DOI: 10.1534/genetics.107.077263

Empirical Bayes Inference of Pairwise F_{ST} and Its Distribution in the Genome

Shuichi Kitada,*,[1] Toshihide Kitakado* and Hirohisa Kishino†

**Faculty of Marine Science, Tokyo University of Marine Science and Technology, Minato, Tokyo 108-8477, Japan and*
†*Graduate School of Agriculture and Life Sciences, University of Tokyo, Bunkyo, Tokyo 113-8657, Japan*

Manuscript received June 11, 2007
Accepted for publication July 17, 2007

ABSTRACT

Populations often have very complex hierarchical structure. Therefore, it is crucial in genetic monitoring and conservation biology to have a reliable estimate of the pattern of population subdivision. F_{ST}'s for pairs of sampled localities or subpopulations are crucial statistics for the exploratory analysis of population structures, such as cluster analysis and multidimensional scaling. However, the estimation of F_{ST} is not precise enough to reliably estimate the population structure and the extent of heterogeneity. This article proposes an empirical Bayes procedure to estimate locus-specific pairwise F_{ST}'s. The posterior mean of the pairwise F_{ST} can be interpreted as a shrinkage estimator, which reduces the variance of conventional estimators largely at the expense of a small bias. The global F_{ST} of a population generally varies among loci in the genome. Our maximum-likelihood estimates of global F_{ST}'s can be used as sufficient statistics to estimate the distribution of F_{ST} in the genome. We demonstrate the efficacy and robustness of our model by simulation and by an analysis of the microsatellite allele frequencies of the Pacific herring. The heterogeneity of the global F_{ST} in the genome is discussed on the basis of the estimated distribution of the global F_{ST} for the herring and examples of human single nucleotide polymorphisms (SNPs).

INFERRING genetic population structure has been a major theme in population biology, ecology, and human genetics. The fixation index F_{ST}, introduced by Wright (1951), is a key parameter for such studies and is most commonly used to measure genetic divergence among subpopulations (Palsbøll *et al.* 2007). It is defined as the correlation between random gametes drawn from the same subpopulation relative to the total population. Another measure used frequently is Cockerham's (1969, 1973) coancestry coefficient, which is the probability that two random genes from different individuals are identical by descent, and the average overall pairs of individuals within the same subpopulation equal Wright's F_{ST} (Excoffier 2003). We use the notation θ_{WC} for the average coancestry coefficient and $\theta_{WC} = F_{ST}$ as shown by Weir and Cockerham (1984). Nei's (1973) G_{ST} is analogous to F_{ST} and identical to F_{ST} for diploid random-mating populations (Excoffier 2003).

Nei and Chesser (1983) proposed an estimator for F_{ST} and G_{ST}. The estimation of these parameters accounts only for the sampling error within subpopulations and therefore assumes that all subpopulations have been sampled (Cockerham and Weir 1986; Excoffier 2003). Weir and Cockerham (1984) developed the moment estimator $\hat{\theta}_{WC}$ for the coancestry coefficient θ_{WC}, which takes the sampling error for the subpopulations into account. Several moment estimators with different weighting schemes have also been derived (Robertson and Hill 1984; Weir and Cockerham 1984). An alternative estimation has been discussed using the method of ordinary least squares (Reynolds *et al.* 1983). Weir and Hill (2002) extended θ_{WC} to a population-specific parameter to allow different levels of coancestry for different populations. They also derived an estimator for θ_{WC} with confidence intervals using a normal theory approach.

Despite the development of methods for assigning individuals to populations (Paetkau *et al.* 1995; Pritchard *et al.* 2000; Huelsenbeck and Andolfatto 2007), the differentiation estimators remain the most commonly used tools for describing population structure (Balloux and Lugon-Moulin 2002). Weir and Cockerham (1984) showed that their estimator $\hat{\theta}_{WC}$ provides the smallest bias among the moment estimators. Goudet *et al.* (1996) confirmed this using simulations and showed that $\hat{\theta}_{WC}$ generates the least-biased estimate of F_{ST} but has the largest variance when F_{ST} is small. Raufaste and Bonhomme (2000) showed that $\hat{\theta}_{WC}$ is nearly unbiased, with minimal variance for large F_{ST}, and that the estimator of Robertson and Hill (1984) $\hat{\theta}_{RH}$ is negatively biased, with minimal variance for small F_{ST}. They proposed a correction for the bias of $\hat{\theta}_{RH}$, but this cannot be corrected properly in the range of [0.05, 0.1]. Therefore, a precise estimate of F_{ST} is crucial, especially for small and moderate levels of genetic differentiation.

[1] *Corresponding author:* Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato, Tokyo, 108-8477, Japan.
E-mail: kitada@kaiyodai.ac.jp

In addition to the estimation of F_{ST} over all subpopulations in a metapopulation (hereafter, we call this global F_{ST}), F_{ST}'s for pairs of sampled localities or subpopulations (pairwise F_{ST}) are usually estimated in conservation biology and ecology. In fact, the computer programs Arlequin (Excoffier *et al.* 2005), FSTAT (Goudet 1995), and Genepop (Raymond and Rousset 1995) estimate these parameters and are used widely in ecological studies. These three software packages produce the same or similar values for pairwise F_{ST} estimates and provide the basic statistics for exploratory analyses of population structure, such as cluster analysis and multidimensional scaling. They are also used as a criterion for population differentiation (Waples and Gaggiotti 2006; Palsbøll *et al.* 2007). However, the estimation of F_{ST}'s is not precise enough to reliably estimate the population structure and the extent of heterogeneity, especially for large gene flow species.

Small numbers of individuals taken from each locality should also affect the precision of $\widehat{F_{ST}}$. Populations often have very complex hierarchical structures, and geographical samples are usually taken from many localities to include a wide area. Therefore, the numbers of individuals from each locality are frequently limited by the large number of sampling points. Small sample sizes can result in biased estimates of the allele frequencies of each subpopulation. This bias may be larger for cases with larger numbers of alleles, such as microsatellite DNA. Uncertainty in the estimates of allele frequencies should affect the estimation of F_{ST}'s. The Bayesian approach provides better estimates of allele frequencies by taking uncertainty into account (Lange 1995; Lockwood *et al.* 2001). Posterior distributions of global F_{ST} were simulated from posterior distributions of allele frequencies, assuming common hyperparameters across all loci (Holsinger 1999; Holsinger *et al.* 2002; Corander *et al.* 2003). However, accurate estimation of pairwise F_{ST}, the essential parameter in ecological studies, has not been fully investigated.

In this article, we propose an empirical Bayes procedure to estimate locus-specific pairwise F_{ST}'s, taking into account the uncertainty of the allele frequencies of subpopulations. The estimation procedure has two stages. First, the hyperparameters of Dirichlet prior distributions for allele frequencies at each locus are estimated from observed allele counts by a maximum-likelihood method. The global F_{ST} is then estimated at each locus. Second, on the basis of the estimates of the hyperparameters, and given the allele counts, posterior distributions of the allele frequencies are generated for each locus, from which the posterior distributions of locus-specific pairwise F_{ST}'s are simulated. The posterior mean of our empirical Bayes pairwise F_{ST} estimates can be interpreted as a shrinkage estimator (Stein 1956; Maritz and Lewin 1989) anchored to the average of the true values among pairs. It performs better than conventional differentiation estimators and robustly estimates the population structure, even for non-Dirichlet cases, as stepping-stone models. The posterior distribution of pairwise F_{ST}'s can be used to calculate a criterion of population differentiation. Our maximum-likelihood estimates of the global F_{ST}'s can also be used as sufficient statistics to estimate the distribution of F_{ST} among loci in the genome. Our model assumes random mating or random sampling of alleles at each locality and that linkage equilibrium holds between loci. It also assumes that allele counts at each locus, given the true allele frequencies, are independent among populations. Our method can be applied to frequency data for common genetic markers, including isozymes, mitochondrial DNA, microsatellites, and single nucleotide polymorphisms (SNPs). We show the efficacy of our model by simulation and by an analysis of microsatellite allele frequencies of the Pacific herring. The heterogeneity of F_{ST} in the genome is discussed on the basis of the estimated distribution of global F_{ST} for the herring and examples of human SNPs.

MODELS AND METHODS

The model: Consider a simple random sampling from multiple localities in a metapopulation. Suppose that K random-mating demes or subpopulations are drawn from the metapopulation. Let $p_{kl} = (p_{kl1}, \ldots, p_{klJ_l})'$ ($k = 1, \ldots, K$; $l = 1, \ldots, L$) be a vector of the true allele frequencies at locus l in subpopulation k, where J_l is the number of different alleles at the locus, and $\sum_{j=1}^{J_l} p_{klj} = 1$. We assume a Dirichlet distribution as the prior distribution of p_{kl}. The probability density function is

$$\pi(p_{kl} \mid \alpha_{lj}) = \frac{\Gamma(\theta_l)}{\prod_{j=1}^{J_l} \Gamma(\alpha_{lj})} \prod_{j=1}^{J_l} p_{klj}^{\alpha_{lj}-1},$$

where $\alpha_l = (\alpha_{l1}, \ldots, \alpha_{lJ_l})'$ are the hyperparameters and $\theta_l = \sum_{j=1}^{J_l} \alpha_{lj}$ is a scale parameter that is specific for the locus. This model describes well a metapopulation that has a continuous structure and consists of an infinite number of subpopulations or demes (Pannell and Charlesworth 2000; Rousset 2003; Hanski and Gaggiotti 2004). Let $\beta_l = (\beta_{l1}, \ldots, \beta_{lJ_l})'$ be the mean allele frequency for the metapopulation at the locus satisfying $\sum_{j=1}^{J_l} \beta_{lj} = 1$. Hence, we have the relation $\beta_{lj} = \alpha_{lj}/\theta_l$.

Under this model, the global F_{ST} (hereafter denoted as F_{ST}^{G}) at each locus is expressed simply by the scale parameter, as

$$F_{ST,l}^{G} = \frac{1}{1 + \theta_l}, \tag{1}$$

as given by Wright (1969), Rannala and Hartigan (1996), Balding and Nichols (1997), Lockwood *et al.* (2001), Balding (2003), and Kitada and Kishino (2004).

In this model, the variance of the jth allele frequency for the locus, p_{klj}, is expressed by

$$V[p_{klj}] = \frac{1}{1+\theta_l}\,\beta_{lj}(1-\beta_{lj}),$$

as given by Weir (1996), Holsinger *et al.* (2002), Balding (2003), and Kitakado *et al.* (2006). The Dirichlet distribution assumes an evolutionary equilibrium and an equal mutation rate for all alleles (Weir and Hill 2002; Ewens 2004). Under this assumption, the scale parameter θ_l refers to the rate of gene flow, as given by Rannala and Hartigan (1996). We use the symbol θ for the scale parameter, following Rannala and Hartigan (1996) and Balding and Nichols (1997). Weir and Cockerham (1984) also used the same symbol θ for the coancestry coefficient ($= F_{\mathrm{ST}}$), so we use θ_{WC} for their θ. Our $F^{\mathrm{G}}_{\mathrm{ST}}$ is equivalent to θ_{WC} (Weir 1996, pp. 47–48) and Holsinger's θ^{B} (Holsinger 1999; Holsinger *et al.* 2002).

Maximum-likelihood estimation of hyperparameters and global F_{ST}: The maximum-likelihood estimation of the hyperparameters has been discussed by Lange (1995), Kitada *et al.* (2000), and Balding (2003). A pseudo-likelihood approach was also taken by Rannala and Hartigan (1996). In the maximum-likelihood framework, a method for the simultaneous estimation of $F^{\mathrm{G}}_{\mathrm{ST}}$ and the linkage disequilibrium coefficient between two SNPs has been proposed (Kitada and Kishino 2004). Kitakado *et al.* (2006) proposed an integrated-likelihood approach to reduce the negative bias of $F^{\mathrm{G}}_{\mathrm{ST}}$, particularly for cases with few sampling points.

Suppose that N_k $(k = 1, \ldots, K)$ alleles of diploid organisms ($N_k/2$ individuals) are counted at locus l and $n_{kl} = (n_{kl1}, \ldots, n_{klJ_l})'$ denotes a vector of observed allele counts at the locus in subpopulation k. We assume that all individuals are successfully genotyped at all loci, so $N_k = N_{kl} = \sum_{j=1}^{J_l} n_{klj}$. The marginal likelihood of the observed allele counts at a locus n_{kl} has a Dirichlet-multinomial distribution (Lange 1995; Rannala and Hartigan 1996; Weir 1996; Balding and Nichols 1997; Kitada *et al.* 2000; Balding 2003; Rousset 2003). The parameters to be estimated are $\alpha_l = (\alpha_{l1}, \ldots, \alpha_{lJ_l})'$. Because we assume the independence of subpopulations, the overall likelihood for these parameters is given by the product of the likelihood functions for K samples, as

$$L(\alpha_l \mid n_{kl}) = \prod_{k=1}^{K}\left\{\frac{N_k!}{\prod_{j=1}^{J_l} n_{klj}!}\,\frac{\Gamma(\theta_l)}{\Gamma(N_k+\theta_l)}\prod_{j=1}^{J_l}\frac{\Gamma(n_{klj}+\alpha_{lj})}{\Gamma(\alpha_{lj})}\right\}. \tag{2}$$

The hyperparameters α_l are estimated by maximizing this marginal likelihood (Lange 1995; Kitada *et al.* 2000). Our method can be used for both allele and haplotype counts without modification, but some notations differ slightly. For haploid organisms, N_k refers to the individuals genotyped; and n_{kl} should be n_k and α_{lj} should be α_j. Henceforth, for simplicity, we focus on diploid organisms.

The locus-specific $F^{\mathrm{G}}_{\mathrm{ST},l}$ can be estimated by substituting $\hat\theta_l\left(= \sum_{j=1}^{J_l}\hat\alpha_{lj}\right)$ for θ_l in Equation 1. The variance estimator for $\hat\theta_l$ is calculated from the Fisher information matrix for α_l, as $\hat V(\hat\theta_l) = \sum_{j=1}^{J_l}\hat V(\hat\alpha_{lj}) + 2\sum_{j<j'}^{J_l}\widehat{\mathrm{Cov}}(\hat\alpha_{lj},\hat\alpha_{lj'})$. The asymptotic variance for $\widehat{F^{\mathrm{G}}_{\mathrm{ST},l}}$ is estimated using the Delta method (Seber 1982) as

$$\hat V(\widehat{F^{\mathrm{G}}_{\mathrm{ST},l}}) \sim \frac{\hat V(\hat\theta_l)}{(1+\hat\theta_l)^4}. \tag{3}$$

In our metapopulation model or infinite-island model, the sampled localities are regarded as a sample from all possible demes or subpopulations, including those not sampled. Hence, Equation 2 estimates the locus-specific genetic differentiation under the random-effect model of population sampling (Weir 1996). The average estimate of $F^{\mathrm{G}}_{\mathrm{ST}}$ for all loci is calculated as an arithmetic mean across the loci.

Empirical Bayes estimation of pairwise F_{ST}: The posterior distribution of allele frequencies p_{kl} at locus l in subpopulation k is again a Dirichlet distribution, with parameters modified by the sample allele counts

$$f(p_{kl} \mid n_{kl}) = \frac{\Gamma(\theta_l+N_k)}{\prod_{j=1}^{J_l}\Gamma(\alpha_{lj}+n_{klj})}\prod_{j=1}^{J_l} p_{klj}^{\alpha_{lj}+n_{klj}-1} \tag{4}$$

(Lange 1995; Weir 1996). Given the estimates of the hyperparameters and the sampled allele counts, random numbers of p_{kl} can be generated through this posterior distribution. The posterior distributions for any parametric functions of p_{kl} can then be simulated by the empirical Bayes procedure (Kitada *et al.* 2000).

When population differentiation between or among specific subpopulations is of interest, the selected populations can be regarded as the entire set of populations. Hence, applying the fixed-effect model of population sampling (Weir 1996) is appropriate. Therefore, we use Nei's G_{ST} formula (Nei and Chesser 1983), which defines quantities with respect to fixed extant populations (Cockerham and Weir 1986), to estimate the posterior distributions of pairwise F_{ST}'s (hereafter denoted as $F^{\mathrm{P}}_{\mathrm{ST}}$), as did Holsinger (1999) and Corander *et al.* (2003) in estimating global F_{ST}. Nei's gene diversity analysis compares expected heterozygosities under Hardy–Weinberg equilibrium (HWE), and the G_{ST} estimator is expressed as a function of allele frequencies. Therefore the posterior distribution of $F^{\mathrm{P}}_{\mathrm{ST}}$ at each locus can easily be generated on the basis of the G_{ST} estimator, without using genotype frequencies. We set the number of each simulation to 10,000, so 10,000 $F^{\mathrm{P}}_{\mathrm{ST}}$'s are calculated at each locus from the 10,000 sets of allele frequencies p_{kl} between a set of two populations. From the posterior distribution of $F^{\mathrm{P}}_{\mathrm{ST}}$, the posterior mean and 95% credible interval are calculated. We use the posterior mean as the empirical Bayes estimator of locus-specific $F^{\mathrm{P}}_{\mathrm{ST}}$. We can also calculate the probability that $F^{\mathrm{P}}_{\mathrm{ST}}$ is smaller than an arbitrary value [*e.g.*, $P(F^{\mathrm{P}}_{\mathrm{ST}} \le c)$], which can be used as the criterion for population differentiation (Waples and Gaggiotti 2006; Palsbøll *et al.* 2007). The average estimate of $F_{\mathrm{ST}}^{\mathrm{P}}$ for overall loci is calculated as an arithmetic mean across the loci.

Rosenberg *et al.* (2003) proposed a general measure for determining the amount of information on individual ancestry on the basis of the Kullback–Leibler information. The informativeness for assignment I_n is defined as

$$I_{n,l} = \sum_{j=1}^{J_l} \left\{ -\bar{p}_{lj} \log \bar{p}_{lj} + \sum_{k=1}^{K} \frac{p_{klj}}{K} \log p_{klj} \right\}, \qquad (5)$$

where $\bar{p}_{lj} = \sum_{k=1}^{K} p_{klj}/K$. The authors showed that I_n and F_{ST} are very closely correlated but that I_n is more informative than the standard SNP-specific pairwise F_{ST}. In an additional analysis, we examine how our empirical Bayes method works to measure this under the same simulation protocol.

Inferring heterogeneity of global F_{ST} among loci: We estimate locus-specific $F_{\mathrm{ST}}^{\mathrm{P}}$'s on the basis of $F_{\mathrm{ST}}^{\mathrm{G}}$ estimated at each locus. Evolutionary forces may differ among sites in the genome. Therefore, it is important to investigate the heterogeneity of F_{ST} among loci. One practical analysis is to test the null hypothesis H_0, the homogeneity of $F_{\mathrm{ST}}^{\mathrm{G}}$ among L loci, $F_{\mathrm{ST},l}^{\mathrm{G}} = F_{\mathrm{ST}}^{\mathrm{G}}(l = 1, \ldots, L)$ against the alternative hypothesis H_1, the heterogeneity of $F_{\mathrm{ST}}^{\mathrm{G}}$ among loci, on the basis of estimates of $F_{\mathrm{ST},l}^{\mathrm{G}}$. When a large number of subpopulations are sampled, the maximum-likelihood estimate $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}$ follows a normal distribution of $N(F_{\mathrm{ST},l}^{\mathrm{G}},\ V(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}))$. The maximum likelihood under H_0 is then given as $\hat{L}_0 = \prod_{l=1}^{L}(1/\sqrt{2\pi\hat{\sigma}_l^2})e^{-(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}} - \bar{F}_{\mathrm{ST}}^{\mathrm{G}})^2/2}$, where $\bar{F}_{\mathrm{ST}}^{\mathrm{G}} = \sum_{l=1}^{L} w_l F_{\mathrm{ST},l}^{\mathrm{G}} / \sum_{l=1}^{L} w_l$, $w_l = (\hat{\sigma}_l^2)^{-1}$, and $\hat{\sigma}_l^2 = \hat{V}(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}})$. The maximum likelihood under H_1 is $\hat{L}_1 = \prod_{l=1}^{L}(1/\sqrt{2\pi\hat{\sigma}_l^2})$, maximizing $F_{\mathrm{ST},l}^{\mathrm{G}}$ by $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}$. The negative twice-log-likelihood ratio is then $\lambda = \sum_{l=1}^{L}(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}} - \bar{F}_{\mathrm{ST}}^{\mathrm{G}})^2/\hat{\sigma}_l^2$, which follows the χ^2-distribution with $(L - 1)$ d.f. under the null hypothesis. We can test the heterogeneity of $F_{\mathrm{ST}}^{\mathrm{G}}$ on the basis of the test statistics.

The other approach to investigate the heterogeneity of $F_{\mathrm{ST}}^{\mathrm{G}}$ is to estimate the distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$ in the genome. In recent years, the number of loci analyzed has been increasing in ecological studies, but is still smaller than those used for human SNPs. For such cases, it would be difficult to directly estimate the specific distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$ from the data. Here, we estimate the distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$ in the genome from estimates of randomly selected loci, $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}(l = 1, \ldots, L)$. When the distribution is expressed by the parametric model $f(F_{\mathrm{ST}}^{\mathrm{G}} \mid \rho)$, the unknown parameter ρ, which defines the distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$, is estimated by maximizing the log marginal likelihood

$$\ell\left(\rho \mid \widehat{F_{\mathrm{ST},1}^{\mathrm{G}}}, \ldots, \widehat{F_{\mathrm{ST},L}^{\mathrm{G}}}\right) = \sum_{l=1}^{L} \log \int p\left(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}} \mid F_{\mathrm{ST},l}^{\mathrm{G}}\right) f\left(F_{\mathrm{ST},l}^{\mathrm{G}} \mid \rho\right) dF_{\mathrm{ST},l}^{\mathrm{G}},$$

where $p(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}} \mid F_{\mathrm{ST},l}^{\mathrm{G}})$ is the distribution of $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}$. Here, we assume the independence of loci $(l = 1, \ldots, L)$. Because the maximum-likelihood estimator $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}$ is a sufficient statistic, it is possible to estimate the distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$ in the genome on the basis of the estimates $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}$ for randomly sampled loci, instead of using a direct estimation from the data.

For the preliminary discussion here, we assume that $F_{\mathrm{ST}}^{\mathrm{G}}$ is normally distributed in the genome. When the distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$ is expected to be different from 0, a simple approximation may be a normal distribution. We then assume $F_{\mathrm{ST}}^{\mathrm{G}}$ follows $N(\mu, \sigma^2)$ as a first step in estimating the distribution of $F_{\mathrm{ST}}^{\mathrm{G}}$ in the genome under the limited number of loci analyzed. In this case, the parameter vector ρ refers to μ and σ^2. The general form of the log marginal likelihood given above becomes

$$\begin{aligned} &\ell\left(\mu, \sigma^2 \mid \widehat{F_{\mathrm{ST},1}^{\mathrm{G}}}, \ldots, \widehat{F_{\mathrm{ST},L}^{\mathrm{G}}}\right) \\ &\quad = -\frac{1}{2}\sum_{l=1}^{L} \log(\sigma^2 + \sigma_l^2) - \frac{1}{2}\sum_{l=1}^{L} \frac{\left(\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}} - \mu\right)^2}{\sigma^2 + \sigma_l^2}. \end{aligned} \qquad (6)$$

Here, $\sigma_l^2(l = 1, \ldots, L)$ is the variance of the estimates, $\widehat{F_{\mathrm{ST},l}^{\mathrm{G}}}(l = 1, \ldots, L)$. We estimate μ and σ^2 numerically, regarding $\hat{\sigma}_l^2$ as σ_l^2. The distribution of $F_{\mathrm{ST}}^{\mathrm{P}}$'s can also be calculated with slight modifications to the above procedure.

RESULTS

Improved precision of our empirical Bayes estimator for pairwise F_{ST}: We investigated the performance of our method of estimating pairwise $F_{\mathrm{ST}}^{\mathrm{P}}$ using numerical simulations. Random vectors of allele frequencies at locus l in subpopulation k, p_{kl}'s, were generated independently from the Dirichlet distribution with the parameter $\alpha_l(= \theta_l\beta_l)$. Here, the number of sampling localities (K) was set at 5, 10, and 50, and the mean allele frequencies at a locus were assumed $\beta_l = (1, 2, \ldots, J_l)/(J_l(J_l + 1)/2)$ with $J_l = 50$. As the true values of the global $F_{\mathrm{ST},l}$'s $(= 1/(1 + \theta_l))$, we chose four different levels: $F_{\mathrm{ST},l}^{\mathrm{G}} = 0.01$, 0.05, 0.1, and 0.2. The sample size $(N_k/2)$ was deemed to be common to all the localities and was set at 20, 30, and 50 individuals. Then, allele counts n_{kl}'s were drawn independently from the multinomial distribution $\mathrm{Multi}(N_k, p_{kl})$ for K localities. The pairwise $F_{\mathrm{ST}}^{\mathrm{P}}$ values between the first and second localities were evaluated by the conventional Nei's G_{ST} estimator and the empirical Bayes method. In the latter procedure, 500 $F_{\mathrm{ST}}^{\mathrm{P}}$'s were simulated on the basis of Nei's G_{ST} formula to save computation time, and the posterior mean was calculated as the estimator of the pairwise F_{ST}. The point estimate for the conventional G_{ST} estimator and the posterior mean of the empirical Bayes estimates were compared with the true $F_{\mathrm{ST}}^{\mathrm{P}}$. These procedures were iterated $R = 1000$ times.

Figure 1 and supplemental Figure S1 at http://www.genetics.org/supplemental/ show the general features of the empirical Bayes estimator compared with those of the conventional G_{ST} estimator. The former examines

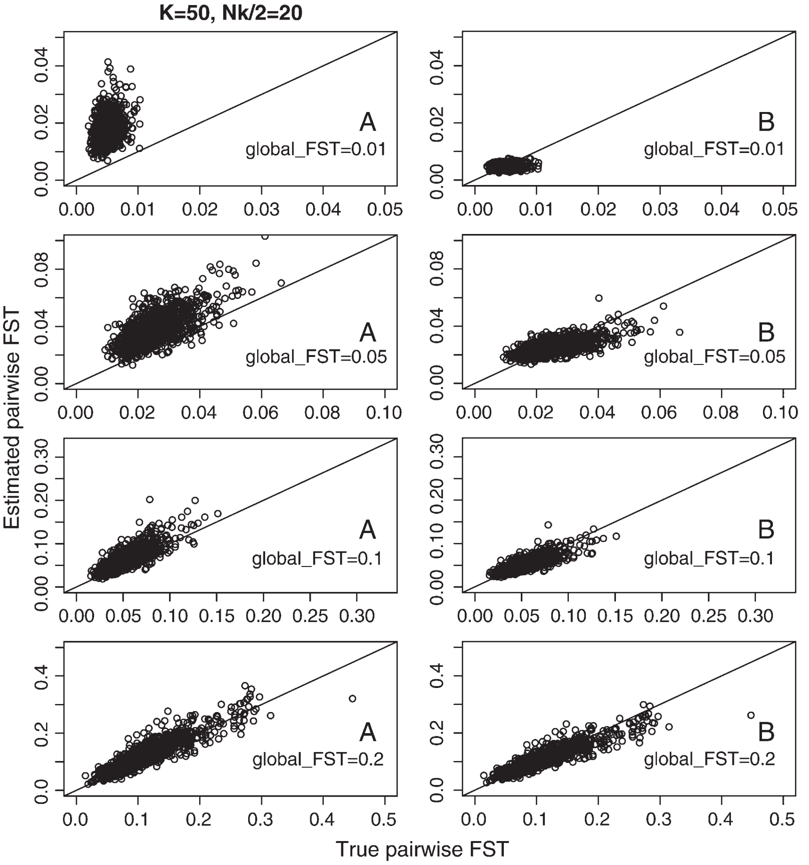

Figure 1.—The conventional G_{ST} (A) and empirical Bayes (B) estimates of pairwise $F_{\mathrm{ST}}^{\mathrm{P}}$ from 1000 simulations under the infinite-island model at various levels of $F_{\mathrm{ST}}^{\mathrm{G}}$ (0.01, 0.05, 0.1, 0.2) over subpopulations. The mean allele frequencies assumed were $\beta = (1, 2, \ldots, J)/(J(J+1)/2)$ with $J = 50$. The number of sampling localities (K) was set at 50. The sample size $N_k/2$ (individuals) was common to all localities and was set at 20 individuals.

the case of a relatively large number of sampling points and a limited sample size from each sampling point ($K = 50$, $N_k/2 = 20$). The latter investigates the case of moderately large samples from a small number of sampling points ($K = 10$, $N_k/2 = 50$), which is common in ecological studies. It is clearly apparent that the conventional G_{ST} estimator greatly overestimates the true values and has a large variance, especially for small $F_{\mathrm{ST}}^{\mathrm{G}}$. In contrast, the empirical Bayes estimates of $F_{\mathrm{ST}}^{\mathrm{P}}$'s shrink toward the average of the true $F_{\mathrm{ST}}^{\mathrm{P}}$ and reduce the positive bias of the conventional estimator. This is reasonable because our empirical Bayes estimator can be interpreted as a shrinkage estimator (Stein 1956; Maritz and Lewin 1989).

The effects of the number of sampling points K and sampled individuals $N_k/2$ are also shown in the supplemental material at http://www.genetics.org/supplemental/. The numbers of sampling points did not affect the positive bias of the conventional G_{ST} estimator because K was fixed at 2. The empirical Bayes procedure provided larger variation for smaller numbers of subpopulations, especially for small $F_{\mathrm{ST}}^{\mathrm{G}}$ values, and vice versa (supplemental Figure S2 at http://www.genetics.org/supplemental/). Conversely, larger numbers of sampled individuals reduced the positive bias of the conventional G_{ST} estimator and the variation of the empirical Bayes estimator (supplemental Figure S3 at http://www.genetics.org/supplemental/).

For a quantitative comparison of the two estimators, we used the following two measures: $(1/R)\sum_{r=1}^{R}(\widehat{F_{\mathrm{ST},r}^{\mathrm{P}}} - F_{\mathrm{ST},r}^{\mathrm{P}})/F_{\mathrm{ST},r}^{\mathrm{P}}$ (relative mean bias) and $\sqrt{(1/R)\sum_{r=1}^{R}((\widehat{F_{\mathrm{ST},r}^{\mathrm{P}}} - F_{\mathrm{ST},r}^{\mathrm{P}})/F_{\mathrm{ST},r}^{\mathrm{P}})^2}$ (root relative mean squared error), where $\widehat{F_{\mathrm{ST},r}^{\mathrm{P}}}$ and $F_{\mathrm{ST},r}^{\mathrm{P}}$ are an estimate and the true value, respectively, for pairwise $F_{\mathrm{ST}}^{\mathrm{P}}$ in the rth iteration. As shown in Table 1, smaller K and sample sizes ($N_k/2$) resulted in larger positive biases for the conventional estimator for various levels of true $F_{\mathrm{ST}}^{\mathrm{G}}$ values. The bias and variation became smaller as $F_{\mathrm{ST}}^{\mathrm{G}}$ became larger. In contrast, the empirical Bayes estimator provided smaller biases and variations for all cases of $F_{\mathrm{ST}}^{\mathrm{G}}$, although smaller K and sample sizes resulted in a slight negative bias.

The negative bias of the empirical Bayes estimator of pairwise F_{ST} is large, especially when gene flow is large and the estimation is based on a sample from a few sampling points. Underestimation of population differences should lead to optimistic management strategies. Therefore, it is recommended in the conservation genetics of birds and fish that samples be collected from as many sampling points as possible. It is noted that the bias of the empirical Bayes estimator is reduced by increasing the number of sampling points, whereas the bias of the conventional G_{ST} estimator is not.

We estimated Rosenberg *et al.*'s informativeness of assignment I_n between the first and second localities

TABLE 1

Relative mean bias and root relative mean squared error (in parentheses) of the conventional G_{ST} and empirical Bayes estimators of pairwise F_{ST} from 1000 simulations at various levels of global F_{ST} (see text)

F_{ST}^{G}	K	$N_k/2$	Conventional	Empirical Bayes
0.01	5	20	2.770 (3.019)	−0.726 (0.815)
		30	1.769 (1.942)	−0.541 (0.630)
		50	1.069 (1.192)	−0.330 (0.412)
	10	20	2.666 (2.880)	−0.325 (0.464)
		30	1.791 (1.955)	−0.205 (0.345)
		50	1.081 (1.214)	−0.122 (0.270)
	50	20	2.728 (2.972)	0.003 (0.294)
		30	1.784 (1.967)	0.010 (0.272)
		50	1.085 (1.217)	0.033 (0.251)
0.05	5	20	0.561 (0.685)	−0.136 (0.283)
		30	0.363 (0.478)	−0.097 (0.233)
		50	0.218 (0.304)	−0.056 (0.175)
	10	20	0.538 (0.662)	−0.039 (0.242)
		30	0.367 (0.475)	−0.015 (0.216)
		50	0.233 (0.317)	0.004 (0.169)
	50	20	0.557 (0.680)	0.047 (0.253)
		30	0.366 (0.480)	0.031 (0.226)
		50	0.206 (0.304)	0.015 (0.181)
0.1	5	20	0.273 (0.418)	−0.053 (0.242)
		30	0.176 (0.284)	−0.036 (0.189)
		50	0.106 (0.194)	−0.019 (0.149)
	10	20	0.255 (0.375)	−0.006 (0.215)
		30	0.183 (0.289)	0.009 (0.186)
		50	0.098 (0.195)	−0.001 (0.151)
	50	20	0.247 (0.373)	0.019 (0.215)
		30	0.182 (0.287)	0.027 (0.183)
		50	0.109 (0.192)	0.020 (0.147)
0.2	5	20	0.125 (0.250)	−0.017 (0.192)
		30	0.083 (0.195)	−0.007 (0.161)
		50	0.051 (0.134)	−0.003 (0.118)
	10	20	0.123 (0.246)	0.009 (0.192)
		30	0.092 (0.197)	0.018 (0.165)
		50	0.049 (0.138)	0.004 (0.122)
	50	20	0.127 (0.267)	0.027 (0.220)
		30	0.087 (0.188)	0.020 (0.155)
		50	0.058 (0.144)	0.018 (0.128)

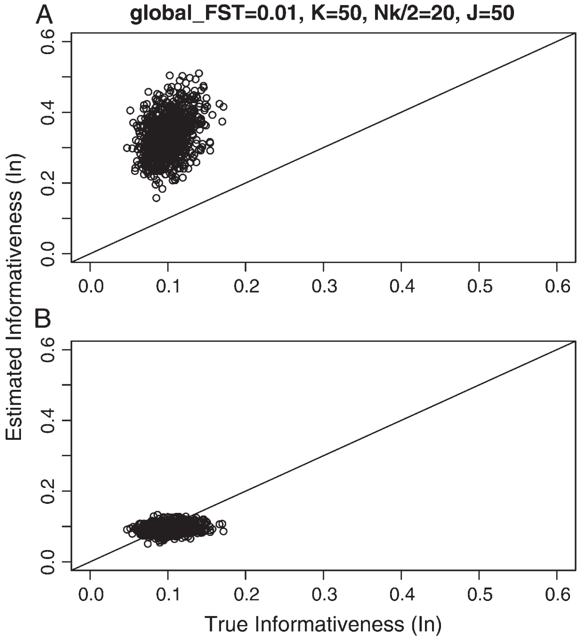

Figure 2.—The conventional informativeness of assignment I_n (Rosenberg *et al.* 2003) (A) and empirical Bayes (B) estimates of I_n from 1000 simulations for the case of $F_{ST}^{G} = 0.01$. The mean allele frequencies assumed were $\beta = (1, 2, \ldots, J)/(J(J+1)/2)$ with $J = 50$. The number of sampling localities (K) was set at 50 and the sample size $N_k/2$ (individuals) was common to all localities and was set at 20 individuals.

using Equation 5 ($K = 2$) and the empirical Bayes method based on the I_n estimator for the case of $F_{ST}^{G} = 0.01$. The mean allele frequencies were assumed to be $\beta = (1, 2, \ldots, J)/(J(J+1)/2)$ with $J = 50$. The number of sampling localities (K) was set at 50 and the sample size ($N_k/2$ individuals) was set at 20 individuals for each sampling point. Figure 2 shows that the conventional estimator for I_n was positively biased, whereas the empirical Bayes estimator of I_n performed much better, consistent with the fact that I_n produces upwardly biased estimates in small samples (Rosenberg *et al.* 2003).

Robustness of our shrinkage estimator for pairwise F_{ST}: The robustness of our F_{ST}^{P} estimator was explored using numerical simulations for non-Dirichlet distributions of the allele frequencies. We considered cases in which genetic differentiation becomes larger with geographic distance. Such a stepping-stone model is biologically realistic and may be common (Palsbøll *et al.* 2007). We set the number of sampled subpopulations (K) to 15 and the F_{ST}^{P} between two adjacent populations to 0.001 (case 1) or 0.0005 (case 2). We considered biallelic cases and set the allele frequencies to (0.5, 0.5) at a locus for the middle population. We then calculated the allele frequencies p_{kl}'s at the locus for another 14 subpopulations by numerical optimization. The sample size ($N_k/2$) was deemed to be common to all localities and was set at 20 individuals. Then, allele counts n_k's were drawn independently from the multinomial distribution Multi(N_k, p_{kl}) for 15 localities. A total of 105 pairwise F_{ST} values between all sets of the two localities were evaluated by the conventional G_{ST} and the empirical Bayes estimator following the simulation protocol described above.

As shown in Figure 3 (case 1) and supplemental Figure S4 at http://www.genetics.org/supplemental/ (case 2, top left), our empirical Bayes procedure provided better estimates than those of the conventional G_{ST} estimator for smaller F_{ST}^{P}'s (~ <0.06 for case 1 and 0.04 for case 2). In conservation, management units should be defined among subpopulations with little genetic differentiation. Palsbøll *et al.* (2007) concluded that $F_{ST}^{P} = 0.0025$ could be used as the criterion for deciding the

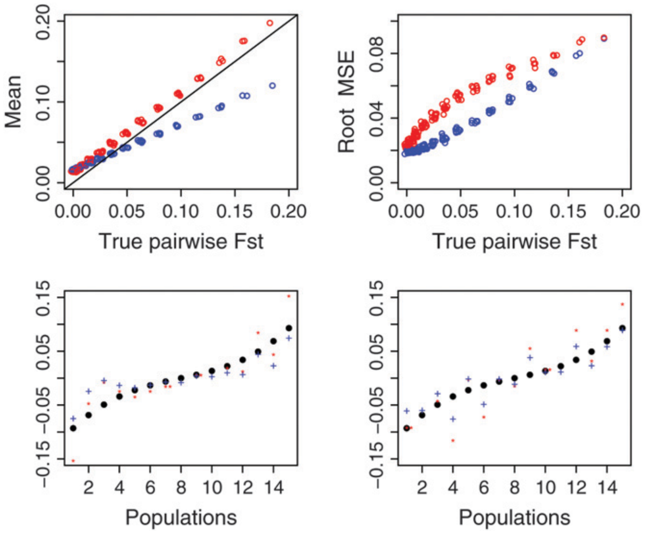

Figure 3.—Mean (top left) and root relative mean squared error (top right) of the conventional G_{ST} (red circle) and empirical Bayes estimators (blue circle) of pairwise F_{ST}^{P} from 1000 simulations under the stepping-stone models. Means and root MSEs were plotted on the true F_{ST}^{P}'s which fluctuated very slightly when small uniform random variables were added to prevent the points overlapping heavily. The number of subpopulations (K) was set at 15 and the pairwise F_{ST}^{P} between two adjacent populations was set at 0.001 (case 1) and 0.0005 (case 2). The sample size $N_k/2$ (individuals) was common to all localities and set at 20 individuals. Only the results for case 1 are shown. The results of the MDS analysis of two data sets are given in the bottom section; black circles show the true population structure, and the estimated population structure based on the conventional (red "*") and empirical Bayes estimates (blue "+") of F_{ST}^{P} is shown.

separate management units of sockeye salmon spawning sites. For such cases with very small genetic differentiation, our method performs more efficiently, even for stepping-stone models. On the contrary, the conventional G_{ST} estimator displayed a positive bias for the whole range of F_{ST}^{P}'s in both cases. Our method reduced the positive bias for small F_{ST}^{P}'s and the bias became negative for larger F_{ST}^{P}'s. Reflecting the characteristics of the shrinkage estimator, the relative mean bias of the empirical Bayes estimator, which was the average of the 105 pairwise F_{ST} estimates, was slightly (1.06 times) larger than that of the conventional estimator for case 1 (Table 2). The precision of our estimates was much better for the whole range of F_{ST}^{P} [Figure 3, top right (case 1) and supplemental Figure S4 (case 2)].

We also investigated the robustness of estimating population structure on the basis of estimated F_{ST}^{P}'s. The results of the multidimensional scaling (MDS) (Torgerson 1952; Young and Hamer 1987) analysis of two data sets in the simulation showed that our method describes the true population structure well. The conventional G_{ST} estimator also worked, despite the larger positive bias and variance (Figure 3, bottom).

The case of Pacific herring: We analyzed six geographical samples of the Pacific herring *Clupea pallasii* from spawning areas in Lake Akkeshi (AK), Yudonuma Lake (YD), and Funka Bay (FK), which are located off the east coast of Hokkaido in Japan, and Obuchinuma Lake (OB), Miyako Bay (MY), and Matsushima Bay (MT), located off the northern Pacific coast of Honshu. Hatchery fish, released and recaptured in Lake Akkeshi (AKH) and Miyako Bay (MYH), were also distinguished from wild fish on the basis of otoliths stained with alizarin complexon. A total of 2055 mature individuals were genotyped at five microsatellite loci. Allele frequencies are given in supplemental Tables S1–S5 at http://www.genetics.org/supplemental/. HWE was satisfied in each sample except at four localities for three loci, and the assumption of our metapopulation model was considered to be satisfied (supplemental Figure S5). On the basis of the estimates of the hyperparameters (supplemental Figure S6), the scale parameter θ_l and $F_{ST,l}^{G}$ were estimated over all subpopulations (Table 3).

We estimated the posterior distributions of F_{ST}^{P} for all sets of subpopulations at all loci. As an example, the posterior distributions between FK and MY at the five loci are shown in Figure 4. The posterior means of F_{ST}^{P} varied from 0.0064 to 0.0245, with the 95% credible intervals in parentheses (Table 4). The P-values in Figures 4 and 5 are for the homogeneity contingency test performed with Genepop (Raymond and Rousset 1995). At all loci, the allelic differences between FK and MY were highly significant ($P < 0.0000$). We used F_{ST}^{P} of 0.01 as a population criterion and defined the probability of

TABLE 2

Relative mean bias and root relative mean squared error of the conventional G_{ST} and empirical Bayes estimators of pairwise F_{ST} from 1000 simulations under the stepping-stone models (see text)

	Relative mean bias		Root relative MSE	
F_{ST}^{P}	Conventional	Empirical Bayes	Conventional	Empirical Bayes
0.001 (case 1)	1.165	1.239	2.352	1.817
0.0005 (case 2)	2.296	1.718	4.362	2.527

The number of subpopulations (K) was set at 15 and the pairwise F_{ST} between two adjacent populations was set at 0.001 (case 1) and 0.0005 (case 2). The sample size $N_k/2$ (individuals) was common to the localities and set at 20 individuals.

TABLE 3

Estimated locus-specific scale parameters θ and F_{ST}^{G} over subpopulations of the Pacific herring, with standard errors in parentheses

Locus	No. alleles	θ	F_{ST}^{G}
*Cha*17	48	63.98 (7.88)	0.0154 (0.0019)
*Cha*20	30	77.26 (13.13)	0.0128 (0.0021)
*Cha*63	35	58.21 (8.60)	0.0169 (0.0025)
*Cha*113	29	41.68 (6.59)	0.0234 (0.0036)
*Cha*123	50	58.78 (6.68)	0.0167 (0.0018)
Mean[a]	38.4	59.98 (3.98)	0.0170 (0.0011)
Mean[b]			0.0160 (0.0010)

[a] Simple mean over all loci.

[b] Weighted mean over all loci, which was equal to the MLE obtained by Equation 3.

TABLE 4

Posterior means and 95% credible and confidence intervals for pairwise F_{ST} of the Pacific herring between Funka Bay and Miyako Bay at each locus

Locus	d.f.	Posterior mean	Empirical Bayes	Weir and Hill[a]
*Cha*17	47	0.0193	[0.0137, 0.0258]	[0.0133, 0.0302]
*Cha*20	29	0.0100	[0.0063, 0.0146]	[0.0063, 0.0180]
*Cha*63	34	0.0064	[0.0046, 0.0086]	[0.0042, 0.0111]
*Cha*113	28	0.0245	[0.0162, 0.0349]	[0.0154, 0.0448]
*Cha*123	49	0.0213	[0.0158, 0.0279]	[0.0149, 0.0331]
Mean	1309	0.0163	[0.0138, 0.0191]	[0.0151, 0.0176]

[a] Estimated by Weir and Hill's (2002) normal theory approach.

$F_{ST}^{P} \leq 0.01$ as P^*. Here, $F_{ST}^{P} = 0.01$ refers to $N_e m \simeq 25$, which means that the effective number of migrants is 25 individuals per generation (Waples and Gaggiotti 2006), where N_e is the effective population size and m is the migration rate. The P^*-value was near 1.0 for *Cha* 63, indicating that the genetic differentiation at the locus was small. In contrast, P^*-values were 0 or near 0 for *Cha* 17, *Cha* 113, and *Cha* 123 and 0.5318 for *Cha* 20 (Figure 4). The posterior distribution of the average F_{ST}^{P} over all the loci was calculated as the mean of the posterior distributions at five loci (Figure 4, bottom right). Both the P- and P^*-values coincided at 0, showing significant genetic differentiation between Funka Bay and Miyako Bay.

The posterior distributions of average F_{ST}^{P} over all loci for all pairs of wild subpopulations were also simulated as simple means of the posterior distributions at five loci (Figure 5). The allelic differences were highly significant with $P = 0.0000$, except between MY and MT. The criterion P^* resulted in a very different evaluation of

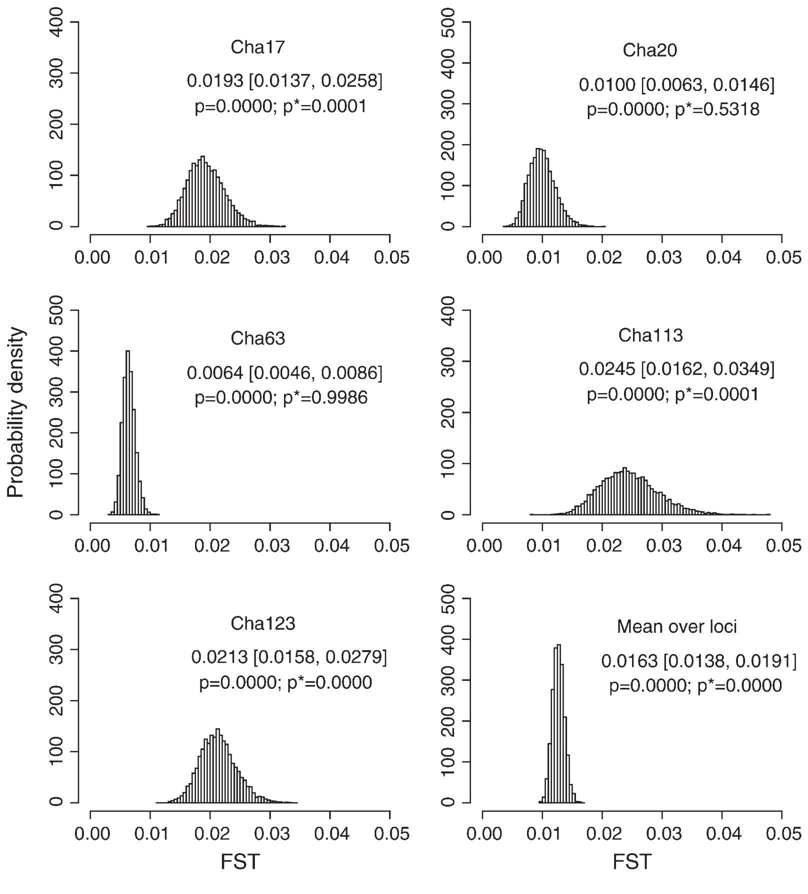

Figure 4.—Posterior distributions of F_{ST}^{P} for the Pacific herring between Funka Bay and Miyako Bay at each locus, and over all loci, which were averaged over F_{ST}^{P} at five loci.

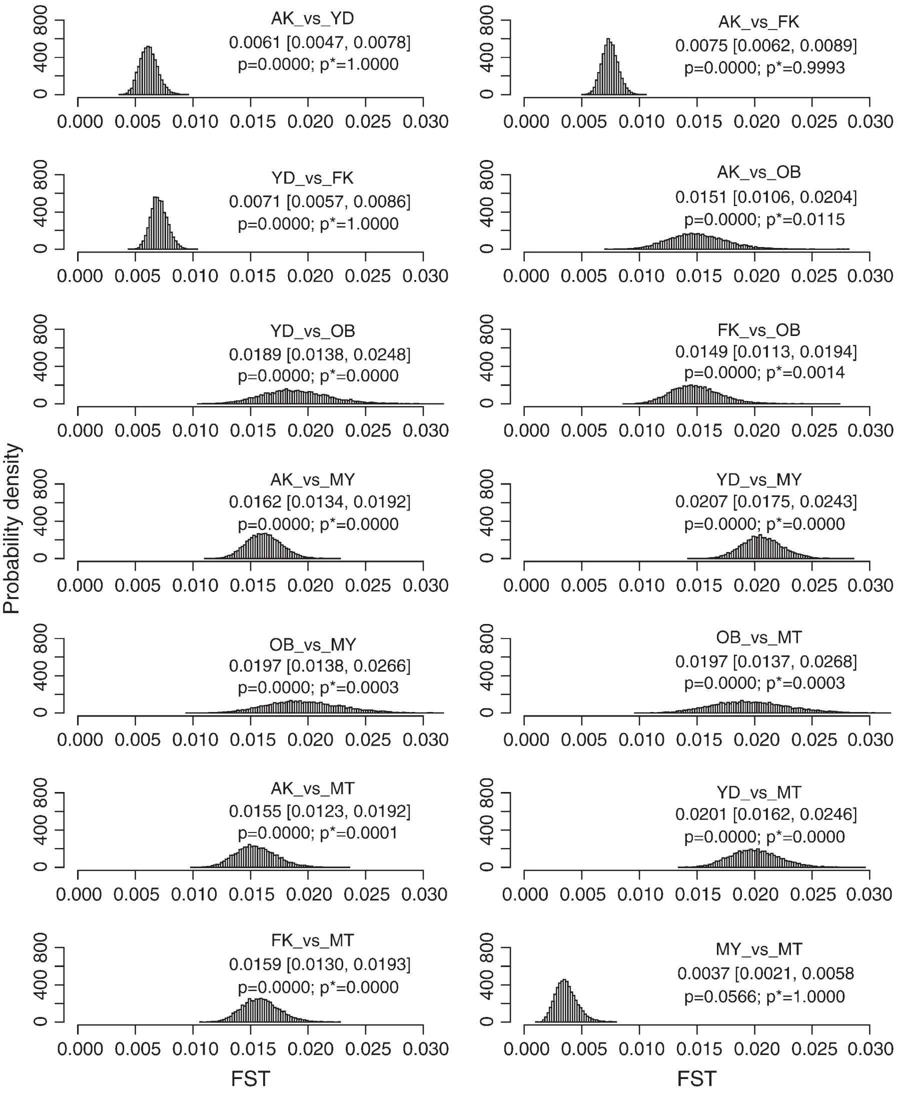

Figure 5.—Posterior distributions of F^{P}_{ST} for the Pacific herring over all loci, which were averaged over F^{P}_{ST} for five loci: AK, Lake Akkeshi; YD, Yudonuma Lake; FK, Funka Bay; OB, Obuchinuma Lake; MY, Miyako Bay; and MT, Matsushima Bay.

population differentiation, even for the same *P*-value of 0.0000. This result clearly shows the difficulty in the hypothesis-testing framework in evaluating the genetic differentiation between subpopulations (*e.g.*, Dizon *et al.* 1995; Ryman and Jorde 2001; Ryman *et al.* 2006). The *P**-based criterion works well in this case and is recommended for use in hypothesis testing.

The variation in $\widehat{F^{G}_{ST,l}}$ over five loci was not trivial, with a coefficient of variation (CV) of 23% (Table 3). However, the negative twice log-likelihood ratio λ was calculated to be 6.9264 and the hypothesis of constant F^{G}_{ST} for all loci was not rejected ($P = 0.8602$, d.f. = 4). We estimated the distribution of F^{G}_{ST} in the genome, assuming normality based on Equation 6. The maximum-likelihood estimates (MLEs) for μ and σ were obtained with 90, 95, and 99% confidence intervals, which define the distribution of F^{G}_{ST} in the genome via the likelihood profile (Figure 6A). The MLE for μ, with the 95% confidence interval, was 0.0160 (0.0128, 0.0206) [Figure 6A (a, e)] and for σ was 0 (0, 0.00716) [Figure 6A (0, c)]. The weighted mean $\bar{F}^{G}_{ST}$ coincided with the MLE of μ (Table 3) and the distribution of the weighted mean $N(\bar{F}^{G}_{ST}, \hat{V}(\bar{F}^{G}_{ST}))$, shown as the blue line in Figure 6B, described well the 95% confidence interval of μ [Figure 6A (a, e)]. Here, $\hat{V}(\bar{F}^{G}_{ST}) = 1/\left(\sum_{l=1}^{L} w_i\right)^2 \sum_{l=1}^{L} w_i^2 \hat{V}(\widehat{F^{G}_{ST,l}})$ and $V(\widehat{F^{G}_{ST,l}})$ were estimated with Equation 3. The log-likelihood profile for σ was monotonic and decreasing, indicating that the point estimate of σ was 0 (supplemental Figure S7 at http://www.genetics.org/supplemental/). Hence, the point estimate of the distribution of F^{G}_{ST} for the Pacific herring is constant, supporting the hypothesis of constant F^{G}_{ST} throughout the genome. However, the distribution of F^{G}_{ST} has uncertainty, which accounts for the confidence regions of μ and σ (Figure 6B).

DISCUSSION

Our empirical Bayes estimator of F^{P}_{ST} performed better than the conventional G_{ST} estimator for various

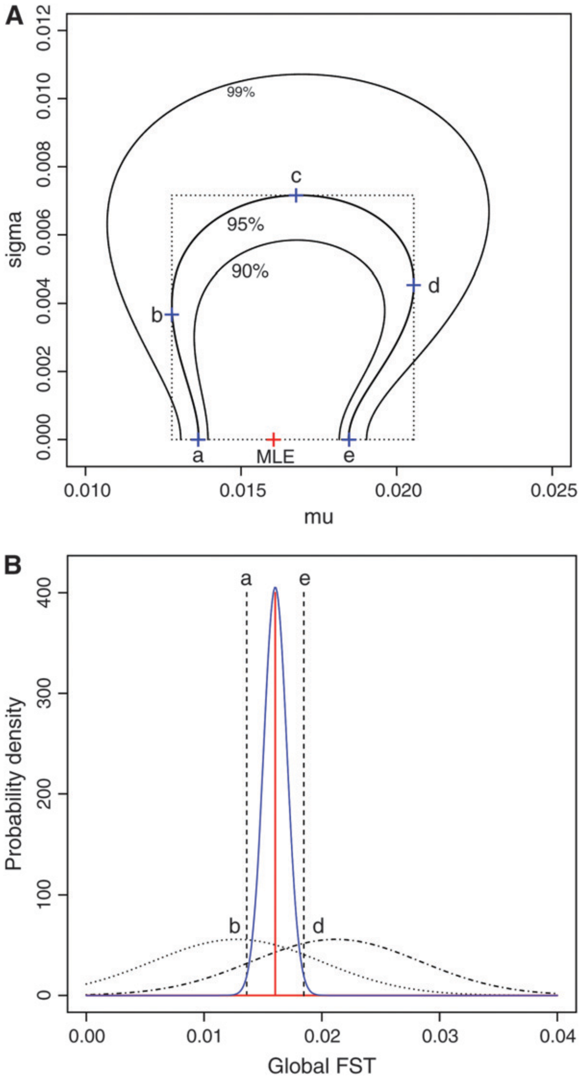

Figure 6.—Confidence regions of the distribution of F_{ST}^{G} in the genome of the Pacific herring, assuming a normal distribution (see the text). (A) The confidence regions of μ and σ^2, which specify the distribution. (B) The MLE distribution (red line, the delta distribution) and the representative distributions on the boundary of the confidence regions (a–e), which correspond to the points in A. The distribution of the weighted mean of $\widehat{F}_{ST}^{G}$ is superimposed (blue line).

levels of F_{ST}^{G} and various sampling conditions under the infinite-island model. Even for non-Dirichlet distributions of allele frequencies, such as stepping-stone models, our method provided better estimates of F_{ST}^{P} than did the conventional G_{ST}, especially for cases with large gene flow.

Integrated likelihood method: The empirical Bayes estimator of F_{ST}^{P}'s is negatively biased, especially when the population has large gene flow and the estimation is based on a sample from only a few sampling points (Table 1, supplemental Figure S2 at http://www.genetics.org/supplemental/). With a small number of sampling points, the MLE of θ is not precise. Therefore, it is recommended in the conservation genetic analysis of such a population that the samples be collected from as many sampling points as possible. When sampling from many localities is not feasible, the integrated-likelihood method (Kitakado *et al.* 2006) can reduce the negative bias of F_{ST}^{G}. We estimated pairwise F_{ST}^{P}'s by the empirical Bayes method based on integrated-likelihood estimates (ILEs) of θ for cases with $F_{ST}^{G} = 0.01$, $K = 5$, and $N_k/2 = 20$, 30, or 50 with the same simulation protocol used for Table 1 (supplemental Figure S8 at http://www.genetics.org/supplemental/). The relative mean biases of the F_{ST}^{P} estimates, with root relative mean squared errors in parentheses, were −0.238 (0.506), −0.132 (0.365), and −0.060 (0.272), respectively. These values were much smaller than those estimated on the basis of the MLE of θ, given in Table 1 (top three rows). The negative bias of $\widehat{F}_{ST}^{P}$'s based on the MLE was reduced to 32.8, 24.4, and 18.2% for $N_k/2 = 20$, 30, and 50, respectively. The integrated-likelihood method uses a uniform prior for the mean allele frequency β_l and eliminates β_l by integration regarding it as a nuisance parameter. By using the ILE of θ instead of the MLE, the empirical Bayes method proposed here provides nearly unbiased estimates of F_{ST}^{P}'s when the sample sizes ($N_k/2$ individuals) are large and works more efficiently, even for cases with a small number of sampling points (supplemental Figure S8).

Weir and Cockerham's $\hat{\theta}_{WC}$: When we estimate genetic differentiation between two specific subpopulations, selected subpopulations can be regarded as the entire set of populations. Nei's G_{ST} formula defines quantities with respect to fixed extant populations (Cockerham and Weir 1986). In addition, G_{ST} is a function of allele frequencies under HWE, and the posterior distribution can easily be simulated from only allele frequencies. Therefore, we used the G_{ST} estimator to estimate the posterior distribution of pairwise F_{ST}.

The citation record suggests that the most widely used estimator for Wright's F_{ST} is Weir and Cockerham's $\hat{\theta}_{WC}$ (Weir and Hill 2002). This moment estimator takes the sampling error for subpopulations into account and essentially estimates the global F_{ST}, F_{ST}^{G}. The estimator is also widely used to estimate pairwise F_{ST}'s among fixed pairs of populations. With the assumption of no local inbreeding, θ_{WC} is estimated only from sample allele frequencies, but these need to be inferred from sample genotype frequencies (Weir and Hill 2002). With J alleles at a locus, the number of possible genotypes is $J(J+1)/2$. In microsatellite DNA analyses, J is generally large. If $J = 50$, the number of genotypes is 1225. Such a situation makes our simulation more complicated and the uncertainty of the genotype counts becomes large under small or moderate sample sizes.

Here, we investigated the properties of $\hat{\theta}_{WC}$ on estimating pairwise F_{ST} on the basis of the relationship between G_{ST} and θ_{WC}. Weir and Cockerham's estimates $\hat{\theta}_{WC}$ can be approximated as a function of G_{ST} by Equation 2 in Weicker *et al.* (2001): $\theta_{WC} = \{G_{ST}(K + (K-1)/(N_k - 1)) - (K-1)/(N_k-1)\}/(G_{ST} + K - 1)$. Using this equation, we calculated the conventional estimates of pairwise Weir and Cockerham's θ_{WC} from G_{ST} estimates with the simulation protocol described in results. We examined cases with a global F_{ST} of $F_{ST}^{G} = 0.01$, 0.05, 0.1, and 0.2. The mean allele frequencies were assumed $\beta = (1, 2, \ldots, J)/(J(J+1)/2)$ with $J = 50$. The number of sampling localities K was set at 10 and the sample size $N_k/2$ (individuals) was deemed to be common to all the localities and was set at 50 individuals. As shown in supplemental Figure S9 at http://www.genetics.org/supplemental/, the two estimators have linear relationships. Hence, our simulation results on the conventional G_{ST} estimator can be extended straightforwardly to pairwise θ_{WC}. In fact, the pairwise $\hat{\theta}_{WC}$-values calculated for the Pacific herring with Genepop (Raymond and Rousset 1995) were 1.93 ± 0.47 times larger than the posterior means of F_{ST}^{P} (supplemental Figure S10 at http://www.genetics.org/supplemental/), which coincides with our simulations for small F_{ST}^{G}'s (Figure 1, supplemental Figures S1 and S11).

Weir and Cockerham's estimator $\hat{\theta}_{WC}$ is nearly unbiased (Raufaste and Bonhomme 2000), although it has a negative bias for the two-allele case (Weir and Hill 2002). Nevertheless, when estimating the pairwise F_{ST}, $\hat{\theta}_{WC}$ is considered to have a large positive bias, especially for species with large gene flows. Nei's G_{ST} and Rosenberg *et al.*'s informativeness of assignment I_n also showed the same phenomenon, suggesting the positive bias is irrelevant to the estimators. This positive bias of the conventional estimators was larger for smaller genetic differentiation. This might be caused by the large variation in the sample allele frequencies, which is larger for smaller sample sizes (individuals) and largely exceeded the real variation between subpopulations. The shrinkage estimator stabilizes such variation and provides better estimates.

Weir and Hill's normal theory: Weir and Hill's (2002) normal theory approach has the same variance as a Dirichlet distribution when $i \neq i'$ in their notation. Hence, their estimator $\hat{\theta}_N$ is equivalent to our $\widehat{F_{ST}^{G}}$. F_{ST}^{G} is the variance of the allele frequencies among subpopulations relative to the total population. Hence, $\widehat{F_{ST}^{G}}$ refers to a sample variance of allele frequencies, and therefore d.f. $\cdot\, \widehat{F_{ST}^{G}}/F_{ST}^{G}$ follows a χ^2-distribution when the number of sampling points K is sufficiently large, as shown by Weir and Hill (2002). The shape of the posterior distribution of F_{ST}^{P} at each locus was unimodal and slightly right tailed, which reminded us of χ^2-distributions (Figure 4). We estimated Weir and Hill's confidence intervals by substituting the posterior mean of F_{ST}^{P} with $\hat{\theta}_N$ as [d.f. $\widehat{F_{ST}^{P}}/\chi^2_{(d,0.975)}$, d.f. $\widehat{F_{ST}^{P}}/\chi^2_{(d,0.025)}$], where d.f. $= (K-1)(J-1)$ is the degrees of freedom. The confidence intervals of F_{ST}^{P} obtained with the χ^2-approximation coincided well with the credible intervals calculated from the posterior distributions, although our credible intervals were narrower than the confidence intervals of the χ^2-approximation, which were slightly right tailed (Table 4). The slight difference in the intervals might have been the effect of the small $K (= 8)$ on the χ^2-approximation, although it was not substantial. On the contrary, the confidence interval for the sample mean over all loci was narrower than our credible interval (Table 4). The distribution of a sample mean is normal when the sample size is large with the variance reduced by the central limit theorem. A χ^2-distribution approaches a normal distribution as the degrees of freedom become larger. For our case of 1309 d.f. [d.f. $= (K-1)\sum_{i=1}^{n}(J_i - 1)$ as given in Weir and Hill 2002], the two distributions are equal. The property of the sample mean should cause the narrower confidence interval of the normal theory approach. The result shows that the posterior distribution of F_{ST}^{P} describes the distribution of $\widehat{F_{ST}^{P}}$ well, both for each locus and for the average over all loci.

LD among loci: The case study of the Pacific herring, based on a few microsatellite markers, did not detect significant variation in F_{ST}^{G} among loci in the genome (Figure 6). The assumption of normality for the MLE of F_{ST}^{G} is valid when more than a few sampling points are surveyed. This assumption might be violated when the data are collected from only a few sampling points or F_{ST}^{G} is close to 0 or 1.0. We also assumed the independence of loci ($l = 1, \ldots, L$). However, it is necessary to take into account the linkage disequilibrium (LD) among loci, when the molecular markers are tightly linked.

Recent progress in whole-genome analysis of human populations provides a new perspective on the inference of F_{ST} and its distribution in genomes (Garte 2003; Hinds *et al.* 2005; Weir *et al.* 2005; Walsh *et al.* 2006). In their Figure 1, Weir *et al.* (2005) showed that values for the single-locus marker $\widehat{F_{ST,l}^{G}}$ over the whole human genome for three (Perlegen) or four (HapMap) populations had a distribution very much like the χ^2-distributions with 2 or 3 d.f. and suggested that values of F_{ST}^{G} are genome-region specific. However, $\widehat{F_{ST}^{G}}$ follows a χ^2-distribution under constant F_{ST}^{G}, as reported by Weir and Hill (2002) and as demonstrated in our analysis of the Pacific herring. Therefore, the distributions of the values for the single-locus marker $\widehat{F_{ST,l}^{G}}$ in Weir *et al.* (2005) do not necessarily support the genome-region-specific F_{ST} hypothesis in the human genome.

Weir *et al.*'s 5-Mb window average values for $\widehat{F_{ST}^{G}}$ were close to normal because of the property of the sample mean. The standard deviations (SDs) of the 5-Mb window average values of $\widehat{F_{ST}^{G}}$ decreased substantially from single-locus estimates of 0.12 to 0.02 for HapMap and from 0.11 to 0.02 for Perlegen (Tables 2 and 3 in Weir *et al.* 2005). The average number of markers in a 5-Mb window was 1114 for HapMap and 1834 for Perlegen (calculated from Table 1 in Weir *et al.* 2005). Hence, these SDs are expected to be $0.00360 (= 0.12/\sqrt{1114})$ and $0.00257 (= 0.11/\sqrt{1834})$, if all SNPs are independent and distributed identically in the genome. The effective sample size (Kish 1965, p. 259) could be much smaller than the real number of SNPs in the 5-Mb windows if the $F_{\mathrm{ST}}^{\mathrm{G}}$'s are correlated, because of LD between the SNPs. A coalescent simulation of human population history implies that linkage equilibrium holds for SNPs separated by >10–100 kb (Figure 1 in Kruglyak 1999). Therefore, we can estimate the effective number of human SNPs per 100-kb window to be ∼1–20. If we assumed it to be ∼10, the effective number of SNPs per 5-Mb window becomes 500. Therefore, the SDs of the 5-Mb window $\widehat{F_{\mathrm{ST}}^{\mathrm{G}}}$ are expected to be $0.00537(= 0.12/\sqrt{500})$ for HapMap and $0.00492(= 0.11/\sqrt{500})$ for Perlegen. The actual value (0.02) is much larger than these, even when the LD among the SNPs is taken into account. This discrepancy can be explained by the large-scale heterogeneity of F_{ST} between the 5-Mb windows. New data show that LD is highly structured into discrete blocks of sequences separated by hot spots of recombination (Goldstein 2001; McVean *et al.* 2004) and differs among species (Hernandez *et al.* 2007). The simultaneous estimation of F_{ST}'s of SNPs and the LD between the SNPs should give us an accurate picture of the distribution of F_{ST} in genomes.

We are grateful to Bruce Weir for valuable comments on the manuscript and encouragement with this work. We also thank the anonymous reviewers for their constructive comments, which improved the earlier version of the manuscript. This work was supported by grants from the Japan Society for the Promotion of Science.

LITERATURE CITED

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

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

Balloux, F., and N. Lugon-Moulin, 2002 The estimation of population differentiation with microsatellite markers. Mol. Ecol. **11:** 155–165.

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–83.

Cockerham, C. C., 1973 Analysis of gene frequencies. Genetics **74:** 679–700.

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

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

Dizon, A. E., B. L. Taylor and G. M. O'Corry-Crowe, 1995 Why statistical power is necessary to link analyses of molecular variation to decisions about population structure, pp. 288–294 in *Evolution and the Aquatic Ecosystem*, edited by J. L. Nielsen and D. A. Powers. American Fisheries Society Symposium 17, Bethesda, MD.

Ewens, W. J., 2004 *Mathematical Population Genetics, I. Theoretical Introduction*, Ed. 2. Springer, Berlin/Heidelberg, Germany/New York.

Excoffier, L., 2003 Analysis of population subdivision, pp. 713–750 in *Handbook of Statistical Genetics*, Ed. 2, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Excoffier, L., G. Laval and S. Schneider, 2005 Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol. Bioinform. Online **1:** 47–50.

Garte, S., 2003 Locus-specific genetic diversity between human populations: an analysis of the literature. Am. J. Hum. Biol. **15:** 814–823.

Goldstein, D. B., 2001 Islands of linkage disequilibrium. Nat. Genet. **29:** 109–111.

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

Goudet, J., M. Raymond, T. de Meeüs and F. Rousset, 1996 Testing differentiation in diploid populations. Genetics **144:** 1933–1940.

Hanski, I., and O. E. Gaggiotti, 2004 *Ecology, Genetics, and Evolution of Metapopulations*. Academic Press, San Diego.

Hernandez, R. D., M. J. Hubisz, D. A. Wheeler, D. G. Smith, B. Ferguson *et al.*, 2007 Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science **316:** 240–243.

Hinds, D. A., L. L. Stuve, G. B. Nelsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072–1079.

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

Holsinger, K. E., P. O. Lewis and D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol. Ecol. **11:** 1157–1164.

Huelsenbeck, J. P., and P. Andolfatto, 2007 Inference of population structure under a Dirichlet process model. Genetics **175:** 1787–1802.

Kish, L., 1965 *Survey Sampling*. Wiley, New York.

Kitada, S., and H. Kishino, 2004 Simultaneous detection of linkage disequilibrium and genetic differentiation of subdivided populations. Genetics **167:** 2003–2013.

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

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

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lange, K., 1995 Application of the Dirichlet distribution to forensic match probabilities. Genetica **96:** 107–117.

Lockwood, J. R., K. Roeder and B. Devlin, 2001 A Bayesian hierarchical model for allele frequencies. Genet. Epidemiol. **20:** 17–33.

Maritz, J. S., and T. L. Lewin, 1989 *Empirical Bayes Methods*, Ed. 2. Chapman & Hall, London.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

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

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

Paetkau, D., W. Calvert, I. Stirling and C. Strobeck, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4:** 347–354.

Palsbøll, P. J., M. Bérubé and F. W. Allendorf, 2007 Identification of management units using population genetic data. Trends Ecol. Evol. **22:** 11–16.

Pannell, J. R., and B. Charlesworth, 2000 Effects of metapopulation processes on measures of genetic diversity. Philos. Trans. R. Soc. Lond. B **355:** 1851–1864.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **156:** 945–959.

Rannala, B., and J. A. Hartigan, 1996 Estimating gene flow in island populations. Genet. Res. **67:** 147–158.

Raufaste, N., and F. Bonhomme, 2000 Properties of bias and variance of two multiallelic estimators of F_{ST}. Theor. Popul. Biol. **57:** 285–296.

Raymond, M., and F. Rousset, 1995 GENEPOP (version 3.4): population genetics software for exact tests and ecumenicism. J. Hered. **86:** 248–249.

Reynolds, J., B. S. Weirs and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105:** 767–779.

Robertson, A., and W. G. Hill, 1984 Deviation from Hardy–Weinberg proportions: sample variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

Rosenberg, N. A., L. M. Li, R. Ward and J. K. Pritchard, 2003 Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. **73:** 1402–1422.

Rousset, F., 2003 Inferences from spatial population genetics, pp. 681–712 in *Handbook of Statistical Genetics*, Ed. 2, edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Ryman, N., and P. E. Jorde, 2001 Statistical power when testing for genetic differentiation. Mol. Ecol. **10:** 2361–2373.

Ryman, N., S. Palm, C. André, G. R. Carvalho, T. G. Dahlgren *et al.*, 2006 Power for detecting genetic divergence: differences between statistical methods and marker loci. Mol. Ecol. **15:** 2031–2045.

Seber, G. A. F., 1982 *The Estimation of Animal Abundance and Related Parameters*, Ed. 2. Griffin, London.

Stein, C., 1956 Inadmissibility of the usual estimator for the mean of a multivariate distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, Vol. 1, pp. 197–206. University of California Press, Berkeley, CA.

Torgerson, W. S., 1952 Multidimensional scaling: 1. theory and methods. Psychometrika **17:** 401–419.

Walsh, R. D., P. Sabeti, H. B. Hutcheson, B. Fry, S. F. Schaffner *et al.*, 2006 Searching for signals of evolutionary selection in 168 genes related to immune function. Hum. Genet. **119:** 92–102.

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

Weicker, J. J., R. T. Brumfield and K. Winker, 2001 Estimating the unbiased estimator θ for population genetic survey data. Evolution **55:** 2601–2605.

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

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

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

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen and W. G. Hill, 2005 Measures of human population structure show heterogeneity among genomic regions. Genome Res. **15:** 1468–1476.

Wright, S., 1951 The general structure of populations. Ann. Eugen. **15:** 323–354.

Wright, S., 1969 *Evolution and the Genetics of Populations: The Theory of Gene Frequencies*, Vol. 2. University of Chicago Press, Chicago.

Young, F. W., and R. M. Hamer, 1987 *Multidimensional Scaling: History, Theory and Applications*. Erlbaum, New York.

Communicating editor: R. W. Doerge