Title:
Empirical Bayes inference of pairwise $F_{ST}$ and its distribution in the genome

Authors and their Affiliations:
Shuichi Kitada*,1, Toshihide Kitakado* and Hirohisa Kishino†

* Faculty of Marine Science, Tokyo University of Marine Science and Technology, Minato, Tokyo 108-8477, Japan
† Graduate School of Agriculture and Life Sciences, University of Tokyo, Bunkyo, Tokyo 113-8657, Japan
Running Title:
Inferring $F_{ST}$ and its distribution

Keywords:
Empirical Bayes, distribution of $F_{ST}$, genome-wide estimation, pairwise FST, shrinkage estimator

1 Corresponding author:
Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato, Tokyo, 108-8477, Japan. Email: kitada@kaiyodai.ac.jp
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 paper 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 based on the estimated distribution of the global $F_{ST}$ for the herring and examples of human 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. 2006). 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 equals Wright’s $F_{ST}$ (Excoffier 2003). We use the notation $\theta_{WC}$ 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 only accounts 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 $\theta_{WC}$, which takes the sampling error for the subpopulations into account. Several moment estimators with different weighting schemes have also been derived (Weir and Cockerham 1984; Robertson and Hill 1984). An alternative estimation has been discussed using the method of ordinary least squares (Reynolds et al. 1983). Weir and Hill (2002) extended $\theta_{WC}$ to a population-specific parameter to allow different levels of coancestry for different populations. They also derived an estimator for $\theta_{WC}$ 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.

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 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. 2006). 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 \( \hat{F}_{ST} \). 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 $F_{STs}$. 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 paper, we propose an empirical Bayes procedure to estimate locus-specific pairwise $F_{STs}$, taking into account the uncertainty of the allele frequencies of subpopulations. The estimation procedure has two stages. Firstly, 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. Secondly, based on 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 $F_{STs}$ 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 $F_{STs}$ can be used to calculate a criterion of population differentiation. Our maximum likelihood estimates of the global $F_{STs}$ 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 based on 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

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

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 $F_{GST}^G$) at each locus is expressed simply by the scale parameter, as

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

(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 $j$th 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 $\theta_l$ refers to the rate of gene flow, as given by Rannala and Hartigan (1996). We use the symbol $\theta$ for the scale parameter, following Rannala and Hartigan (1996) and Balding and Nichols (1997). Weir and Cockerham (1984) also used the same symbol $\theta$ for the coancestry coefficient ($= F_{ST}$), so we use $\theta_{WC}$ for their $\theta$. Our $F_{GST}^G$ is equivalent to $\theta_{WC}$ (Weir 1996, pp. 47–48) and Holsinger’s $\theta^B$ (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 $F_{GST}^G$ 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 $F_{GST}^G$, 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 $n_{kl} = (n_{kl1}, \ldots, n_{klJ})'$ 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 $\alpha_l = (\alpha_{l1}, \ldots, \alpha_{lJ})'$. 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

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

The hyperparameters $\alpha_l$ 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 $\alpha_{lj}$ should be $\alpha_j$. Henceforth, for simplicity, we will focus on diploid organisms.

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

$$\hat{V}(\hat{F}^{G}_{ST,l}) \sim \frac{\hat{V}(\hat{\theta}_l)}{(1 + \hat{\theta}_l)^4}. \quad (3)$$

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 $F^{G}_{ST}$ 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 (Lange 1995; Weir 1996)

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

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 $F_{STS}$s (hereafter denoted as $F_{PST}^P$), 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 $F_{PST}^P$ at each locus can easily be generated based on $G_{ST}$ estimator, without using genotype frequencies. We set the number of each simulation to 10,000, so 10,000 $F_{PST}^P$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 $F_{PST}^P$, the posterior mean and 95% credible interval are calculated. We use the posterior mean as the empirical Bayes estimator of locus-specific $F_{PST}^P$. We can also calculate the probability that $F_{PST}^P$ is smaller than an arbitrary value (e.g., $P(F_{PST}^P \leq c)$), which can be used as the criterion
for population differentiation (Waples and Gaggiotti 2006; Palsbøll et al. 2006). The average estimate of $F_{ST}^P$ 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 based on the Kullback–Leibler information. The informativeness for assignment $I_n$ is defined as:

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

(5)

where $\bar{p}_{lj} = \frac{\sum_{k=1}^{K} p_{klj}}{K}$. 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 $F_{ST}^P$'s based on $F_{ST}^G$ 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 $F_{ST}^G$ among $L$ loci, $F_{ST,l}^G = F_{ST}^G(l = 1, \ldots, L)$ against the alternative hypothesis $H_1$, the heterogeneity of $F_{ST}^G$ among loci, based on estimates of $F_{ST,l}^G$. When a large number of subpopulations are sampled, the maximum likelihood estimate $\hat{F}_{ST,l}^G$ follows a normal distribution of $N(\bar{F}_{ST}^G, \hat{\sigma}_{l}^2)$. The maximum likelihood under $H_0$ is then given as

$$\hat{L}_0 = \prod_{l=1}^{L} \frac{1}{\sqrt{2\pi \hat{\sigma}_{l}^2}} e^{-\frac{(F_{ST,l}^G - \bar{F}_{ST}^G)^2}{2/\hat{\sigma}_{l}^2}}$$

where $\bar{F}_{ST}^G = \sum_{l=1}^{L} w_l F_{ST,l}^G / \sum_{l=1}^{L} w_l$, $w_l = (\hat{\sigma}_{l}^2)^{-1}$ and \( \hat{\sigma}_{l}^2 = \hat{V}(F_{ST,l}^G). \) The maximum likelihood under $H_1$ is $\hat{L}_1 = \prod_{l=1}^{L} \frac{1}{\sqrt{2\pi \hat{\sigma}_{l}^2}}$ maximizing $F_{ST,l}^G$ by $\hat{F}_{ST,l}^G$. The negative twice log likelihood ratio is then

$$\lambda = \sum_{l=1}^{L} \frac{(\hat{F}_{ST,l}^G - F_{ST,l}^G)^2}{\hat{\sigma}_{l}^2},$$

which follows the $\chi^2$ distribution with degree of freedom $(L-1)$ under the null hypothesis. We can test the heterogeneity of $F_{ST}^G$ based on the test statistics.

The other approach to investigate the heterogeneity of $F_{ST}^G$ is to estimate the dist-
tribution of $F_{ST}^G$ 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 $F_{ST}^G$ from the data. Here, we estimate the distribution of $F_{ST}^G$ in the genome from estimates of randomly selected loci, $\hat{F}_{ST,l}^G(l = 1, \ldots, L)$. When the distribution is expressed by the parametric model $f(F_{ST}^G|\rho)$, the unknown parameter $\rho$, which defines the distribution of $F_{ST}^G$, is estimated by maximizing the log marginal likelihood

$$\ell(\rho|\hat{F}_{ST,1}^G, \ldots, \hat{F}_{ST,L}^G) = \sum_{l=1}^L \log \int p(\hat{F}_{ST,l}^G|F_{ST,l}^G) f(F_{ST,l}^G|\rho) dF_{ST,l}^G,$$

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

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

$$\ell(\mu, \sigma^2|\hat{F}_{ST,1}^G, \ldots, \hat{F}_{ST,L}^G) = -\frac{1}{2} \sum_{l=1}^L \log (\sigma^2 + \sigma_l^2) - \frac{1}{2} \sum_{l=1}^L \frac{(\hat{F}_{ST,l}^G - \mu)^2}{\sigma^2 + \sigma_l^2}.$$  

(6)

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

RESULTS

12
Improved precision of our empirical Bayes estimator for pairwise $F_{ST}$: We investigated the performance of our method of estimating pairwise $F_{ST}^P$ using numerical simulations. Random vectors of allele frequencies at locus $l$ in subpopulation $k$, $p_{kl}$, 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 $F_{ST,l}$s ($= 1/(1 + \theta_l)$), we chose four different levels: $F_{ST,l}^G = 0.01, 0.05, 0.1$, 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 $\text{Multi}(N_k, p_{kl})$ for $K$ localities. The pairwise $F_{ST}^P$ values between the first and second localities were evaluated by the conventional Nei’s $G_{ST}$ estimator and the empirical Bayes method. In the later procedure, 500 $F_{ST}^P$-s were simulated based on 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 $F_{ST}^P$. These procedures were iterated $R = 1,000$ times.

Figures 1 and S1 in the Supplementary material show the general features of the empirical Bayes estimator compared with those of the conventional $G_{ST}$ estimator. The former examines 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 $F_{ST}^G$. In contrast, the empirical Bayes estimates of $F_{ST}^P$-s shrink.
towards the average of the true $F_{ST}^P$ and reduces 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 Supplementary material. The numbers of sampling points did not affect the positive bias of the conventional $G_{ST}$ estimator because $K$ was fixed at two. The empirical Bayes procedure provided larger variation for smaller numbers of sub-populations especially for small $F_{ST}^G$ values, and vice versa (FigureS2, Supplementary material). Conversely, larger numbers of sampled individuals reduced the positive bias of the conventional $G_{ST}$ estimator and the variation of the empirical Bayes estimator (Figure S3).

For a quantitative comparison of the two estimators, we used the following two measures: $\frac{1}{R} \sum_{r=1}^{R} (\hat{F}_{ST,r}^P - F_{ST,r}^P) / F_{ST,r}^P$ (relative mean bias) and $\sqrt{\frac{1}{R} \sum_{r=1}^{R} (\hat{F}_{ST,r}^P - F_{ST,r}^P)^2 / F_{ST,r}^P}$ (root relative mean squared error), where $\hat{F}_{ST,r}^P$ and $F_{ST,r}^P$ are an estimate and the true value, respectively for pairwise $F_{ST}^P$ in the $r$th 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 $F_{ST}^G$ values. The bias and variation became smaller as $F_{ST}^G$ became larger. In contrast, the empirical Bayes estimator provided smaller biases and variations for all cases of $F_{ST}^G$, 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 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 numbers 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. 2006). 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 $n_{kl}$s were drawn independently from the multinomial distribution $\text{Multi}(N_k, p_{kl})$ 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 Figures 3 (case 1) and S4 (case 2) in the Supplementary material (top left), our empirical Bayes procedure provided better estimates than those of the conventional $G_{ST}$ estimator for smaller $F_{ST}^{P}$ (approximately smaller than 0.06 for case 1 and 0.04 for case 2). In conservation, management units should be defined among sub-populations with little genetic differentiation. Palsbøll et al. (2007) concluded that $F_{ST}^{P} = 0.0025$ could be used as the criterion for deciding the 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 $F_{ST}^{P}$ in both cases. Our method reduced the positive bias for small $F_{ST}^{P}$ and the bias became negative for larger $F_{ST}^{P}$. 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, resulted in 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 S4 in the Supplementary material [Case 2]).

We also investigated the robustness of estimating population structure based on estimated $F_{ST}^{P}$. The results of the multidimensional scaling (MDS) (Torgerson 1952; Young and Harmer 1978) 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 based on otoliths stained with alizarin complex. A total of 2,055 mature individuals were genotyped at five microsatellite loci. Allele frequencies are given in Tables S1–S5 (Supplementary material). 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 (Figure S5, Supplementary material). Based on the estimates of the hyperparameters (Figure S6, Supplementary material), the scale parameter $\theta_l$ 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 $F_{ST}^P \leq 0.01$ 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 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 $\hat{F}_{ST,l}^G$ over five loci was not trivial, with a coefficient of variation (CV) of 23% (Table 3). However, the negative twice log likelihood ratio $\lambda$ was calculated to be 6.9264 and the hypothesis of constant $F_{ST}^G$ for all loci was not rejected ($p = 0.8602$, df=4). We estimated the distribution of $F_{ST}^G$ in the genome, assuming normality based on Equation 6. The maximum likelihood estimates (MLE) for $\mu$ and $\sigma$ were obtained with 90%, 95%, and 99% confidence intervals, which define the distribution of $F_{ST}^G$ in the genome via the likelihood profile (Figure 6, left). The MLE for $\mu$, with the 95% confidence interval, was 0.0160 (0.0128, 0.0206) (Figure 6, left [a,e]), and for $\sigma$ was 0 (0, 0.00716) (Figure 6, left [0,c]). The weighted mean $\bar{F}_{ST}^G$ coincided with the MLE of $\mu$ (Table 3) and the distribution of the weighted mean $N(\hat{F}_{ST}^G, \hat{V}(\hat{F}_{ST}^G))$, shown as the blue line in Figure 6, right, described well the 95% confidence interval of $\mu$ (Figures 6 [a,e]). Here, $\hat{V}(\hat{F}_{ST}^G) = 1/(\sum_{l=1}^{L} w_i)^2 \sum_{l=1}^{L} w_i^2 \hat{V}(\hat{F}_{ST,l}^G)$ and $V(\hat{F}_{ST,l}^G)$ were estimated with Equation 3. The log likelihood profile for $\sigma$ was monotonic and decreasing, indicating that the point estimate of $\sigma$ was 0 (Figure S7, Supplementary material). Hence, the point estimate of the distribution of $F_{ST}^G$ for the Pacific herring is constant, supporting the hypothesis of constant $F_{ST}^G$ throughout the genome. However, the distribution of
$F_{ST}^G$ has uncertainty, which accounts for the confidence regions of $\mu$ and $\sigma$ (Figure 6, right).

DISCUSSION

Our empirical Bayes estimator of $F_{ST}^P$ performed better than the conventional $G_{ST}$ estimator for various 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 $F_{ST}^P$ 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, Figure S2, Supplementary material). With a small number of sampling points, the maximum likelihood estimate (MLE) of $\theta$ 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 (ILE) of $\theta$ 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 (Figure S8). 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 based on MLE of $\theta$, given in Table 1 (top three rows). The negative bias of $\hat{F}_{ST}^P$s based on 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 $\beta_l$ and eliminates $\beta_l$ by integration regarding as nuisance parameters. By using ILE of $\theta$ instead of 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 (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 $G_{ST}$ 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, $\theta_{WC}$ 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 1,225. Such 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 $F_{ST}$ based on the relationship between $G_{ST}$ and $\theta_{WC}$. 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):

$$\theta_{WC} = \frac{\left\{G_{ST} \left( K + \frac{K-1}{N_k-1} \right) - \frac{K-1}{N_k-1} \right\} / (G_{ST} + K - 1)}{G_{ST} + K - 1}.$$  

Using this equation, we calculated the conventional estimates of pairwise Weir and Cockerham’s $\theta_{WC}$ from $G_{ST}$ estimates with the simulation protocol described in RESULTS. We examined cases with a global $F_{ST}$ of $F^G_{ST} = 0.01, 0.05, 0.1$ 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 Figure S9 in the Supplementary material, the two estimators have linear relationships. Hence, our simulation results on the conventional $G_{ST}$ estimator can be extended straightforwardly to pairwise $\theta_{WC}$. 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 $F^P_{ST}$ (Figure S10, Supplementary material), which coincides with our simulations for small $F^G_{ST}$s (Figures 1, S1, and S11, Supplementary material).

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 \( \hat{F}_{G_{ST}}^C \). \( F_{G_{ST}}^C \) is the variance of the allele frequencies among subpopulations relative to the total population. Hence, \( \hat{F}_{G_{ST}}^C \) refers to a sample variance of allele frequencies, and therefore \( \text{df} \cdot \frac{\hat{F}_{G_{ST}}^C}{F_{G_{ST}}^C} \) follows a \( \chi^2 \) 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 \( F_{G_{ST}}^P \) at each locus was unimodal and slightly right tailed, which reminded us of \( \chi^2 \) distributions (Figure 4). We estimated Weir and Hill’s confidence intervals by substituting the posterior mean of \( F_{G_{ST}}^P \) with \( \hat{\theta}_N \) as \( [\text{df}\hat{F}_{G_{ST}}^P/\chi^2_{(d,0.975)}], [\text{df}\hat{F}_{G_{ST}}^P/\chi^2_{(d,0.025)}] \), where \( \text{df} = (K - 1)(J - 1) \) is the degree of freedom. The confidence intervals of \( F_{G_{ST}}^P \) obtained with the \( \chi^2 \) approximation coincided well with the credible intervals calculated from the posterior distributions, although our credible intervals were narrower than the confidence intervals of the \( \chi^2 \) 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 \( \chi^2 \) 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 \( \chi^2 \) distribution approaches a normal distribution as the degree of freedom becomes larger. For our case of 1,309 df (\( \text{df} = (K - 1) \sum_{i=1}^n (J_i - 1) \)) 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 \( F_{G_{ST}}^P \) describes the distribution of \( \hat{F}_{G_{ST}}^P \) 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 \( F_{G_{ST}}^C \) 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 of 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 $\hat{F}_{ST,l}^G$ over the whole human genome for three (Perlegen) or four (HapMap) populations had a distribution very much like the $\chi^2$ distributions with two or three degrees of freedom, and suggested that values of $F_{ST}^G$ are genome-region specific. However, $\hat{F}_{ST}^G$ follows a $\chi^2$ 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 $\hat{F}_{ST,l}^G$ 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 $\hat{F}_{ST}^G$ were close to normal because of the property of the sample mean. The standard deviations (SD) of the 5 Mb window average values of $\hat{F}_{ST}^G$ 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 1,114 for HapMap and 1,834 for Perlegen (calculated from Table 1 in Weir et al. 2005). Hence, these SDs are expected to be 0.00360 ($= 0.12/\sqrt{1,114}$) and 0.00257 ($= 0.11/\sqrt{1,834}$), 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 $F_{ST}$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 more than 10–100 kb (Figure 1 in Kruglyak 1999). Therefore, we can estimate the effective number of human SNPs per 100 kb window to be about 1–20. If we assumed it to be around 10, the effective number of SNPs per 5 Mb window becomes 500. Therefore, the SDs of the 5 Mb window $\hat{F}_{ST}$ 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 have improved the earlier version of the manuscript. This work was supported by grants from the Japan Society for the Promotion of Science.

LITERATURE CITED


Excoffier, L., P. Smouse, and J. Quattro, 1992 Analysis of molecular variance inferred


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


### 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)

<table>
<thead>
<tr>
<th>$F_{ST}^G$</th>
<th>$K$</th>
<th>$N_k/2$</th>
<th>Conventional</th>
<th>Empirical Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>5</td>
<td>20</td>
<td>2.770 (3.019)</td>
<td>0.726 (0.815)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>1.769 (1.942)</td>
<td>-0.541 (0.630)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.069 (1.192)</td>
<td>-0.330 (0.412)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td>2.666 (2.880)</td>
<td>-0.325 (0.464)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>1.791 (1.955)</td>
<td>-0.205 (0.345)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.081 (1.214)</td>
<td>-0.122 (0.270)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>20</td>
<td>2.728 (2.972)</td>
<td>0.003 (0.294)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>1.784 (1.967)</td>
<td>0.010 (0.272)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1.085 (1.217)</td>
<td>0.033 (0.251)</td>
</tr>
<tr>
<td>0.05</td>
<td>5</td>
<td>20</td>
<td>0.561 (0.685)</td>
<td>-0.136 (0.283)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.363 (0.478)</td>
<td>-0.097 (0.233)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.218 (0.304)</td>
<td>-0.056 (0.175)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td>0.538 (0.662)</td>
<td>-0.039 (0.242)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.367 (0.475)</td>
<td>-0.015 (0.216)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.233 (0.317)</td>
<td>0.004 (0.169)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>20</td>
<td>0.557 (0.680)</td>
<td>0.047 (0.253)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.366 (0.480)</td>
<td>0.031 (0.226)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.206 (0.304)</td>
<td>0.015 (0.181)</td>
</tr>
<tr>
<td>0.1</td>
<td>5</td>
<td>20</td>
<td>0.273 (0.418)</td>
<td>-0.053 (0.242)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.176 (0.284)</td>
<td>-0.036 (0.189)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.106 (0.194)</td>
<td>-0.019 (0.149)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td>0.255 (0.375)</td>
<td>-0.006 (0.215)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.183 (0.289)</td>
<td>0.009 (0.186)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.098 (0.195)</td>
<td>-0.001 (0.151)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>20</td>
<td>0.247 (0.373)</td>
<td>0.019 (0.215)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.182 (0.287)</td>
<td>0.027 (0.183)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.109 (0.192)</td>
<td>0.020 (0.147)</td>
</tr>
<tr>
<td>0.2</td>
<td>5</td>
<td>20</td>
<td>0.125 (0.250)</td>
<td>-0.017 (0.192)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.083 (0.195)</td>
<td>-0.007 (0.161)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.051 (0.134)</td>
<td>-0.003 (0.118)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td>0.123 (0.246)</td>
<td>0.009 (0.192)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.092 (0.197)</td>
<td>0.018 (0.165)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.049 (0.138)</td>
<td>0.004 (0.122)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>20</td>
<td>0.127 (0.267)</td>
<td>0.027 (0.220)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.087 (0.188)</td>
<td>0.020 (0.155)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.058 (0.144)</td>
<td>0.018 (0.128)</td>
</tr>
</tbody>
</table>
**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 1,000 simulations under the stepping-stone models (see text)

<table>
<thead>
<tr>
<th>$F^P_{ST}$</th>
<th>Relative mean bias</th>
<th>Root relative MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conventional</td>
<td>Empirical Bayes</td>
</tr>
<tr>
<td>0.001 (Case 1)</td>
<td>1.165</td>
<td>1.239</td>
</tr>
<tr>
<td>0.0005 (Case 2)</td>
<td>2.296</td>
<td>1.718</td>
</tr>
</tbody>
</table>

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 $\theta$ and $F_{ST}^G$ over subpopulations of the Pacific herring, with standard errors in parentheses

<table>
<thead>
<tr>
<th>Locus</th>
<th>No. alleles</th>
<th>$\theta$</th>
<th>$F_{ST}^G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cha17</td>
<td>48</td>
<td>63.98 (7.88)</td>
<td>0.0154 (0.0019)</td>
</tr>
<tr>
<td>Cha20</td>
<td>30</td>
<td>77.26 (13.13)</td>
<td>0.0128 (0.0021)</td>
</tr>
<tr>
<td>Cha63</td>
<td>35</td>
<td>58.21 (8.60)</td>
<td>0.0169 (0.0025)</td>
</tr>
<tr>
<td>Cha113</td>
<td>29</td>
<td>41.68 (6.59)</td>
<td>0.0234 (0.0036)</td>
</tr>
<tr>
<td>Cha123</td>
<td>50</td>
<td>58.78 (6.68)</td>
<td>0.0167 (0.0018)</td>
</tr>
<tr>
<td>Mean$^a$</td>
<td>38.4</td>
<td>59.98 (3.98)</td>
<td>0.0170 (0.0011)</td>
</tr>
<tr>
<td>Mean$^b$</td>
<td></td>
<td></td>
<td>0.0160 (0.0010)</td>
</tr>
</tbody>
</table>

$^a$: 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

<table>
<thead>
<tr>
<th>Locus</th>
<th>df</th>
<th>Posterior mean</th>
<th>Empirical Bayes</th>
<th>Weir and Hill$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cha17</td>
<td>47</td>
<td>0.0193</td>
<td>[0.0137, 0.0258]</td>
<td>[0.0133, 0.0302]</td>
</tr>
<tr>
<td>Cha20</td>
<td>29</td>
<td>0.0100</td>
<td>[0.0063, 0.0146]</td>
<td>[0.0063, 0.0180]</td>
</tr>
<tr>
<td>Cha63</td>
<td>34</td>
<td>0.0064</td>
<td>[0.0046, 0.0086]</td>
<td>[0.0042, 0.0111]</td>
</tr>
<tr>
<td>Cha113</td>
<td>28</td>
<td>0.0245</td>
<td>[0.0162, 0.0349]</td>
<td>[0.0154, 0.0448]</td>
</tr>
<tr>
<td>Cha123</td>
<td>49</td>
<td>0.0213</td>
<td>[0.0158, 0.0279]</td>
<td>[0.0149, 0.0331]</td>
</tr>
<tr>
<td>mean</td>
<td>1,309</td>
<td>0.0163</td>
<td>[0.0138, 0.0191]</td>
<td>[0.0151, 0.0176]</td>
</tr>
</tbody>
</table>

$^a$: Estimated by Weir and Hill’s (2002) normal theory approach
**Figure legends**

**FIGURE 1.** The conventional $G_{ST}$ (A) and empirical Bayes (B) estimates of pairwise $F_{ST}^P$ from 1,000 simulations under the infinite island model at various levels of $F_{ST}^G$ (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.

**FIGURE 2.** The conventional informativeness of assignment $I_n$ (Rosenberg et al. 2003) (A) and empirical Bayes (B) estimates of $I_n$ from 1,000 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.

**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 1,000 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. The figure only shows the results for case 1. The results of the MDS analysis of two data sets are given in the bottom panel; black circle shows the true population structure, and the estimated population structure based on the conventional (red) and empirical Bayes estimates (blue) of $F_{ST}^P$. 
FIGURE 4. Posterior distributions of $F^P_{ST}$ for the Pacific herring between Funka Bay and Miyako Bay at each locus, and over all loci, which were averaged over $F^P_{ST}$ 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.

FIGURE 6. Confidence regions of the distribution of $F^G_{ST}$ in the genome of the Pacific herring, assuming a normal distribution (see the text). (A) The confidence regions of $\mu$ and $\sigma^2$, 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 $\hat{F}^G_{ST}$ is superimposed, and labeled f (blue line).
Fig. 1 by Kitada et al.
Fig. 2 by Kitada et al.
Fig. 3 by Kitada et al.
Fig. 4 by Kitada et al.
Fig. 5 by Kitada et al.
Fig. 6 by Kitada et al.