}{r-1}\sum_{i=1}^{r}(\tilde{p}_i - \bar{p})^2 + \frac{1}{r}\sum_{i=1}^{r}\tilde{p}_i(1-\tilde{p}_i)},$$

and, for a large number of samples,

$$\hat{\phi}_i \approx 1 - \frac{\tilde{p}_i(1-\tilde{p}_i)}{\bar{p}(1-\bar{p})}.$$

Averaging over samples recovers the "classical" estimate (27)

$$\hat{\phi} \approx \frac{\sum_{i=1}^{r}(\tilde{p}_i - \bar{p})^2}{r\,\bar{p}(1-\bar{p})}.$$

Care is needed in interpreting the values of the estimates $\hat{\phi}_i$, as differences may reflect differences among the sample sizes n_i or among the coefficients θ_i, or both.

When the populations are not independent, $\theta_{ii'} \neq 0$, the estimate of ϕ_i shown in Equation 6 is actually estimating $(\phi_i - \theta_A)/(1-\theta_A)$, where

$$\theta_A = \frac{\sum_{\substack{i,i'=1 \\ i\neq i'}}^{r} n_i n_{i'} \theta_{ii'}}{\sum_{\substack{i,i'=1, \\ i\neq i'}}^{r} n_i n_{i'}}.$$

The weighted average θ_A reduces to the simple arithmetic mean, θ_a, of the $\theta_{ii'}$'s when the sample sizes are equal. An estimate of $\beta_{ii'} = (\theta_{ii'} - \theta_A)/(1-\theta_A)$ is given by

$$\beta_{ii'} = \frac{\theta_{ii'} - \theta_A}{1-\theta_A} \hat{=} 1 - \frac{\left(\sum_{i=1}^{r} n_{ic}\right)\sum_{u=1}^{m}\left[\tilde{p}_{iu}(1-\tilde{p}_{i'u}) + \tilde{p}_{i'u}(1-\tilde{p}_{iu})\right]}{2\sum_{u=1}^{m}\sum_{i=1}^{r}\left[n_i(\tilde{p}_{iu} - \bar{p}_u)^2 + n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})\right]}. \quad 7.$$

where $\hat{=}$ denotes "is estimated by." These estimates sum to zero. In the case of only two samples, this estimate is zero as required. The corresponding single-population equation is

$$\beta_i = \frac{\theta_i - \theta_A}{1-\theta_A} \hat{=} 1 - \frac{\left(\sum_{i=1}^{r} n_{ic}\right)\sum_{u=1}^{m}\frac{n_i}{n_i - 1}\tilde{p}_{iu}(1-\tilde{p}_{iu})}{\sum_{u=1}^{m}\sum_{i=1}^{r}[n_i(\tilde{p}_{iu} - \bar{p}_u)^2 + n_{ic}\tilde{p}_{iu}(1-\tilde{p}_{iu})]}. \quad 8.$$

This is to replace Equation 6, although the difference between them is trivial for large sample sizes.

By analogy to θ_A, the weighted average θ_W can be defined as

$$\theta_W = \frac{\sum_{i=1}^{r} n_i \theta_i}{\sum_{i=1}^{r} n_i},$$

which reduces to the simple arithmetic average, θ_w, when the sample sizes are equal. The quantity $\beta_W = (\theta_W - \theta_A)/(1 - \theta_A)$ can be estimated as

$$\hat{\beta}_W = 1 - \frac{\left(\sum_{i=1}^{r} n_{ic}\right) \sum_{u=1}^{m} \frac{n_i^2}{n_i - 1} \tilde{p}_{iu}(1 - \tilde{p}_{iu})}{\left(\sum_{i=1}^{r} n_i\right) \sum_{u=1}^{m} \sum_{i=1}^{r} \left[n_i (\tilde{p}_{iu} - \bar{p}_u)^2 + n_{ic} \tilde{p}_{iu}(1 - \tilde{p}_{iu})\right]}. \quad 9.$$

For equal sample sizes this reduces to the estimator in Equation 5 given by Weir & Cockerham (100). Because it serves as an estimator in the case of unequal sample sizes, however, it may be preferred to the Weir & Cockerham estimator.

There are two unsatisfactory aspects of this development. In the first place, it is seen that the quantities being estimated depend on the sample sizes, unless those sizes are equal. A more serious problem is the involvement of the average between-population relatedness quantity θ_A. Unless there are grounds for assuming this quantity is zero, all estimates are relative to that value. This does not prevent a comparison among the values of θ_i or $\theta_{ii'}$, but it does prevent their absolute value being estimated. There is the same need for a reference population when inbreeding coefficients F_{IT} are to be estimated. The issue is similar to that faced in the reconstruction of phylogenetic trees. Trees cannot be rooted unless there is information from an outgroup.

Finally, for large numbers of large samples,

$$\frac{\theta_i - \theta_A}{1 - \theta_A} \hat{=} 1 - \frac{\sum_{u=1}^{m} \tilde{p}_{iu}(1 - \tilde{p}_{iu})}{\sum_{u=1}^{m} \bar{p}_u(1 - \bar{p}_u)} \quad 10.$$

$$\frac{\theta_{ii'} - \theta_A}{1 - \theta_A} \hat{=} 1 - \frac{\sum_{u=1}^{m} [\tilde{p}_{iu}(1 - \tilde{p}_{i'u}) + \tilde{p}_{i'u}(1 - \tilde{p}_{iu})]}{2 \sum_{u=1}^{m} \bar{p}_u(1 - \bar{p}_u)}. \quad 11.$$

NORMAL THEORY APPROACH

Moment estimators have the property of being unbiased but little else is known about their sampling properties. If the sampling distribution for the data is known, then likelihood methods can be employed. If individuals, and hence genotypes, are sampled randomly from a single population their counts follow a multinomial distribution among samples from the same population. When there is random union of gametes in the population, allele counts are also multinomially distributed over samples from the population. For large samples, the multinomial distribution can be approximated by the multivariate normal distribution, and it will now be assumed that the normal distribution applies also across populations. Normality has also been assumed by previous authors (51, 60, 81, 87). If $\tilde{\mathbf{P}}$ is the vector of sample allele frequencies:

$$\tilde{\mathbf{P}} \sim \text{MVN}(\mathbf{P}, \mathbf{V}),$$

where

$$\tilde{\mathbf{P}} = \begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \\ \cdots \\ \tilde{\mathbf{p}}_r \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} \mathbf{p} \\ \mathbf{p} \\ \cdots \\ \mathbf{p} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1r} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2r} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{V}_{r1} & \mathbf{V}_{r2} & \cdots & \mathbf{V}_{rr} \end{bmatrix}.$$

The vectors $\tilde{\mathbf{p}}_i$ and $\mathbf{p}$ have $(m-1)$ components $\tilde{p}_{iu}$ and p_u, one for each of $(m-1)$ of the alleles at the locus. The $(m-1) \times (m-1)$ matrices $\mathbf{V}_{ii'}$ have elements $V_{ii'uu'}$. When $i = i'$ and $u = u'$ these elements are the variances of $\tilde{p}_{iu}$, otherwise they are the covariances of $\tilde{p}_{iu}$ and $\tilde{p}_{i'u'}$. Their values are:

$$V_{ii'uu'} = \begin{cases} p_u(1-p_u)\phi_i & i = i', u = u' \\ -p_u p_{u'}\phi_i & i = i', u \neq u' \\ p_u(1-p_u)\theta_{ii'} & i \neq i', u = u' \\ -p_u p_{u'}\theta_{ii'} & i \neq i', u \neq u'. \end{cases}$$

Overall Estimate

If there is no relationship among alleles from different populations, $\theta_{ii'} = 0$, then the vectors $\tilde{\mathbf{p}}_i$ are independent. These vectors also have the same expected value, but they have the same variances only if the ϕ_i values are the same. Unless the sample sizes are very large, this requires not only equal θ_i values, but also equal sample sizes n_i. Suppose now that $\phi_i = \phi$, because $\theta_i = \theta$ and because the n_i's are either equal or so large that they are approximately equal. The sample allele frequency vectors $\tilde{\mathbf{p}}_i$ are then independently and identically distributed and, from

standard theory, the quadratic form

$$Q = \sum_{i=1}^{r} (\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})' \mathbf{V}_{ii}^{-1} (\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})$$

$$= \frac{1}{\phi} \sum_{i=1}^{r} \sum_{u=1}^{m} \frac{(\tilde{p}_{iu} - \bar{p}_u)^2}{\bar{p}_u}$$

has a chi-square distribution

$$Q \sim \phi \chi^2_{(r-1)(m-1)}.$$

The mean allele frequencies are $\bar{p}_u = \sum_{i=1}^{r} n_i \tilde{p}_{iu} / \sum_{i=1}^{r} n_i$ as before, and the estimate of the common value θ is

$$\hat{\theta}_N = \frac{1}{n-1} \left(\frac{n}{(r-1)(m-1)} \sum_{i=1}^{r} \sum_{u=1}^{m} \frac{(\tilde{p}_{iu} - \bar{p}_u)^2}{\bar{p}_u} - 1 \right) \qquad 12.$$

when the sample sizes are equal, or

$$\hat{\theta}_N = \frac{1}{(r-1)(m-1)} \sum_{i=1}^{r} \sum_{u=1}^{m} \frac{(\tilde{p}_{iu} - \bar{p}_u)^2}{\bar{p}_u} \qquad 13.$$

when the sample sizes are large (70). If data are available from L independent loci, the lth of which has m_l alleles, the sum over loci of the quadratic forms has a chi-square distribution with $d = (r-1) \sum_{l=1}^{L} (m_l - 1)$ df, and the estimates are simply averaged over loci.

From the properties of the chi-square distribution

$$\mathcal{E}(\hat{\theta}_N) = \theta$$

$$\mathrm{Var}(\hat{\theta}_N) = \frac{2[1 + (n-1)\theta]^2}{(n-1)^2 d} \approx \frac{2\theta^2}{d}.$$

Similar expressions were given by Foulley & Hill (31).

The chi-square distribution also provides confidence intervals. For example, if $X_{0.025}$ and $X_{0.975}$ are the 2.5*th* and 97.5*th* percentiles of the χ^2_d distribution, a 95% confidence interval is

$$\left(\frac{d}{X_{0.975}} \left[\hat{\theta}_N + \frac{1}{n-1} \right] - \frac{1}{n-1}, \frac{d}{X_{0.025}} \left[\hat{\theta}_N + \frac{1}{n-1} \right] - \frac{1}{n-1} \right)$$

for equal sample sizes, and

$$\left(\frac{d\hat{\theta}_N}{X_{0.975}}, \frac{d\hat{\theta}_N}{X_{0.025}} \right)$$

for large sample sizes.

Population-Specific Estimates

When the populations are independent, $\theta_{ii'} = 0$ for all $i \neq i'$, but with different values of θ_i, the variance matrix $\mathbf{V}$ can be written as a Kronecker product:

$$\mathbf{V} = \mathbf{\Pi} \otimes \mathbf{\Phi},$$

where

$$\mathbf{\Pi} = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots \\ -p_1p_2 & p_2(1-p_2) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1 & 0 & \cdots \\ 0 & \phi_2 & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}.$$

If there are r samples and m alleles at the locus, $\mathbf{V}$ has determinant

$$|\mathbf{V}| = \left(\prod_{i=1}^{r} \phi_i\right)^m \left(\prod_{u=1}^{m} p_i\right)^r$$

and inverse

$$\mathbf{V}^{-1} = \mathbf{\Phi}^{-1} \otimes \mathbf{\Pi}^{-1},$$

where

$$\mathbf{\Pi}^{-1} = \begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_m} & \frac{1}{p_m} & \cdots \\ \frac{1}{p_m} & \frac{1}{p_2} + \frac{1}{p_m} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

$$\mathbf{\Phi}^{-1} = \begin{bmatrix} \frac{1}{\phi_1} & 0 & \cdots \\ 0 & \frac{1}{\phi_2} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}.$$

Ignoring terms that do not include the parameters of interest in likelihood expressions, the log-likelihood function is

$$\begin{aligned} \ln L &= -\frac{1}{2}\ln(|\mathbf{V}|) - \frac{1}{2}(\tilde{\mathbf{P}} - \mathbf{P})'\mathbf{V}^{-1}(\tilde{\mathbf{P}} - \mathbf{P}) \\ &= -\frac{m}{2}\sum_{i=1}^{r}\ln(\phi_i) - \frac{r}{2}\sum_{u=1}^{m}\ln(p_u) - \frac{1}{2}\sum_{i=1}^{r}\sum_{u=1}^{m}\frac{(\tilde{p}_{iu} - p_u)^2}{\phi_i p_u}. \end{aligned}$$

Because the p_u's sum to one, it is necessary to add a Lagrangian term before maximizing this function in order to find the maximum likelihood estimates of the

p_u's and ϕ_i's. The modified function and its derivatives are

$$\ln L = -\frac{m}{2}\sum_{i=1}^{r}\ln(\phi_i) - \frac{r}{2}\sum_{u=1}^{m}\ln(p_u) - \frac{1}{2}\sum_{i=1}^{r}\sum_{u=1}^{m}\frac{\tilde{p}_{iu}^2}{\phi_i p_u}$$

$$+\frac{1}{2}\sum_{i=1}^{r}\frac{1}{\phi_i} + \lambda\left(\sum_{u=1}^{m} p_u - 1\right)$$

$$\frac{\partial \ln L}{\partial \phi_i} = -\frac{m}{2\phi_i} - \frac{1}{2}\sum_{u=1}^{m}\frac{\tilde{p}_{iu}^2}{\phi_i^2 p_u} - \frac{1}{2\phi_i^2}$$

$$\frac{\partial \ln L}{\partial p_u} = -\frac{r}{2p_u} + \frac{1}{2}\sum_{i=1}^{r}\frac{\tilde{p}_{iu}^2}{\phi_i p_u^2} + \lambda$$

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{u=1}^{m} p_u - 1.$$

Setting the derivatives to zero provides equations that need to be solved numerically. One approach would be to iterate

$$\phi_i = \frac{1}{m}\sum_{u=1}^{m}\frac{(\tilde{p}_{iu} - p_u)^2}{p_u} \qquad 14.$$

$$p_u = \frac{\sum_{i=1}^{r}\left(1 - \frac{\tilde{p}_{iu}^2}{\phi_i p_u}\right)}{\sum_{u=1}^{m}\sum_{i=1}^{r}\left(1 - \frac{\tilde{p}_{iu}^2}{\phi_i p_u}\right)}.$$

The θ_i's are then recovered from the ϕ_i's.

In the special case of equal ϕ_i's (which implies equal sample sizes as well as equal θ_i's), the log-likelihood becomes

$$\ln L = -\frac{rm}{2}\ln(\phi) - \frac{r}{2}\sum_{u=1}^{m}\ln(p_u) - \frac{1}{2\phi}\sum_{i=1}^{r}\sum_{u=1}^{m}\frac{\tilde{p}_{iu}^2}{p_u} + \frac{r}{2\phi} + \lambda\left(\sum_{u=1}^{m} p_u - 1\right)$$

This leads to the iterative equations

$$\phi = \frac{1}{rm}\sum_{i=1}^{r}\sum_{u=1}^{m}\frac{(\tilde{p}_{iu} - p_u)^2}{p_u}$$

$$p_u = \frac{\sum_{i=1}^{r}\left(1 - \frac{\tilde{p}_{iu}^2}{\phi p_u}\right)}{\sum_{u=1}^{m}\sum_{i=1}^{r}\left(1 - \frac{\tilde{p}_{iu}^2}{\phi p_u}\right)}.$$

A comparison with the estimate of θ in Equations 12 and 13 emphasizes that the maximum likelihood estimates of allele frequencies are not the sample allele frequencies (see Appendix), although the two will be equal for large m and r. It appears to be satisfactory in practice (simulation results not shown) to replace p_u in the estimates of ϕ_i and ϕ by the sample average values $\bar{p}_u$ and change the m divisor to $(m-1)$:

$$\hat{\theta}_{iN} = \frac{1}{n-1}\left(\frac{rn}{(r-1)(m-1)}\sum_{u=1}^{m}\frac{(\tilde{p}_{iu}-\bar{p}_u)^2}{\bar{p}_u}-1\right). \qquad 15.$$

Averaging the estimates from Equation 15 over samples gives the estimate in Equation 12 and there is a corresponding simplification for large sample sizes n. This approximation requires independent populations.

The advantage of the likelihood approach is that hypotheses about the ϕ_i's can be tested. The hypothesis $H_0: \phi_i = \phi$ can be tested by comparing the likelihoods maximized under no constraint and under the constraint of the hypothesis.

NUMERICAL RESULTS

The moment estimators discussed here were applied to the simple case of three populations having the tree structure shown in Figure 1. Data were simulated assuming a pure drift model, and means and standard deviations of estimates from 1000 replicates are shown in Table 1. The simulation was for a single locus with $m = 5$ alleles, all equally frequent initially. Population $i = 0$, of size 500 alleles, resulted from 5 generations of random mating. Population $i = 3$ was of size 300 alleles, and $t_1 + t_2$ was 20 generations. Population $i = 4$, of 500 alleles, resulted from $t_2 = 10$ generations of random mating from population $i = 0$.

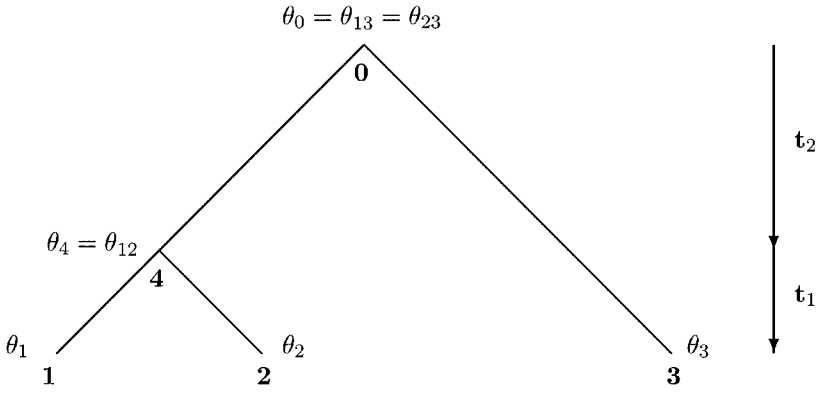

Figure 1 Three-population tree.

TABLE 1 Moment estimates, using Equations 7 and 8, for populations in Figure 1. (Parameter values given in text.)

	Populations						
	1	**2**	**3**	**1&2**	**1&3**	**2&3**	β_W
θ parameter	.210	.053	.076	.032	.010	.010	
β parameter*	.196	.036	.060	.015	−.007	−.007	.097
β estimate	.195	.033	.060	.017	−.008	−.008	.096
SD of estimate	.130	.047	.066	.049	.046	.037	.052

*$\beta = (\theta - \theta_A)/(1 - \theta_A)$

Populations $i = 1$ and $i = 2$, of 50 and 500 alleles, respectively, resulted from $t_1 = 10$ generations of random mating after population $i = 4$. All sample sizes were $n_i = 100$, $i = 1, 2, 3$.

The moment methods were then applied to data made publicly available by the FBI (12). Three samples, each of about 200 people, were collected from the United States and typed at 13 microsatellite markers, the "CODIS" set of loci. Sample properties for these loci are shown in Table 2: the locus name, the number of alleles m_l and the adjusted sample size terms n_{lc} for the lth locus. Estimates of the within-population coancestries θ_i are shown in Table 3, and of the between-population coancestries $\theta_{ii'}$ in Table 4.

TABLE 2 Sample properties of FBI data (12)

			Heterozygosity		
Locus	**No. Alleles**	**Sample size**	**AA**	**CA**	**HI**
D3S135	10	414.6	.763	.795	.719
vWA	10	385.5	.809	.811	.769
FGA	22	385.5	.863	.860	.878
D8S117	13	385.5	.778	.797	.792
D21S11	21	384.8	.861	.853	.811
D18S51	17	385.5	.873	.876	.875
D5S818	10	384.9	.739	.682	.718
D13S31	9	384.8	.688	.771	.827
D7S820	10	414.6	.782	.806	.772
CSF1PO	11	414.6	.781	.734	.707
TPOX	11	414.0	.763	.621	.607
THO1	8	414.6	.727	.783	.757
D16S53	8	412.6	.798	.767	.771

AA: African American, CA: Caucasian, HI: Hispanic.

TABLE 3 Single-population estimates, from Equation 8, for FBI data (12)

	β_i				
Locus	**AA**	**CA**	**HI**	**Average**	$\hat{\beta}_W$
D3S135	.010	−.030	.069	.017	.019
vWA	.000	−.002	.050	.017	.019
FGA	.007	.012	−.008	.003	.006
D8S117	.026	.003	.009	.012	.015
D21S11	−.012	−.003	.047	.012	.014
D18S51	.011	.008	.010	.010	.012
D5S818	−.018	.061	.012	.019	.021
D13S31	.132	.028	−.042	.036	.040
D7S820	.014	−.016	.026	.008	.011
CSF1PO	−.048	.015	.051	.006	.008
TPOX	−.118	.090	.112	.027	.030
THO1	.078	.008	.041	.043	.045
D16S53	−.011	.028	.024	.014	.016
All loci	.010	.017	.032	.020	.020

AA: African American, CA: Caucasian, HI: Hispanic.

The development based on normal theory shown above suggests that sample variances decrease with the number of alleles per locus, the number of loci, and the number of samples. The simulation results shown in Table 1 show rather large standard deviations for the case of only three samples, and this may account for the very large variation among loci for the results in Tables 3 and 4. Of course it may also be that the different loci are not providing replicates of the same evolutionary history. Loci may have been subjected to different selection pressures, for example, and variation among θ values has been suggested as a means of detecting selection, as recently reviewed by Vitalis et al. (91) and applied by Marshall & Ritland (55). If loci can be regarded as providing replication of the same process, however, then averaging over loci is appropriate. The variation among loci is much reduced when the three population-specific estimates are averaged, or when only a common value is estimated.

DISCUSSION

This review has extended Weir & Cockerham (100) in two directions. Most significantly, it has allowed the separate estimation of population- and population-pair specific values of θ. Previously it was assumed that populations were independent

TABLE 4 Two-population estimates, from Equation 7, for FBI data (12)

	$\hat{\beta}_{ii'}$			$\hat{\beta}_W$		
Locus	**AA&CA**	**AA&HI**	**CA&HI**	**AA&CA**	**AA&HI**	**CA&HI**
D3S135	−.018	.026	−.009	.010	.016	.030
vWA	−.018	.006	.010	.019	.021	.017
FGA	.006	−.002	−.004	.006	.004	.008
D8S117	−.012	.002	.009	.029	.018	.000
D21S11	−.008	−.008	.015	.003	.029	.010
D18S51	−.004	−.008	.011	.016	.021	.001
D5S818	.003	−.039	.033	.021	.037	.006
D13S31	.058	−.021	−.032	.026	.067	.026
D7S820	.001	.004	−.006	.000	.019	.013
CSF1PO	−.024	−.009	.034	.010	.012	.002
TPOX	−.043	−.053	.097	.030	.049	.007
THO1	−.034	.028	.006	.077	.035	.021
D16S53	−.009	−.009	.018	.020	.018	.011
Total	−.008	−.006	.014	.021	.023	.020

AA: African American, CA: Caucasian, HI: Hispanic.

and that either each population had the same value of θ or a population-average value was being estimated. The other extension has been the adoption of multivariate normal methods as an alternative to the method of moments. There may be an increase in computational burden and increase in bias with these methods, but there is the gain of a distributional form for the estimates.

Natural populations of the same species are unlikely to have the same value of θ, if only because they have different sizes. Although the reconstruction of intra-specific trees can proceed satisfactorily on the basis of the usual estimates of average θ values (65, 98), there are occasions when population-specific values are needed. There is the immediate issue of degrees of freedom. For r populations, there are r within-population values and $r(r-1)/2$ between-population values to be estimated. As there are $m-1$ independent allele frequencies for a locus with m alleles, there are only $r(m-1)$ independent observations in all, so only loci with large numbers of alleles can be used. With L loci, there is an increase in the number of observed allele frequencies to $Lr(m-1)$ and an increase to $r(r+1)/2 + L(m-1)$ parameters, so that even diallelic SNPs can be used. The constraints are less severe if the between-population coefficients $\theta_{ii'}$ are ignored, but it needs to be recognized that the estimates are then actually for a combination of within- and between-population values.

Under a pure drift model, values of θ are simple functions of population size and time. For a pair of populations, the values of θ within each can be expressed

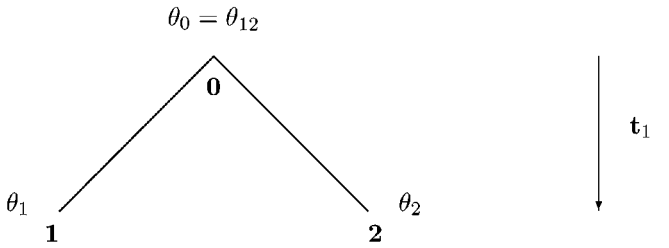

Figure 2 Two populations.

in terms of θ for their most recent common ancestral population. For the situation in Figure 2:

$$\theta_i = 1 - (1 - \theta_{12})X_i^{t_1}, \quad i = 1, 2,$$

where $X_i = (2N_i - 1)/2N_i$ and N_i is the constant population size for populations $i = 1, 2$. Therefore,

$$\beta_i = \frac{\theta_i - \theta_{12}}{1 - \theta_{12}} = 1 - X_i^{t_1} \approx \frac{t_1}{2N_i}.$$

The β parameters estimated by Equation 8 for a pair of populations are therefore furnishing estimates of the time since those populations diverged from an ancestral population. Although the two times must be the same, the pure drift model shows that the estimates will be different when the two population sizes are different. The estimate of Weir & Cockerham (100) is for

$$\beta_W = \frac{\theta_W - \theta_{12}}{1 - \theta_{12}} = 1 - \frac{X_1^{t_1} + X_2^{t_1}}{2}$$
$$\approx \frac{1}{2}\left(\frac{1}{2N_1} + \frac{1}{2N_2}\right)t_1 = \frac{t_1}{2N_h},$$

where N_h is the harmonic mean of the two population sizes. The quantity β_W is proportional to the divergence time t_1 (65).

If populations $i = 1, 2, 3, 4$ in Figure 1 have sizes N_i, and if $X_i = (2N_i - 1)/2N_i$:

$$\theta_{12} = 1 - (1 - \theta_0)X_4^{t_2}$$
$$\theta_i = 1 - (1 - \theta_{12})X_i^{t_1} = 1 - (1 - \theta_0)X_i^{t_1}X_4^{t_2}, \quad i = 1, 2$$
$$\theta_3 = 1 - (1 - \theta_0)X_3^{t_1 + t_2}$$
$$\theta_{13} = \theta_{23} = \theta_0.$$

The β parameters being estimated from the three extant populations 1, 2 and 3 involve the average between-population quantity $\theta_A = (\theta_{12} + 2\theta_0)/3$ although this

cancels out of the expressions needed to estimate the times:

$$\frac{\beta_i - \beta_{12}}{1 - \beta_{12}} = 1 - X_i^{t_1} \approx \frac{t_1}{2N_i}, \quad i = 1, 2$$

$$\frac{\beta_i - \beta_{i3}}{1 - \beta_{13}} = 1 - X_3^{t_1+t_2} \approx \frac{t_1 + t_2}{2N}, \quad i = 1, 2.$$

The θ's of interest can be expressed in terms of the estimable β's:

$$\frac{\theta_i - \theta_{12}}{1 - \theta_{12}} = \frac{\beta_i - \beta_{12}}{1 - \beta_{12}}, \quad i = 1, 2.$$

If θ_0 is assumed to be zero, the outgroup population 3 allows estimation of all three measures θ_1, θ_2 and θ_{12} for populations 1 and 2 since then $\beta_{12} = 2\theta_{12}/(3 - \theta_{12})$ and $\beta_i = (3\theta_i - \theta_{12})/(3 - \theta_{12}), i = 1, 2$.

Moment estimates of the θ's involve only the second moments of sample allele frequencies, whereas likelihood or Bayesian methods use the whole distribution. Higher-order moments can be expressed in terms of analogs of θ (96). Ignoring sample-size terms

$$\mathcal{E}(\tilde{p}_{iu} - p_u)^2 = p_u(1 - p_u)\theta$$

$$\mathcal{E}(\tilde{p}_{iu} - p_u)^3 = p_u(1 - p_u)(1 - 2p_u)\gamma$$

$$\mathcal{E}(\tilde{p}_{iu} - p_u)^4 = p_u(1 - p_u)(1 - 2p_u)(1 - 3p_u)\delta + 3p_u^2(1 - p_u)^2\Delta.$$

The normal distribution assumption implies that $\gamma = \delta = 0$, $\Delta = \theta^2$, or that there are no dependencies among a set of four alleles in addition to those between any pair of them. Assuming that allele frequencies have a Dirichlet distribution over populations, or that p_{iu} has a Beta distribution with parameters $(1 - \theta)p_u/\theta$ and $(1-\theta)(1-p_u)/\theta$ (4) implies that $\gamma = 2\theta^2/(1+\theta)$, $\delta = 6\theta^3/[(1+\theta)(1+2\theta)]$, $\Delta =$ (99). These relations hold for the infinite-allele mutation model, but not for the stepwise mutation model (35).

ACKNOWLEDGMENTS

This work was supported in part by NIH Grant GM 45344. Very helpful discussions were held with Peter Donnelly and George Nicholson, and the review was completed while the first author enjoyed the hospitality of both the Mathematical Genetics group in the Department of Statistics and the Wellcome Trust Centre for Human Genetics at the University of Oxford.

APPENDIX

The failure of the maximum likelihood estimate of mean allele frequencies to equal their observed values reflects, in part, the approximation of a multinomial distribution by a multivariate normal. In the general setting of a population with

proportions Q_u in the uth of m categories, the probability of category counts n_u in a sample of size $n = \sum_{u=1}^{m} n_u$ is

$$\Pr(\{n_u\}) = \frac{n!}{\prod_{u=1}^{m} n_u!} \prod_{u=1}^{m} (Q_u)^{n_u},$$

and the means, variances, and covariances of the counts are

$$\mathcal{E}(n_u) = nQ_u$$

$$\mathrm{Var}(n_u) = nQ_u(1 - Q_u)$$

$$\mathrm{Cov}(n_u, n_{u'}) = -nQ_uQ_{u'}, \quad u \neq u'.$$

The log-likelihood for the category probabilities is

$$\ln(L(\{Q_u\}) = \sum_{u=1}^{m} n_u \ln(Q_u).$$

To accommodate the dependency caused by $\sum_{u=1}^{m} Q_u = 1$, the Lagrange multiplier term $\lambda(1 - \sum_{u=1}^{m} Q_u)$ is added to the log-likelihood. Differentiating with respect to Q_u gives

$$\frac{\partial \ln(L)}{\partial Q_u} = \frac{n_u}{Q_u} - \lambda,$$

which leads to the maximum likelihood estimates (MLEs) $\hat{Q}_u = \tilde{Q}_u$ where $\tilde{Q}_u = n_u/n$.

For large sample sizes, the multivariate normal distribution provides a good approximation to the multinomial. The appropriate normal distribution for category counts will have variance matrix $n\mathbf{V}$ where $\mathbf{V}$ has uth diagonal element $Q_u(1 - Q_u)$ and off-diagonal elements $-Q_uQ_{u'}, u \neq u'$. Omitting the mth row and column removes the singularity of this matrix. The mean vector is then $n\mathbf{Q} = n[Q_1, Q_2, \ldots, Q_{m-1}]'$. The determinant of the reduced matrix is $\prod_{u=1}^{m} Q_u$ and its inverse has uth diagonal element $[1/(Q_u) + 1/(Q_m)]$ and all off-diagonal elements equal to $1/(Q_m)$. These results lead to the log-likelihood

$$\ln(L) = -\frac{1}{2} \ln\left(\prod_{u=1}^{m} Q_u\right) - \frac{1}{2} \sum_{u=1}^{m} \frac{(n_u - nQ_u)^2}{nQ_u} - \lambda\left(1 - \sum_{u=1}^{m} Q_u\right),$$

where the Lagrange multiplier λ allows all m unknowns Q_u to be included. Setting the derivative with respect to each Q_u equal to zero gives

$$\frac{1}{n}\left(\lambda - \frac{1}{2Q_u}\right) + \left(\frac{\tilde{Q}_u - Q_u}{Q_u} + \frac{(\tilde{Q}_u - Q_u)^2}{2Q_u^2}\right) = 0.$$

Only for large n will these equations are satisfied by $Q_u = \tilde{Q}_u$, so that $\hat{Q}_u = \tilde{Q}_u$ are

approximations to the MLEs in the normal approximation formulation. In general, however, the MLEs are not simply the observed values.

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

LITERATURE CITED

1. Anaya JM, Correa PA, Mantilla RD. 2002. Rheumatoid arthritis association in Colombian population is restricted to HLA-DRB1*04 QRRAA alleles. *Genes Immun.* 3:56–58
2. Balding DJ, Bishop M, Cannings C, eds. 2001. *Handbook of Statistical Genetics*. New York: Wiley. 890 pp.
3. Balding DJ, Nichols RA. 1994. DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.* 64:125–40
4. Balding DJ, Nichols RA. 1995. A method for characterizing differentiation between populations at multi-allelic loci and its implications for establishing identity and paternity. *Genetica* 96:3–12
5. Balloux F, Goudet J. 2002. Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Mol. Ecol.* 11:771–83
6. Balloux F, Lugon-Moulin N. 2002. The estimation of population differentiation with microsatellite markers. *Mol. Ecol.* 11:155–65
7. Banks SC, Skerratt LF, Taylor AC. 2002. Female dispersal and relatedness structure in common wombats (*Vombatus ursinis*). *J. Zool.* 256:389–99
8. Basset P, Balloux F, Perrin N. 2001. Testing demographic models of effective population size. *Proc. R. Soc. London Ser. B* 268:311–17
9. Blundell GM, Ben-David M, Groves P, Bowyers RT, Geffen E. 2002. Characteristics of sex-biased dispersal and gene flow in coastal river otters: implications for natural recolonization of extirpated populations. *Mol. Ecol.* 11:289–303
10. Brenner CH. 1998. Difficulties in the estimation of ethnic affiliation. *Am. J. Hum. Genet.* 62:1559–60
11. Broughton RE, Stewart LB, Gold JR. 2002. Microsatellite variation suggests substantial gene flow between king mackerel (*Scomberomorus cavalla*) in the western Atlantic Ocean and Gulf of Mexico. *Fish. Res.* 54:305–16
12. Budowle B, Moretti T. 1999. Genotype profiles for six population groups at the 13 CODIS short randem repeat core loci and other PCR-based loci. *Forensic Sci. Comm.* http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm
13. Charbonnel N, Angers B, Rastavonjizay R, Bremond P, Jarne P. 2002. Evolutionary aspects of the metapopulation dynamics of *Biomphataria pfeifferi*, the intermediate host of *Schistosoma mansoni*. *J. Evol. Biol.* 15:248–61
14. Chiappero MB, Sabattini MS, Blanco A, Calderon GE, Gardenal CN. 2002. Gene flow among *Calomys musculinis* (Rodentia, Muridae) populations in Argentina. *Genetica* 114:63–72
15. Cockerham CC, Weir BS. 1987. Correlations, descent measures: drift with migration and mutation. *Proc. Natl. Acad. Sci. USA* 84:8512–14
16. Cockerham CC, Weir BS. 1993. Estimation of gene flow from *F*-statistics. *Evolution* 47:855–63
17. Cornuet J-M, Piry S, Luikart G, Estoup A, Solignac M. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* 153:1989–2000
18. Davies N, Bermingham E. 2002. The historical biogeography of two Caribbean

butterflies (*Lepidoptera: Heliconiidae*) as inferred from genetic variation at multiple loci. *Evolution* 56:573–89
19. Dallimer M, Blackburn C, Jones PJ, Pemberton JM. 2002. Genetic evidence for male biased dispersal in the red-billed quelea *Quelea quelea*. *Mol. Ecol.* 3:529–33
20. Dawson KJ, Belkhir K. 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* 78:59–77
21. Dodds KG. 1986. *Resampling methods in genetics and the effect of family structure in genetic data*. PhD thesis, North Carolina State Univ. 110 pp.
22. Dufresne F, Bourget E, Bernatchez L. 2002. Differential patterns of spatial divergence in microsatellite and allozyme alleles: further evidence for locus-specific selection in the acorn barnacle, *Semibalanus balanoides*? *Mol. Ecol.* 11:113–23
23. Endsley MA, Baker MD, Krafsur ES. 2002. Microsatellite loci in the house fly *Musca domestica* L (Diptera: Muscidae). *Mol. Ecol. Notes* 2:72–74
24. Ernest HB, Rubin ES, Boyce WM. 2002. Fecal DNA analysis and risk assessment of mountain lion predation of bighorn sheep. *J. Wildlife Manage.* 66:75–85
25. Evett IW, Pinchin R, Buffery C. 1992. An investigation of the feasibility of inferring ethnic origin from DNA profiles. *J. Forensic Sci.* 32:301–6
26. Evett IW, Weir BS. 1998. *Interpreting DNA Evidence*. Sunderland, MA: Sinauer. 285 pp.
27. Excoffier L. 2001. Analysis of population subdivision. See Ref. 2, pp. 271–307
28. Fievet E, Eppe R. 2002. Genetic differentiation among populations of the amphidromous shrimp *Atya innocous* (HERBST) and obstacles to their upstream migration. *Arch. Hydrobiol.* 153: 287–300
29. Fisher RA. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1:3
30. Foreman LA, Lambert JA. 2000. Genetic differentiation within and between for UK ethnic groups. *Forensic Sci. Int.* 114:7–20
31. Foulley JL, Hill WG. 1999. On the precision of estimation of genetic distance. *Genet. Sel. Evol.* 31:457–64
32. Frydenberg J, Pertoldi C, Dahlgaard J. 2002. Genetic variation in original and colonizing *Drosophila buzzatii* populations analyzed by microsatellite loci isolated with a new PCR screening method. *Mol. Ecol.* 11:181–90
33. Gold JR, Turner TF. 2002. Population structure of red drum (*Sciaenops ocellatus*) in the northern Gulf of Mexico, as inferred from variation in nuclear encoded microsatellites. *Mar. Biol.* 140:249–65
34. Gomez A, Adcock GJ, Lunt DH, Carvalho GR. 2002. The interplay between colonization history and gene flow in passively dispersing zooplankton: microsatellite analysis of rotifer resting egg banks. *J. Evol. Biol.* 15:158–71
35. Graham J, Curran J, Weir BS. 2000. Conditional genotypic probabilities for microsatellite loci. *Genetics* 155:1973–80
36. Grimaldi MC, Crouau, Roy B, Contu L, Amoros JP. 2002. Molecular variation of HLA class I genes in the Corsican population: approach to its origin. *Eur. J. Immun.* 29:101–7
37. Hawkins SL, Varnavskya NV, Matzak EA, Efremov VV, Guthrie CM III, et al. 2002. Population structure of odd-broodline Asian pink salmon and its contrast to even–broodline structure. *J. Fish. Biol.* 60:370–88
38. Heath DD, Busch C, Kelly J, Atagi DY. 2002. Temporal change in genetic structure and effective population size in steelhead trout (*Oncorhynchus mykiss*). *Mol. Ecol.* 11:197–214
39. Holsinger KE. 1999. Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Hereditas* 130:245–55
40. Holsinger KE, Lewis PO, Dey DK. 2002. A Bayesian approach to inferring

population structure from dominant markers. *Mol. Ecol.* 11:1157–64
41. Hundertmark KJ, Shields GF, Udina IG. 2002. Mitochondrial phylogeography of moose (*Alces alces*): Late Pleistocene divergence and population expansion. *Mol. Phylogenet. Evol.* 22:375–87
42. Ivandic V, Hackett CA, Nevo E, Keith R, Thomas WTB, Forster BP. 2002. Analysis of simple sequence repeats (SSRs) in wild barley from the Fertile Crescent: associations with ecology, geography and flowering time. *Plant Mol. Biol.* 48:511–27
43. Jiang C. 1987. *Estimation of F-statistics in subdivided populations*. PhD thesis, North Carolina State Univ. 95 pp.
44. Jonsdottir ODB, Imsland AK, Danielsdottir AK, Marteinsdttir G. 2002. Genetic heterogeneity and growth properties of different genotypes of Atlantic cod (*Gadus morhua L.*) at two spawning sites off south Iceland. *Fish. Res.* 55:37–47
45. Kim KS, Choi CB. 2002. Genetic structure of Korean native pig using microsatellite markers. *Korean J. Genet.* 24:1–7
46. Kinnison MT, Bentzen B, Unwin MJ, Quinn TP. 2002. Reconstructing recent divergence: evaluating nonequilibrium population structure in New Zealand chinook salmon. *Mol. Ecol.* 11:739–54
47. Krafsur ES. 2002. Population structure of the tsetse fly *Gossina pallidipes* estimated by allozyme, microsatellite and mitochondrial gene diversities. *Insect Mol. Biol.* 11:37–45
48. Laikre L, Jarvi T, Johansson L, Palm S, Rubin JF, et al. 2002. Spatial and temporal population structure of sea trout at the Island of Gotland, Sweden, delineated from mitochondrial DNA. *J. Fish. Biol.* 60:49–71
49. Lange K. 1995. Applications of the Dirichlet distribution to forensic match probabilities. *Genetica* 96:107–17
50. Li Y-J. 1996. *Characterizing the structure of genetics populations*. PhD thesis, North Carolina State Univ. 106 pp.
51. Long J. 1986. The allelic correlation structure of Gainj- and Kalam-speaking people I. The estimation and interpretation of Wright's F-statistics. *Genetics* 112:629–47
52. Lowe AL, Urquhart A, Foreman LA, Evett IW. 2001. Inferring ethnic origin by means of an STR profile. *Forensic Sci. Int.* 119:17–22
53. Manzano C, de la Rua C, Iriondo M, Mazn LI, Vicario A, Aguirre A. 2002. Structuring the genetic heterogeneity of the Basque population: a view from classical polymorphisms. *Hum. Biol.* 74:51–74
54. Margis R, Felix D, Caldas JF, Salgueiro F, de Araujo DSD, et al. 2002. Genetic differentiation among three neighboring Brazil-cherry (*Eugenia uniflora L.*) populations within the Brazilian Atlantic rain forest. *Biodivers. Conserv.* 11:149–63
55. Marshall HD, Ritland K. 2002. Genetic diversity and differentiation of Kermode bear populations. *Mol. Ecol.* 11:685–97
56. Merilä J, Crnokrak P. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. *J. Evol. Biol.* 14:892–903
57. Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142:1061–64
58. Monaghan MT, Spaak P, Robinson CT, Ward JV. 2002. Population genetic structure of 3 alpine stream insects: influences of gene flow, demographics, and habitat fragmentation. *J. N. Am. Benthol. Soc.* 21:114–31
59. Morand ME, Brachet S, Rossignol P, Dufour J, Frascaria-Lacoste N. 2002. A generalized heterozygote deficiency assessed with microsatellites in French common ash populations. *Mol. Ecol.* 11:377–85
60. Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P.

2002. Assessing population differentiation and isolation from single nucleotide polymorphism data. *Proc. R. Stat. Soc.* In press
61. Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–96
62. Raufaste N, Bonhomme F. 2000. Properties of bias and variance of two multiallelic estimators of F_{ST}. *Theoret. Pop. Biol.* 57:285–96
63. Raymond M, Rousset F. 1995. An exact test for population differentiation. *Evolution* 49:1280–83
64. Reusch TBH. 2002. Microsatellites reveal high population connectivity in eelgrass (*Zostera marina*) in two contrasting coastal areas. *Limnol. Oceanogr.* 47:78–85
65. Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105:767–79
66. Ribeiro MM, LeProvost G, Gerber S. 2002. Origin identification of maritime pine stands in France using chloroplast simple-sequence repeats. *Ann. For. Sci.* 59:53–62
67. Rios C, Sanz S, Saavedra C, Pea JB. 2002. Allozyme variation in populations of scallops, *Pecten jacobaeus* (L.) and *P. maximus* (L.) (Bivalvia: Pectinidae), across the Almeria-Oran front. *J. Exp. Mar. Biol. Ecol.* 267:223–44
68. Richardson BA, Brunsfeld J, Klopfenstein NB. 2002. DNA from bird-dispersed seed and wind-disseminated pollen provides insights into postglacial colonization and population genetic structure of whitebark pine (*Pinus albicaulis*). *Mol. Ecol.* 11:215–27
69. Robertson AR. 1962. Weighting in the estimation of variance components in the unbalanced single classification. *Biometrics* 18:413–17
70. Robertson A, Hill WG. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* 107:703–18
71. Roeder K, Escobar M, Kadane J, Balasz I. 1998. Measuring heterogeneity in forensic databases using hiererachical Bayes models. *Biometika* 85:269–87
72. Roques S, Sevigny JM, Bernatchez L. 2002. Genetic structure of deep-water redfish, *Sebastes mentella*, populations across the North Atlantic. *Mar. Biol.* 140:297–307
73. Rothman ED, Sing CF, Templeton AR. 1974. A model for analysis of population structure. *Genetics* 76:943–60
74. Rousset F. 2001. Inferences from spatial population genetics. See Ref. 2, pp. 239–69
75. Rousset F. 2002. Inbreeding and relatedness coefficients: What do they measure? *Heredity* 88:371–80
76. Schultheis AS, Hendricks AC, Weigt LA. 2002. Genetic evidence for 'leaky' cohorts in the semivoltine stonefly *Peltoperla tarteri* (Plecoptera: Peltoperlidae). *Freshwater Biol.* 47:367–76
77. Shriver MD, Smith MW, Li Jin. 1998. Reply to Brenner. *Am. J. Hum. Genet.* 62:1560–61
78. Shriver MD, Smith MW, Li Jin, Marcini A, Akey JM, et al. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* 60: 957–64
79. Slatkin M. 1995. A measure of population subdivision based on microsatellite frequencies. *Genetics* 139:457–62
80. Smouse PE, Spielman RS. 1977. How allocation of individuals depends on genetic differences among populations. In *Human Genetics*, ed. S Armendares, R. Lisker, pp. 255–60. Amsterdam: Excerpta Medica
81. Smouse PE, Williams RC. 1982. Multivariate analysis of HLA-disease associations. *Biometrics* 38:757–68
82. Spielman RS, Smouse PE. 1975. Multivariate classification of human populations. I. Allocation of Yanomama Indians to villages. *Am. J. Hum. Genet.* 28:317–31

83. Spitze K. 1993. Population structure in *Daphnia abtusa*: quantitative genetic and allozymic variation. *Genetics* 135:367–74
84. Squire T, Newman RA. 2002. Fine-scale population structure in the wood frog (*Rana sylvatica*) in a northern woodland. *Herpetologica* 58:119–30
85. Travis SE, Proffitt CE, Lowenfield RC, Mitchell TW. 2002. A comparative assessment of genetic diversity among differently-aged populations of *Spartina alterniflora* on restored versus natural wetlands. *Restor. Ecol.* 10:37–42
86. Tsigenopoulos CS, Kotlik P, Berrebi P. 2002. Biogeography and pattern of gene flow among Barbus species (*Teleostei: Cyprinidae*) inhabiting the Italian Peninsula and neighbouring Adriatic drainages as revealed by allozyme and mitochondrial sequence data. *Biol. J. Linn. Soc.* 75:83–99
87. Tufto J, Engen S, Hindar K. 1996. Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* 144:1911–21
88. Van der Strate HJ, Van de Zande L, Stam WT. 2002. The contribution of haploids, diploids and clones to fine-scale population structure in the seaweed *Cladophoropsis membranacea* (Chlorophyta). *Mol. Ecol.* 11:329–45
89. Viard F, Bernard J, Despalnque B. 2002. Crop-weed interactions in the *Beta vulgaris* complex at a local scale: allelic diversity and gene flow within sugar beet fields. *Theor. Appl. Genet.* 104:688–97
90. Vitalis R, Couvet D. 2001. Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* 157:911–25
91. Vitalis R, Dawson K, Boursot P. 2001. Interpretation of variation across marker loci as evidence of selection. *Genetics* 158:1811–23
92. von der Schulenburg JHJ, Hurst GDD, Tetzlaff D, Booth GE, Zakharov IA, Majerus MEN. 2002. History of infection with different male-killing bacteria in the two-spot ladybird beetle *Atalia bipunctata* revealed through mitochondrial DNA sequence analysis. *Genetics* 160:1075–86
93. Wakeley J. 2001. The coalescent in an island model of population subdivision with variation among demes. *Theor. Pop. Biol.* 59:133–44
94. Was A, Wenne R. 2002. Genetic differentiation in hatchery and wild sea trout (*Salmo trutta*) in the Southern Baltic at microsatellite loci. *Aquaculture* 204:493–506
95. Weicker JJ, Brumfield RT, Winker K. 2001. Estimating the unbiased estimator θ for population genetic survey data. *Evolution* 55:2601–5
96. Weir BS. 1994. The effects of inbreeding on forensic calculations. *Annu. Rev. Genet.* 28:597–621
97. Weir BS. 1996. *Genetic Data Analysis II*. Sunderland, MA: Sinauer 376 pp.
98. Weir BS. 2000. What is the structure of human populations? *Evol. Biol.* 32:195–202
99. Weir BS. 2001. Forensics. See Ref. 2, pp. 721–39
100. Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–70
101. Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity* 82:1385–70
102. Wolańska-Nowak P. 2000. Application of subpopulation theory to evaluation of DNA evidence. *Forensic Sci. Int.* 113:63–69
103. Wondji C, Simard F, Fontenille D. Evidence for genetic differentiation between the molecular forms M and S within the Forest chromosomal form of *Anopheles gambiae* in an area of sympatry. *Insect Mol. Biol.* 11:11–19
104. Wright S. 1951. The genetical structure of populations. *Ann. Eugen.* 15:323–54
105. Yang R-C. 1998. Estimating hierarchical *F*-statistics. *Evolution* 52:950–56

Annual Review of Genetics
Volume 36, 2002

CONTENTS