MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) **17**, 1210–1222 doi: 10.1111/1755-0998.12663

The empirical Bayes estimators of fine-scale population structure in high gene flow species

SHUICHI KITADA,* REIICHIRO NAKAMICHI*,[1] and HIROHISA KISHINO†
*Graduate School of Marine Science and Technology, Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato-ku, Tokyo 108-8477, Japan, †Graduate School of Agriculture and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

Abstract

An empirical Bayes (EB) pairwise F_{ST} estimator was previously introduced and evaluated for its performance by numerical simulation. In this study, we conducted coalescent simulations and generated genetic population structure mechanistically, and compared the performance of the EBF_{ST} with Nei's G_{ST}, Nei and Chesser's bias-corrected G_{ST} (G_{ST_NC}), Weir and Cockerham's θ (θ_{WC}) and θ with finite sample correction (θ_{WC_F}). We also introduced EB estimators for Hedrick' G'_{ST} and Jost' D. We applied these estimators to publicly available SNP genotypes of Atlantic herring. We also examined the power to detect the environmental factors causing the population structure. Our coalescent simulations revealed that the finite sample correction of θ_{WC} is necessary to assess population structure using pairwise F_{ST} values. For microsatellite markers, EBF_{ST} performed the best among the present estimators regarding both bias and precision under high gene flow scenarios ($F_{ST} \leq 0.032$). For 300 SNPs, EBF_{ST} had the highest precision in all cases, but the bias was negative and greater than those for G_{ST_NC} and θ_{WC_F} in all cases. G_{ST_NC} and θ_{WC_F} performed very similarly at all levels of F_{ST}. As the number of loci increased up to 10 000, the precision of G_{ST_NC} and θ_{WC_F} became slightly better than for EBF_{ST} for cases with $F_{ST} \geq 0.004$, even though the size of the bias remained constant. The EB estimators described the fine-scale population structure of the herring and revealed that ~56% of the genetic differentiation was caused by sea surface temperature and salinity. The R package FINEPOP for implementing all estimators used here is available on CRAN.

Keywords: Atlantic herring, empirical Bayes, microsatellite, pairwise F_{ST}, SNP

Received 8 September 2016; revision received 9 February 2017; accepted 21 February 2017

Introduction

Wright's F_{ST} is the most widely used measure of genetic divergence among populations in the fields of population and evolutionary genetics (Weir & Hill 2002; Holsinger & Weir 2009), conservation and management (Palsbøll *et al.* 2006), and seascape (Selkoe *et al.* 2008) and landscape genetics (Storfer *et al.* 2010). Wright (1951) defined F_{ST} as the correlation between randomly sampled gametes relative to the total drawn from the same population. Nei (1973) derived a formula to measure the genetic differentiation between populations denoted by G_{ST}, which is identical to F_{ST} (Appendix 1). The numerator of G_{ST} represents the variance in allele frequencies between populations. Therefore, its estimate is biased, even though the estimated allele frequencies are unbiased (Appendix 2). To overcome this problem, Nei & Chesser (1983) derived unbiased estimators for the numerator and denominator of G_{ST}, and corrected the bias in G_{ST} (hereafter G_{ST_NC}). Weir & Cockerham (1984) also proposed a bias-corrected moment estimator $\hat{\theta}$ (θ_{WC}) for the coancestry coefficient in the analysis of variance framework. θ_{WC} is the ratio of the unbiased estimators of the between-population variance of allele frequencies to the total variance component and is an estimator of F_{ST} (Weir & Cockerham 1984). These F_{ST} estimators were originally developed to estimate the mean F_{ST} over a metapopulation based on a set of population samples, which is often called global F_{ST} (e.g. Pérez-Lezaun *et al.* 1997). G_{ST} considers inference on observed set of populations sampled, while θ_{WC} considers replicates of a set of populations (Weir & Cockerham 1984). In addition to the global F_{ST}, F_{ST} values between pairs of population samples (pairwise F_{ST}) are routinely used to estimate population structure.

In high gene flow species, such as marine fish, the weak genetic signal of population differentiation hinders

Correspondence: Shuichi Kitada, Fax: +81-3-5463-0536;
E-mail: kitada@kaiyodai.ac.jp
[1]Present address: Research Center for Bioinformatics and Biosciences, National Research Institute of Fisheries Science, Yokohama 236-8648, Japan

the precise estimation of population genetic parameters (Waples 1998). Larger sampling variances for smaller sample sizes would also make it more difficult to correctly estimate F_{ST}. Thus, there is a high risk of obtaining biased F_{ST} values, resulting in the detection of spurious population structures. Because allele frequencies are very similar among populations in such cases, estimation of the between-population heterozygosity is not precise, especially when highly polymorphic markers such as microsatellite loci are used. To address this problem, we previously proposed an empirical Bayes (EB) method, which generates posterior distributions of pairwise F_{ST} using a Dirichlet distribution [or a beta for single-nucleotide polymorphisms (SNPs)] based on the G_{ST} formula (Kitada *et al.* 2007). The mean of the posterior distribution is defined as EBF_{ST} estimator. However, the performance testing was limited to G_{ST}, and θ_{WC} was evaluated via a function of G_{ST} using parametric simulations based on a Dirichlet distribution.

In this study, we explored the performance of our EBF_{ST} estimator relative to other established methods using coalescent simulations that generate genetic population structure mechanistically. In addition, we introduced new EB estimators (hereafter EBG_{ST_H} and EBD_J) for G'_{ST} (Hedrick 2005; G_{ST_H}) and D (Jost 2008; D_J). We applied these estimators to publicly available data set of Atlantic herring (*Clupea harengus*) SNP genotypes and inferred the population structure. We also evaluated the power to detect environmental effects, such as those of sea temperature, salinity, and geographical distance, on the herring F_{ST}, taking the correlation between F_{ST} values into account based on regression analyses using bootstrapping. Atlantic herring is distributed across a wide geographical area with steep gradients of salinity and sea surface temperature from the North Sea to the inner Baltic Sea, but its F_{ST} values were reported to be very small (Bekkevold *et al.* 2005; Gaggiotti *et al.* 2009). Thus, herring is one of the best species to test the performance of the EB estimators in high gene flow scenarios.

Materials and methods

Performance of F_{ST} estimators by coalescent simulations

To test the performance of F_{ST} estimators, we conducted coalescent simulations using the software `ms` (Hudson 2002) and generated genotype data under Wright's island model. The number of populations sampled was set to 30. In each population, microsatellite genotypes were obtained from 50 individuals and SNP genotypes were obtained from 25 individuals. The number of markers was set to 10 and 60 for microsatellites, and 300 and 10 000 for SNPs. We generated genotypes for eight levels of the true F_{ST} value, 0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064 and 0.128, which cover the extent of population differentiation from marine fish to human. The true F_{ST} values for microsatellite genotypes were computed under the infinite allele model given by Eq. (3) in the paper by Rousset (1996) as:

$$F_{ST} = \gamma d/(\gamma d + N_0(2 - \sigma a\gamma)(1 - \gamma d)),$$

where $\gamma = (1 - \mu)^2, a = (1 - m)^2 + m^2/(r - 1), b = (1 - a)/(r - 1)$, and $d = a - b = (1 - m[r/(r - 1)])^2$.

Here, μ is the mutation rate per generation for all alleles, m is the migration rate per generation, and r is the number of subpopulations sampled from a metapopulation. We substituted $\sigma = 0$ for hermaphroditic populations. The true pairwise F_{ST} values for SNP genotypes were computed as $F_{ST} = 1/(4N_0m + 1)$ (Wright 1951). In the coalescent simulations, we set the diploid population size to $N_0 = 500$ (which corresponds to an effective population size of N_e=1 000), and the migration rate was given by $m = (1/F_{ST} - 1)/4N_0$. The mutation rate for the entire microsatellite locus was set to $\mu = 5 \times 10^{-5}$ per locus per generation to generate the mean number of alleles (~20) for marine fish (DeWoody & Avise 2000), which is an order of magnitude smaller than $\mu = 10^{-3} - 10^{-4}$ for human microsatellites (Sun *et al.* 2012). For SNPs, we set $4N_0\mu = 0.3$ to generate a heterozygosity value of ~0.3, which is consistent with observations of heterozygosity in Atlantic herring, namely 0.31 ± 0.01 (Limborg *et al.* 2012a).

We computed pairwise F_{ST} values for G_{ST}, G_{ST_NC}, θ_{WC} and EBF_{ST} estimators based on the generated genotype data. Additionally, we used a modified calculation of θ_{WC}, termed θ_{WC_F}, to account for the fixed sampling of population pairs because θ_{WC} accounts for the replication of sampled populations (r). Our finite sample correction replaces a with $a(r - 1)/r$ in Eq. (2) on p. 1359 of the paper by Weir & Cockerham (1984). This was done because the pairwise F_{ST} value is calculated for specific population pairs, so applying the fixed-effect model of population sampling (Weir 1996) is appropriate. We assumed that the scale parameter (θ) of a Dirichlet (for microsatellite loci) or a beta (SNPs) distribution is common to all loci, but that mean allele frequencies differ for each locus in the EB F_{ST} estimation. The scale parameter (θ) was estimated numerically by maximizing the marginal likelihood function under this assumption [Eq. (2) in the paper by Kitada *et al.* 2007]. The simulation procedure was replicated 10 times, and a total of $B = 10 \times 30(30 - 1)/2 = 4350$ pairwise F_{ST} values were obtained for each F_{ST} estimator. The mean bias (MB) $\frac{1}{B}\sum_{i=1}^{B}(\hat{F}_{ST,i} - F_{ST})$ and root mean squared error (RMSE) $\sqrt{\frac{1}{B}\sum_{i=1}^{B}(\hat{F}_{ST,i} - F_{ST})^2}$ were compared.

EB estimators of other differentiation estimators

Posterior distributions for any parametric functions of gene frequencies can be generated by the EB procedure (Kitada *et al.* 2000). Therefore, the posterior distributions of the new G_{ST}-related measures G_{ST_H} (Hedrick 2005) and D_J (Jost 2008) are easily introduced as,

$$G_{ST\_H}^{pos} = \frac{(H_T^{pos} - H_S^{pos})}{H_T^{pos}} \frac{(1 + H_S^{pos})}{1 - H_S^{pos}} = G_{ST}^{pos} \frac{(1 + H_S^{pos})}{1 - H_S^{pos}}$$

$$D_J^{pos} = \frac{(H_T^{pos} - H_S^{pos})}{1 - H_S^{pos}} \frac{r}{r-1} = G_{ST}^{pos} \frac{H_T^{pos}}{1 - H_S^{pos}} \frac{r}{r-1},$$

respectively, where G_{ST}^{pos}, H_T^{pos}, and H_S^{pos} are the posterior distributions of G_{ST}, H_T, and H_S, and r is the number of subpopulations sampled. These are generated from a Dirichlet and/or a beta distribution given the estimate of the Dirichlet or beta scale parameter ($\hat{\theta}$). The mean of the posterior distribution is an EB estimator of G_{ST_H} (EBG_{ST_H}) and D_J (EBD_J), and the mean of G_{ST}^{pos} is our EBF_{ST} estimator.

Population structure of Atlantic herring

We analysed the publicly available SNP genotype data over the 281 loci in 21 Atlantic herring samples (n = 607) (Limborg *et al.* 2012a,b). Genotype data obtained during different years from the same sampling locations were combined because there was no difference among years (Limborg *et al.* 2012a), resulting in 18 samples. The 18 sampling locations are abbreviated as follows: NOR (Norway), ICE (Iceland), SHE (Shetland), WIR (Western Ireland), CLS (Celtic Sea), IRS (Irish Sea), EC (English Channel 1999/2009), CNS (Central North Sea), RF (Ringkøbing Fjord), LIM (Limfjord), SKA (Skagerrak), KAT (Kattegat), RUG (Rügen 2003/2009), HB (Hanö Bay), GD (Gdansk), GR (Gulf of Riga 2002/2008), GF (Gulf of Finland) and BB (Bothnian Bay). We calculated the pairwise EBF_{ST}, G_{ST_NC} and θ_{WC_F} values based on the 281 SNPs, including 16 outlier loci that were significantly correlated with environmental factors, such as annual mean temperature and salinity (Limborg *et al.* 2012a). G_{ST} was also calculated to determine the effect of the bias correction on the estimators compared with the original F_{ST} definition. We calculated θ_{WC} using GENEPOP4.2 (Raymond & Rousset 1995; Rousset 2008). We also calculated the new differentiation estimators, G_{ST_H} and D_J (D_{est} in Jost 2008), based on the unbiased estimators of H_T and H_S (Nei & Chesser 1983), and EBG_{ST_H} and EBD_J using FINEPOP1.3. Based on these pairwise F_{ST} and the new differentiation estimates, we depicted the population structure by drawing UPGMA trees.

Detecting effects of environmental factors on genetic differentiation

We performed regression analyses of the pairwise F_{ST} values against geographical distance and the differences in sea surface temperature and sea surface salinity to examine the effect of environmental variables on population differentiation using the 281 SNPs from the work of Limborg *et al.* (2012a) (Appendix S1, S2, Supporting information). We evaluated the predictive power of explanatory (environmental) variables and their combinations, instead of testing correlations between each explanatory valuable and the pairwise F_{ST} values by the partial Mantel test, to avoid the potential bias caused by correlations among the elements of distance matrices (Guillot & Rousset 2013). If we do not take account of correlations in F_{ST} values between pairs of sampling points, the standard errors of the regression coefficients may be underestimated. This would result in a radical significance test of environmental variables. To overcome this problem, we conducted bootstrapping to increase the precision of the regression coefficients. We resampled locations with replacement (18 local samples, n = 607). We also resampled the member individuals with replacement from the sampled populations.

We calculated pairwise EBF_{ST} and θ_{WC_F} values using FINEPOP1.3 for each bootstrap sample and estimated regression coefficients for the F_{ST} values. This procedure was iterated 100 times, and the standard deviation (SD) of the regression coefficients was calculated. We then computed the Z-value by dividing the estimated mean coefficient by its SD for each regression coefficient. The Z-value follows a normal distribution N (0, 1) and therefore provides a P-value for the significance of each regression coefficient. All possible model combinations for the environmental explanatory variables were examined, including their interactions with EBF_{ST} and θ_{WC_F}. The full model was as follows:

$$F_{ST} = \beta_1 D + \beta_2 T + \beta_3 S + \beta_4 D \times T + \beta_5 D \times S + \beta_6 T \times S + \beta_7 D \times T \times S$$

(see Table S3, Supporting information).

Here, D is the shortest ocean path, and T and S are the absolute differences in sea surface temperature and sea surface salinity between-population pairs, respectively. The parameters $\beta_1, \ldots, \beta_7$ are the partial regression coefficients. As the objective variables (pairwise F_{ST} values) were correlated, the effective sample size was less than the actual number of pairs; thus, it was necessary to modify the Akaike Information Criterion (AIC) with the likelihood assuming *iid* error terms (Akaike 1973) to select the explanatory variables. We used the Takeuchi Information Criterion (TIC; Takeuchi 1976;

Burnham & Anderson 2002), which considers the effective sample size (Kish 1965; Skinner *et al.* 1989) as an extension of the AIC:

$$\mathrm{TIC} = -2 \times [\text{maximum log likelihood}] + 2 \times \mathrm{trace}[A^{-1}B]$$

where A is the variance–covariance matrix of the regression coefficients assuming *iid* for the error terms. B is the variance of the estimated regression coefficients based on bootstrap resampling of the locations and individual sample members. The term of trace $[A^{-1}B]$ is the effective number of parameters. The best fit model with the minimum TIC value was selected for EBF_{ST} and for θ_{WC}, and we compared the performance of the two methods based on the R^2 value.

Results

Coalescent simulations

Our coalescent simulations revealed that the finite sample correction of θ_{WC} is necessary to properly assess the population structure using pairwise F_{ST} values (Fig. 1).

For microsatellite genotypes, the mean ± SD (range) number of alleles was between 21.4 ± 4.3 (12–31) and 24.2 ± 4.7 (15–33). The results for G_{ST} indicated the performance of the F_{ST} estimator without bias correction, and the bias was positive (Fig. 1a, b, Table S1, Supporting information). EBF_{ST} performed the best among the estimators regarding both bias and precision when F_{ST} ≤0.032. The bias and variance of θ_{WC} were greater than for G_{ST} when $F_{ST} > 0.008$, and the median of θ_{WC} values was approximately double those of G_{ST_NC}, θ_{WC_F}, and EBF_{ST}. G_{ST_NC} and θ_{WC_F} performed the same for all levels of F_{ST}. The RMSE of the EB F_{ST} estimator was half to one-third the size of the RMSE of the other estimators for the high gene flow scenarios of $F_{ST} < 0.016$. The difference in RMSE diminishes with a decreased level of gene flow and becomes almost the same among the estimators when $F_{ST} > 0.128$. Increasing the number of loci from 10 to 60 was ineffective at reducing the bias but improved the precision for all F_{ST} estimators.

For SNP genotypes at 300 loci, the RMSE of the EBF_{ST} estimator was the smallest in all cases, and half or one-third the size of those of the other estimators for the high gene flow scenarios where $F_{ST} < 0.016$, although the EBF_{ST} estimator had greater negative bias than G_{ST_NC} and θ_{WC_F} (Fig. 1c, d, Table S2, Supporting information). Consistent with our results obtained using microsatellite markers, the difference in RMSE diminishes with a decreased level of gene flow. Additionally, the relative bias diminishes with a decreased level of gene flow. The variance of the estimators decreased with an increase in the number of loci, whereas the bias remained constant. As a result, the unbiased estimators G_{ST_NC} and θ_{WC_F} may outperform the EB estimator in high-throughput data because of the effect of shrinkage. Still, in the simulation using 10 000 loci, the RMSE of the EBF_{ST} estimator was half that of the other estimators for a high gene flow scenario where $F_{ST} = 0.001$ and comparable for scenarios with lower levels of gene flow.

Population structure of Atlantic herring

The means ± SDs of the pairwise F_{ST} estimates were 0.01427 ± 0.00383 for G_{ST}, 0.00619 ± 0.00357 for G_{ST_NC}, 0.01185 ± 0.00710 for θ_{WC}, 0.00595 ± 0.00356 for θ_{WC_F} and 0.00482 ± 0.00050 for EBF_{ST} (Fig. 2a). The mean G_{ST} (without bias correction) was 2.3 times larger than G_{ST_NC}, 1.2 times for θ_{WC}, 2.4 times for θ_{WC_F} and 3.0 times for EBF_{ST}. The mean EBF_{ST} decreased to 41% of that of θ_{WC}, 78% of that of G_{ST_NC} and 81% of that of θ_{WC_F}. The SDs for G_{ST}, G_{ST_NC} and θ_{WC_F} were ~0.004 and that for θ_{WC} was ~0.007, whereas that for EBF_{ST} was an order of magnitude smaller (0.0005). Interestingly, the new differentiation estimators showed very similar values to F_{ST} estimators. G_{ST_H} values had a similar distribution to θ_{WC}, D_J was very close to θ_{WC_F}, and EBD_J was close to EBF_{ST} (Fig. 2a). Estimates of G_{ST_NC} were highly correlated with those of G_{ST_H} ($r = 0.9999$), θ_{WC_F} ($r = 0.9989$) and D_J ($r = 0.9997$) (Fig. 2b). The EBF_{ST} values decreased, but the correlations were quite strong with G_{ST_NC} ($r = 0.9550$) and θ_{WC_F} ($r = 0.9541$). The significance was very high for all combinations ($P < 2.2e^{-16}$). The EBF_{ST} values were also strongly correlated with other EB estimators ($r > 0.99$). When we fitted a linear model of $y = \alpha x$, the proportion estimates were $\hat{\alpha} = 1.88$ for EBG_{ST_H} and $\hat{\alpha} = 0.89$ for EBD_J ($R^2 = 1$, $P < 2.2e^{-16}$).

All estimators consistently described four large clusters, where the Baltic Sea (green) was associated with the Baltic–North Sea transition area (blue), and the North Sea (magenta)/British Isles (red) was associated with the North Atlantic (orange). An exception was G_{ST}, which localized the Baltic Sea apart from the other three clusters (Fig. 3a, b). All estimators identified a subcluster of SHL and CNS in the North Sea/British Isles except θ_{WC_F}. Interestingly, G_{ST_NC}, D_J and G_{ST_H} described the same population structure, although the differentiation for G_{ST_H} was approximately twice as large (Fig. 3b, c). θ_{WC} described the same population structure, but with slight differences in the Baltic Sea. θ_{WC_F} showed a similar pattern, with the difference that WIR associated with SHL and CNS. As for the EB estimators, EBF_{ST} and EBD_J provided the same population structure. EBG_{ST_H} also showed almost the same pattern, but with a slight difference in the Baltic–North Sea transition area, which was consistent with G_{ST_NC}, G_{ST_H}, θ_{WC} and D_J.

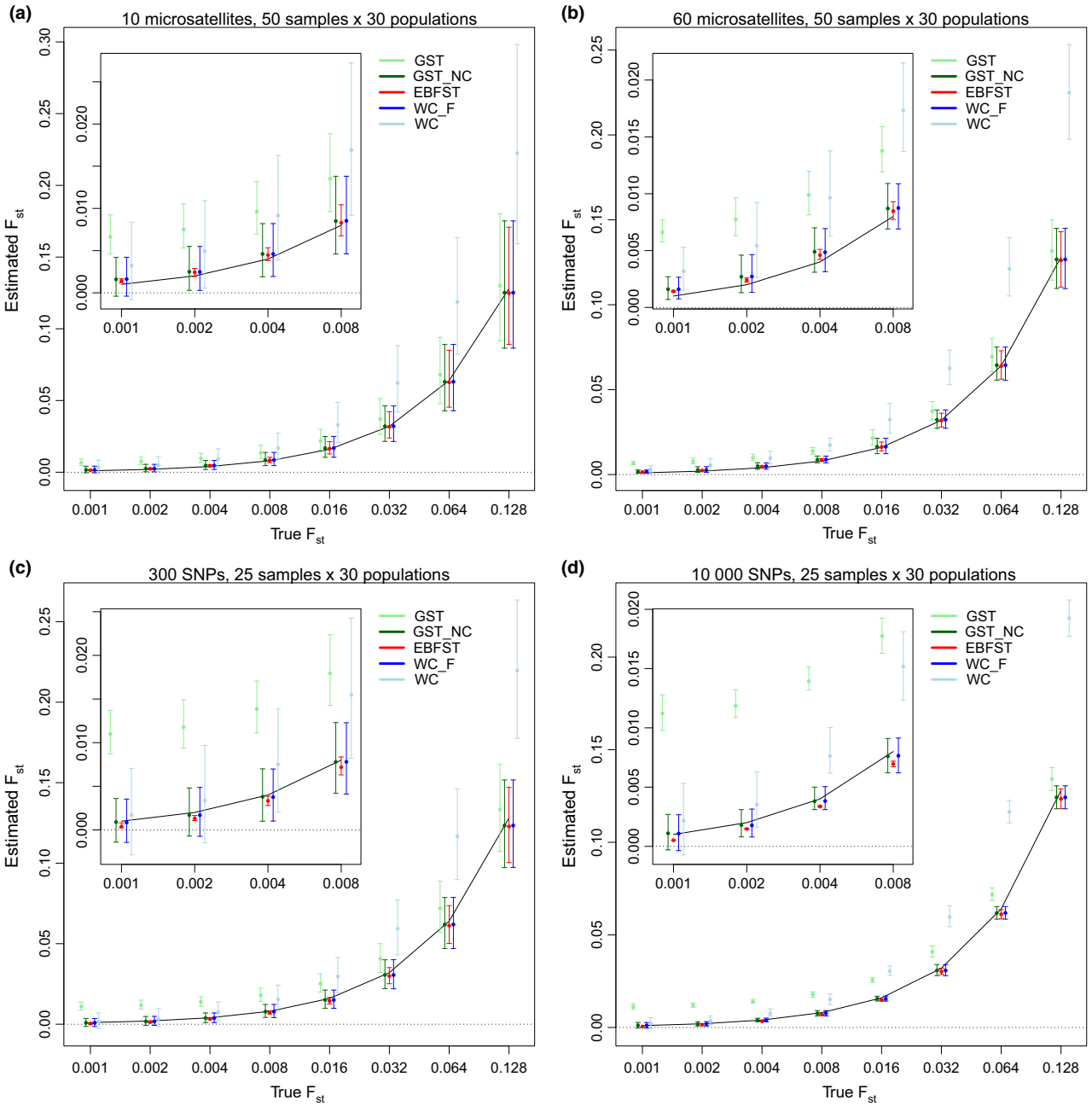

Fig. 1 Performance of F_{ST} estimators in estimating pairwise F_{ST}. Results of the coalescent simulations at various levels of F_{ST} (0.001–0.128) based on (a) 10 and (b) 60 microsatellite loci, and (c) 300 and d) 10 000 SNPs. Filled circles represent medians with 95% confidence intervals. The solid lines indicate values for true F_{ST}, and dotted lines for $F_{ST} = 0$. We generated genotypes under the Wright island model using coalescent simulations of 30 populations sampled with a sample size of 50 (for microsatellites) and 25 (for SNPs) individuals in each population, and estimated F_{ST} between $30 \times 29/2 = 435$ pairs of populations based on G_{ST}, G_{ST_NC}, EBF_{ST}, θ_{WC_F} and θ_{WC}. The procedure was repeated 10 times (see text).

Effects of environmental factors on genetic differentiation

The best fit model for both EBF_{ST} and θ_{WC_F} included geographical distance, salinity and their interaction (Model 8) (Tables 1 and S3, Supporting information). TIC was slightly smaller when using annual mean sea surface temperature and salinity. The model fitting was much better in θ_{WC_F} (TIC = 314.48, $R^2 = 0.61$) than in EBF_{ST} (TIC = 329.36, $R^2 = 0.56$), showing that 56% of the EBF_{ST} fine-scale population structure (Fig. 3b) was explained by sea surface temperature and salinity. The regression

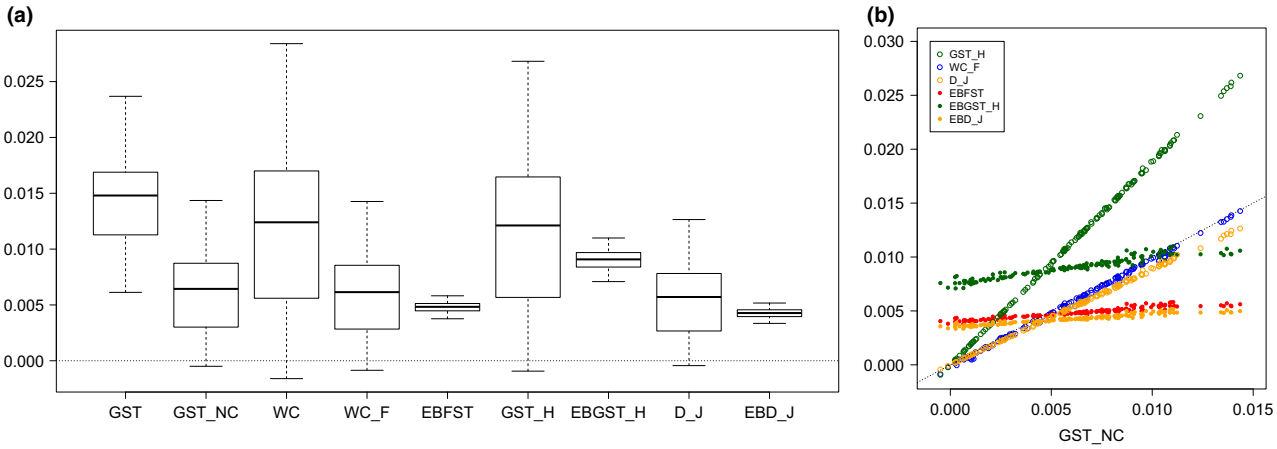

Fig. 2 Pairwise values of population differentiation of existing estimators for Atlantic herring inferred from the 281 SNP genotypes (Limborg *et al.* 2012a,b). (a) Distribution of pairwise estimates for G_{ST}, G_{ST_NC}, θ_{WC}, θ_{WC_F}, EBF_{ST}, G_{ST_H}, EBG_{ST_H}, D_J and EBD_J. (b) G_{ST_NC} values versus values of θ_{WC_F}, EBF_{ST}, G_{ST_H}, EBG_{ST_H}, D_J and EBD_J (see text). The dotted lines in the panels show (a) $F_{ST} = 0$ and (b) $y = x$.

coefficients were consistent in both θ_{WC_F} and EBF_{ST}. Those for geographical distance were positive and highly significant, and those for salinity were also positive and significant, while interaction between geographical distance and salinity was negative and not significant.

Discussion

Our coalescent simulations revealed the need for the finite sample correction of θ_{WC} when assessing population structure using pairwise F_{ST} values, and demonstrated that the EBF_{ST} estimator performed the best with respect to bias and precision in high gene flow scenarios ($F_{ST} \leq 0.032$) when highly polymorphic markers, such as microsatellites, were used. For SNPs, the EBF_{ST} estimator had greater negative bias than G_{ST_NC} and θ_{WC_F}, but the precision was the highest in all cases when 300 SNPs were used. However, when using 10 000 SNPs, the precision became better for G_{ST_NC} and θ_{WC_F} under scenarios where $F_{ST} \geq 0.004$. The EBF_{ST} estimate always takes positive values based on G_{ST}, while other estimators and their lower 95% confidence limits can take negative values when the true F_{ST} is very small. The empirical data analyses of the Atlantic herring SNPs demonstrated that the EBF_{ST} estimator identified fine-scale population structure and that 56% of the genetic differentiation was explained by geographical distance and sea surface salinity. The new EB estimators, EBD_J and EBG_{ST_H}, identified the same and very similar population structures compared with that from the EBF_{ST} estimator.

Bias-corrected G_{ST_NC} and θ_{WC_F} performed very similarly at all levels of F_{ST}. In contrast, θ_{WC} provided pairwise F_{ST} estimates ~2 times greater than those of G_{ST_NC} and θ_{WC_F}. Originally, both G_{ST} and θ_{WC} were developed to estimate F_{ST} (global F_{ST}) in a metapopulation based on a set of randomly selected population samples. The major difference in the two estimators G_{ST_NC} and θ_{WC} is the bias correction under the fixed- and random-effect models of population sampling (Weir 1996). When estimating pairwise F_{ST}, the number of populations is two ($r = 2$ in θ_{WC}), which yields the correction term $(r-1)$ of θ_{WC} in estimating the variance of allele frequencies over populations (s^2 on p. 1360 in the paper by Weir & Cockerham 1984) as one. This should provide a between-population variance that is twice G_{ST_NC}, which uses r instead of $r - 1$. Another difference is that θ_{WC} considers the variance component ($c = \bar{h}/2$) for the third-stage sampling of gametes in the denominator of the total variance ($a + b + c$) (see Eqs. (2)–(4) on p. 1359–1360 of the paper by Weir & Cockerham 1984). When all sample sizes (n) of individuals are equal ($n_i = n$), $a + b = s^2/r + \tilde{p}(1-\tilde{p}) - \bar{h}/2$. Therefore, the sum of their correction term regarding $\bar{h}$ in the denominator $a + b + c$ becomes 0. As for the numerator, $a = s^2(nr-1)/\{r(n-1)\} - \tilde{p}(1-\tilde{p})/(n-1) + \bar{h}/\{4(n-1)\}$. The term $\bar{h}$ is the average heterozygote frequency, and $\bar{h} = \sum_{i=1}^{r} \bar{h}_i$ for $n_i = n$. Therefore, the correction term $\bar{h}/\{4(n-1)\}$ in the numerator should take small values, and the effect of the third-stage sampling variance component could be negligible when sample sizes (n individuals) are large enough.

In contrast, our EBF_{ST} estimator uses the original G_{ST} formula, and the bias is not corrected explicitly. However, the EBF_{ST} estimator accounts for sampling variances of populations (first-stage sampling) and individuals (second-stage sampling) by generating the posterior distributions of allele frequencies given the observed allele counts in sampled populations. An EB estimator of a population mean and/or rate that incorporates variance component structures is useful for small-area estimation (Ghosh & Lahiri 1987). The idea is to

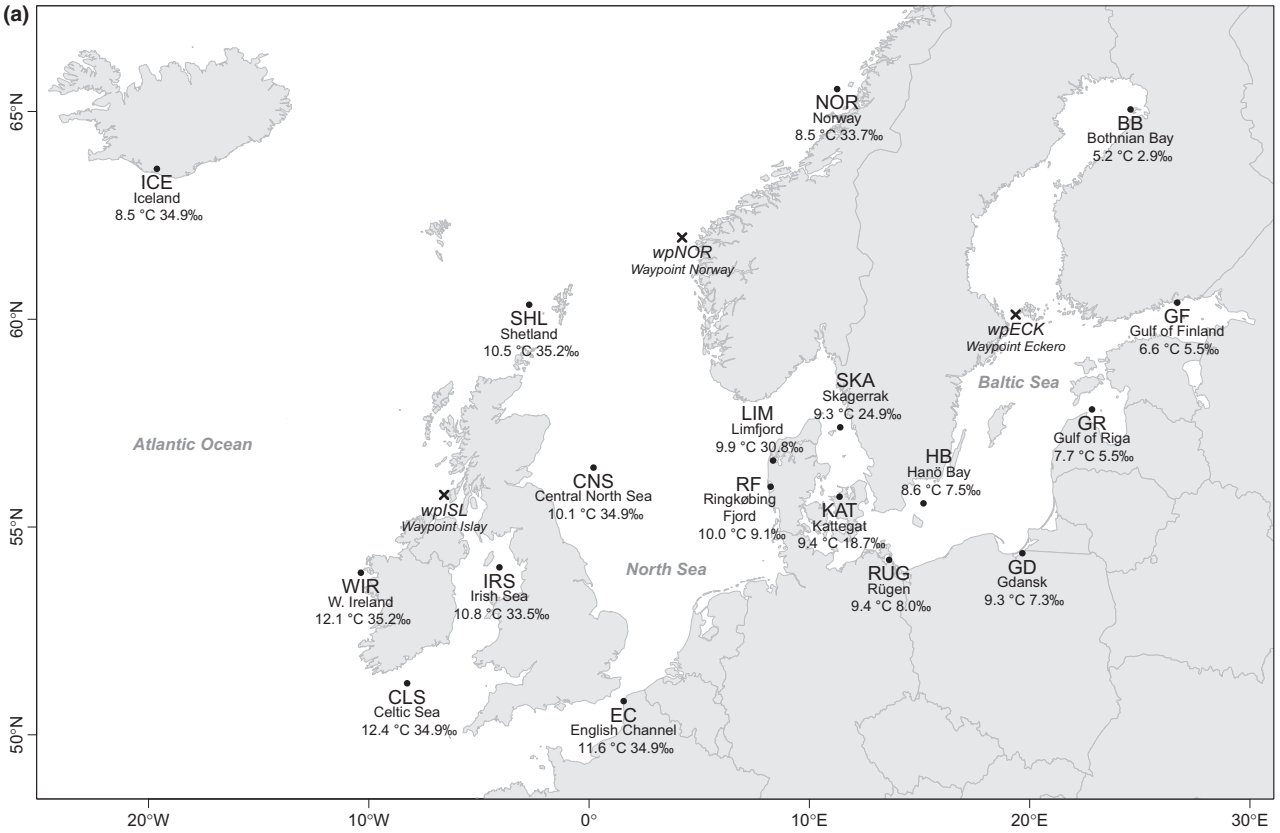

Fig. 3. Atlantic herring population structure inferred from the 281 SNP genotypes (Limborg *et al*. 2012a,b). (a) Sampling locations and obligatory way points (see text) with annual average sea surface temperatures (°C) and salinities (‰). Population structure based on (b) pairwise F_{ST} estimators of G_{ST}, G_{ST_NC}, θ_{WC_F}, θ_{WC}, and EBF_{ST}, and (c) new G_{ST}-related estimators of G_{ST_H}, D_J, EBG_{ST_H} and EBD_J. The broken lines in panels indicate $F_{ST} = 0$. Colours refer to the four genetically distinct groups: green: Baltic Sea, blue: Baltic–North Sea transition area, red: North Sea/British Isles and orange: North Atlantic.

Table 1 Estimated regression coefficients of the best fit models for EBF_{ST} and θ_{WC_F} based on the bootstrap samples generated from 281 SNPs in the work of Limborg *et al.* (2012a)

Model		EBF_{ST}				θ_{WC_F}			
No.	Variable	Estimate[a]	SD[b]	Z (=a/b)	P	Estimate[a]	SD[b]	Z (=a/b)	P
8		R^2=0.5564				R^2=0.6064			
	D	0.5153	0.1714	3.0060	**0.0003**	0.5532	0.1682	3.2895	**0.0010**
	S	0.3505	0.1785	1.9770	**0.0480**	0.3462	0.1658	2.0876	**0.0368**
	$D \times S$	−0.1244	0.0695	−1.7887	0.0737	−0.1155	0.0751	−1.5376	0.1242

[a]Based on F_{ST} estimates and explanatory variables.
[b]Obtained from bootstrapping. D, geographical distance (shortest ocean distance); S, mean annual sea surface salinity. Bold values indicate significance.

'borrow strength' from related areas to find more accurate estimates for a given area or, simultaneously, for several areas. The posterior distributions of allele frequencies generated in our EB estimation procedure gain strength from the set of sampled populations and shrink towards the true allele frequencies in a metapopulation. The EBF_{ST} can therefore be interpreted as a shrinkage estimator (Stein 1956). The results of our coalescent simulations suggest that the shrinkage is effective for highly polymorphic markers to correct estimates of allele frequencies even under small sample sizes, but not for SNPs because the allele frequencies of two alleles might be more precisely estimated than microsatellites given the sample size. The EB estimators, EBF_{ST}, EBG_{ST_H} and EBD_J, consistently identified the Atlantic herring population structure, which consisted of four large groups: (i)

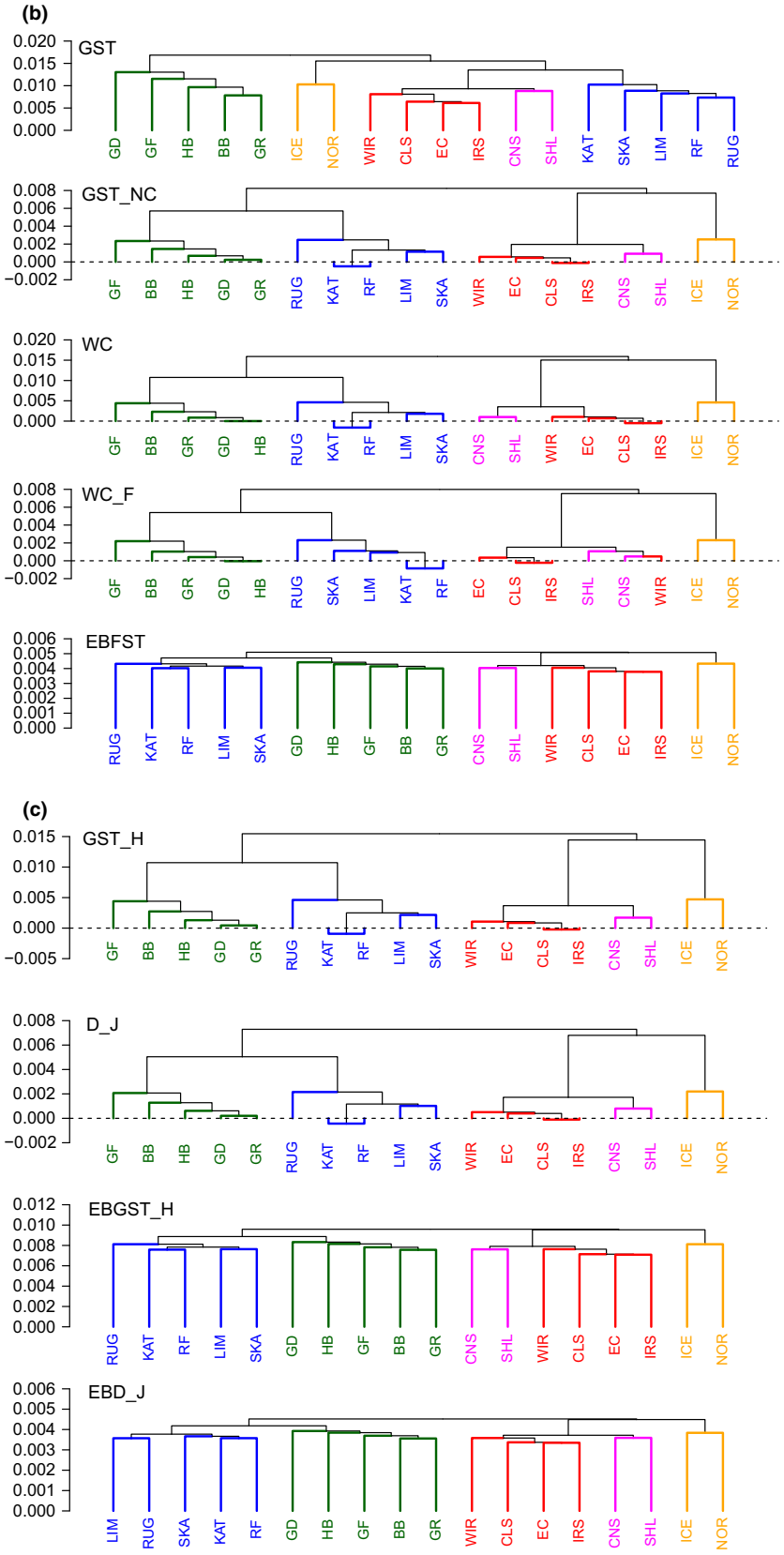

Fig. 3. Continued

Baltic Sea; (ii) Baltic–North Sea transition area; (iii) North Sea/British Isles; and (iv) North Atlantic (Fig. 3). The population structure generally agrees with that inferred by the original study (Limborg *et al.* 2012a) and that obtained from the top 156 loci among the 281 loci ranked by their contribution to divergence of the four large clusters (Bekkevold *et al.* 2015). Our EB estimators provided a finer scale population structure without any prior information.

As for the new differentiation estimators, G_{ST_H} performed similarly to θ_{WC}, and D_J performed similarly to G_{ST_NC} and θ_{WC_F} for the Atlantic herring SNPs. The observed heterozygosity (H_o) was 0.31 ± 0.01 and was very similar in all samples [Table 1 in the paper by Limborg *et al.* (2012a)]. G_{ST_H} (Hedrick 2005) and D_J (Jost 2008) were developed for cases in which heterozygosity is high within each subpopulation but the subpopulations have significantly differentiated. In such cases, G_{ST} takes small values even though the actual differentiation is large, especially for cases with highly polymorphic markers such as microsatellite loci. Our analysis of the high gene flow Atlantic herring using SNPs might not be an appropriate example to test the characteristics of the new differentiation estimators. However, our results of coalescent simulations should be straightforward because they are functions of G_{ST}. There has been extensive discussion on the new differentiation measures (Heller & Siegismund 2009; Ryman & Leimar 2009; Gerlach *et al.* 2010; Leng & Zhang 2011; Whitlock 2011; Wang 2015). However, further study is needed for various levels of heterozygosity and genetic differentiation to comprehensively evaluate the performance of the new differentiation estimators including EBG_{ST_H}, EBD_J and EBF_{ST}.

The R package FINEPOP 1.3.0 implements all estimators used in this study. It can be applied to genotype/haplotype data derived from common markers, including isozymes, mitochondrial DNA, microsatellites and SNPs. Accepted data formats include GENEPOP and a frequency format for allele and haplotype frequencies in text files. The function `read.genepop` or `read.frequency` loads the data file and the population label file. `EBFST` calculates EBF_{ST} values and outputs the pairwise F_{ST} matrix. `GstN`, `GstNC`, and `thetaWC.pair` calculate pairwise F_{ST} values for G_{ST} (Nei 1973), G_{ST_NC} (Nei & Chesser 1983) and θ_{WC_F} (θ_{WC} of Weir & Cockerham (1984) with finite sample correction), respectively. `GstH` and `DJ` calculate the new differentiation measures; pairwise G'_{ST} (Hedrick 2005) and D_J (Jost 2008) values are based on the unbiased estimators of H_T and H_S (Nei & Chesser 1983). `EBGstH` and `EBDJ` calculate the EB estimates for them. R script used in the herring case studies is provided to exemplify usages of FINEPOP functions (Appendix S3, Supporting Information). The function of regression analysis of genetic population structure on environmental factors will also be included in the coming version.

Acknowledgements

We thank the editor, Michael M. Hansen and anonymous reviewers for their constructive comments, which improved our manuscript significantly. We appreciate Morten T. Limborg and his colleagues, whose analysis supported by intensive survey (Limborg et al. 2012a,b) motivated our TIC-based variable selection. This study was supported by JSPS Grant-in-Aid for Scientific Research (B) awards 22380110 to SK, and 25280006 and 16H02788 to HK.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (eds Petrov B. N., Csaki F.), pp. 267–281. Akademiai Kiado, Budapest.

Bekkevold D, Andre C, Dahlgren TG *et al.* (2005) Environmental correlates of population differentiation in Atlantic herring. *Evolution*, **59**, 2656–2668.

Bekkevold D, Helyar SJ, Limborg MT *et al.* (2015) Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science*, **72**, 1790–1801.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.

DeWoody JA, Avise JC (2000) Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *Journal of Fish Biology*, **56**, 461–473.

Gaggiotti OE, Bekkevold D, Jørgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, **63**, 2939–2951.

Gerlach G, Jueterbock A, Kraemer P, Deppermann J, Harmand P (2010) Calculations of population differentiation based on G_{ST} and D: forget G_{ST} but not all of statistics!. *Molecular Ecology*, **19**, 3845–3852.

Ghosh M, Lahiri P (1987) Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, **82**, 1153–1162.

Guillot G, Rousset F (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution*, **4**, 336–344.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Heller R, Siegismund HR (2009) Relationship between three measures of genetic differentiation G_{ST}, D_{EST} and G'_{ST}: How wrong have we been? *Molecular Ecology*, **18**, 2080–2083.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jost LOU (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Kish L (1965) *Survey Sampling*. John Wiley and Sons, New York.

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

Kitada S, Kitakado T, Kishino H (2007) Empirical Bayes inference of pairwise F_{ST} and its distribution in the genome. *Genetics*, **177**, 861–873.

Leng L, Zhang DX (2011) Measuring population differentiation using G_{ST} or D? A simulation study with microsatellite DNA markers under a finite island model and nonequilibrium conditions. *Molecular Ecology*, **20**, 2494–2509.

Limborg MT, Helyar SJ, de Bruyn M *et al.* (2012a) Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, **21**, 3686–3703.

Limborg MT, Helyar SJ, de Bruyn M *et al*. (2012b) Data from: Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Dryad Digital Repository*, http://dx.doi.org/10.5061/dryad.2n763
Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Science of the United States of America*, **70**, 3321–3323.
Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversity. *Annals of Human Genetics*, **47**, 253–259.
Palsbøll PJ, Bérubé M, Allendorf FW (2006) Identification of management units using population genetic data. *Trends in Ecology and Evolution*, **22**, 11–16.
Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J (1997) Microsatellite variation and the differentiation of modern humans. *Human Genetics*, **99**, 1–7.
Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.
Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, **142**, 1357–1362.
Rousset F (2008) GENEPOP'007: a complete reimplementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.
Ryman N, Leimar O (2009) G_{ST} is still a useful measure of genetic differentiation—a comment on Jost's *D*. *Molecular Ecology*, **18**, 2084–2087.
Selkoe KA, Henzler CM, Gaines SD (2008) Seascape genetics and the spatial ecology of marine populations. *Fish and Fisheries*, **9**, 363–377.
Skinner CJ, Holt D, Smith TF (1989) *Analysis of Complex Surveys*. John Wiley and Sons, New York.
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*, vol. **1**, pp. 197–206. University of California Press, Berkeley.
Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? *Molecular Ecology*, **19**, 3496–3514.
Sun JX, Helgason A, Masson G *et al.* (2012) A direct characterization of human mutation based on microsatellites. *Nature Genetics*, **44**, 1161–1165.
Takeuchi K (1976) Distribution of information statistics and criteria for adequacy of Models. *Mathematical Science*, **153**, 12–18 (in Japanese).
Wang J (2015) Does G_{ST} underestimate genetic differentiation from marker data? *Molecular Ecology*, **24**, 3546–3558.
Waples RS (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, **89**, 438–450.
Weir BS (1996) *Genetic Data Analysis*. Sinauer, Sunderland.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Weir BS, Hill WG (2002) Estimating *F*-statistics. *Annual Review of Genetics*, **36**, 721–750.
Whitlock MC (2011) G'_{ST} and *D* do not replace F_{ST}. *Molecular Ecology*, **20**, 1083–1091.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.
Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution*, **19**, 395–420.

Appendix 1: Relationship between G_{ST} and F_{ST}

Nei's G_{ST} formula (Nei 1973) defines F_{ST} between populations as:

$$G_{ST} = \frac{H_T - H_S}{H_T} \qquad \text{(eqn 1)}$$

where H_T and H_S are the total-population and within-population heterozygosity values at a locus. These statistics are defined as:

$$H_S = 1 - \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{m} p_j^{(i)^2}$$

$$H_T = 1 - \sum_{i=1}^{m}\left(\frac{1}{r}\sum_{i=1}^{r} p_j^{(i)}\right)^2,$$

where m is the number of alleles and $p_j^{(i)}$ is the frequency of allele j in population i ($i = 1, \ldots, r$). Nei (1973) defined Wright's F-statistics as $F_{IS} = 1 - H_0/H_S$, $F_{IT} = 1 - H_0/H_T$ and $F_{ST} = 1 - H_S/H_T$, where H_0 is the frequency of all heterozygotes. These equations satisfy Wright's definition: $1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$ (Nei & Chesser 1983).

F_{ST} is defined as 'the correlation between random gametes, drawn from the same subpopulation, relative to the total' (Wright 1951; p. 328). F_{ST} is also defined as the ratio of the between-population variance to the total variance of allele frequencies (e.g. Weir & Cockerham 1984). Here, we consider cases with multiple alleles as

$$F_{ST} = \frac{\sigma_p^2}{\sum_{j=1}^{m} p_j(1 - p_j)} \qquad \text{(eqn 2)}$$

where p_j is the mean allele frequency of allele j, and σ_p^2 is the variance of allele frequencies over subpopulations.

We explicitly show the relationship between G_{ST} and F_{ST} between two populations ($r = 2$) with multiple alleles. The numerator of G_{ST} can be written as

$$\begin{aligned} H_T - H_S &= \frac{1}{2}\sum_{j=1}^{m}\left(p_j^{(1)^2} + p_j^{(2)^2}\right) - \frac{1}{4}\sum_{j=1}^{m}\left(p_j^{(1)} + p_j^{(2)}\right)^2 \\ &= \frac{1}{2}\sum_{j=1}^{m}\left[\left(p_j^{(1)} - p_j^{(2)}\right)^2 + 2p_j^{(1)}p_j^{(2)}\right] \\ &\quad - \frac{1}{4}\sum_{j=1}^{m}\left[\left(p_j^{(1)} - p_j^{(2)}\right)^2 + 4p_j^{(1)}p_j^{(2)}\right] \\ &= \frac{1}{4}\sum_{j=1}^{m}\left(p_j^{(1)} - p_j^{(2)}\right)^2. \end{aligned}$$

H_T is

$$\begin{aligned} H_T &= 1 - \frac{1}{4}\sum_{j=1}^{m}\left(p_j^{(1)} + p_j^{(2)}\right)^2 \\ &= 1 - \sum_{j=1}^{m}\left(\frac{p_j^{(1)} + p_j^{(2)}}{2}\right)^2 \\ &= 1 - \sum_{j=1}^{m}\bar{p}_j^2. \end{aligned}$$

Therefore, we have

$$G_{ST} = \frac{\frac{1}{4}\sum_{j=1}^{m}(p_j^{(1)} - p_j^{(2)})^2}{1 - \sum_{j=1}^{m}\bar{p}_j^2}.$$

The denominator of G_{ST} can be expanded as

$$\begin{aligned} 1 - \sum_{j=1}^{m}\bar{p}_j^2 &= 1 - \left(\sum_{j=1}^{m-1}\bar{p}_j^2 + (1 - \sum_{j=1}^{m-1}p_j)^2\right) \\ &= 2\left[\sum_{j=1}^{m-1}p_j(1-p_j) - \sum_{j<j'}^{m-1}p_j p_{j'}\right]. \end{aligned}$$

The second term is as follows:

$$-\sum_{j<j'}^{m-1}p_j p_{j'} = p_m(1-p_m) - \sum_{j=1}^{m-1}p_j(1-p_j),$$

and we have

$$\begin{aligned} &2\left[\sum_{j=1}^{m-1}p_j(1-p_j) - \sum_{j<j'}^{m-1}p_j p_{j'}\right] \\ &= 2\sum_{j=1}^{m-1}p_j(1-p_j) - \sum_{j=1}^{m-1}p_j(1-p_j) + p_m(1-p_m) \\ &= \sum_{j=1}^{m}p_j(1-p_j). \end{aligned}$$

Thus, the denominator of G_{ST} equals that of Eq. (2):

$$1 - \sum_{j=1}^{m}\bar{p}_j^2 = \sum_{j=1}^{m}p_j(1-p_j).$$

In general, the variance of observed random variables x ($x_1, \ldots, x_n$) is expressed in Eq. (3). Therefore, the numerator of F_{ST} between two populations is expressed as

$$\frac{1}{4}\sum_{j=1}^{m}(p_j^{(1)} - p_j^{(2)})^2.$$

Thus, it was confirmed that G_{ST} is equivalent to F_{ST} (Eq. (1) = Eq. (2)).

For biallelic cases (m=2), $H_T = 2p(1-p)$, and $H_T - H_S = (p^{(1)} - p^{(2)})^2/2$. Then, we have Wright's F_{ST} definition (Wright 1951, 1965) of

$$F_{ST} = \frac{V(p)}{\bar{p}(1-\bar{p})}$$

where v and $V(p)$ are the mean allele frequency and the variance of allele frequencies over subpopulations, respectively.

$$\begin{aligned} V[x] &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}(x_i - x_j)\right)^2 \\ &= \frac{1}{(n-1)n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(x_i - x_j)(x_i - x_k) \\ &= \frac{1}{(n-1)n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(x_i^2 - x_i x_k - x_i x_j + x_j x_k) \\ &= \frac{1}{(n-1)n^2}\left(n^2\sum_{i=1}^{n}x_i^2 - n\left(\sum_{i=1}^{n}x_i^2 + \sum_{i\neq k}x_i x_k\right)\right. \\ &\quad \left. -n\left(\sum_{i=1}^{n}x_i^2 + \sum_{i\neq j}x_i x_j\right) + n\left(\sum_{i=1}^{n}x_i^2 + \sum_{j\neq k}x_j x_k\right)\right) \\ &= \frac{1}{(n-1)n}\left((n-1)\sum_{i=1}^{n}x_i^2 - \sum_{i\neq j}x_i x_j\right) \\ &= \frac{1}{2(n-1)n}\sum_{i\neq j}(x_i - x_j)^2 \end{aligned} \quad \text{(eqn 3)}$$

Appendix 2: Relative bias and mean square error of the numerator of the F_{ST} estimator

We hereafter express the G_{ST} numerator (Eq. 1) as $F_{ST}^{Num} = H_T - H_S$. After Taylor series expansion, we obtain the F_{ST} estimator around the true value:

$$\begin{aligned} \hat{F}_{ST} &= \frac{\hat{F}_{ST}^{Num}}{\hat{H}_T} \\ &\approx \frac{F_{ST}^{Num}}{H_T} + \frac{1}{H_T}\left(\hat{F}_{ST}^{Num} - F_{ST}^{Num}\right) - \frac{F_{ST}^{Num}}{H_T^2}\left(\hat{H}_T - H_T\right). \end{aligned}$$

Therefore,

$$\frac{\hat{F}_{ST} - F_{ST}}{F_{ST}} \approx \frac{\hat{F}_{ST}^{Num} - F_{ST}^{Num}}{F_{ST}^{Num}} - \frac{\hat{H}_T - H_T}{H_T} \quad \text{(eqn4)}$$

Eqaution (4) shows that the $\hat{F}_{ST}$ relative bias is determined by the relative bias of $\hat{F}_{ST}^{Num}$ and that of $\hat{H}_T$. In the

case of high gene flow, that is, when F_{ST}^{Num} is small, the $\hat{F}_{ST}^{Num}$ relative bias becomes large. Using Eq. (4), the relative mean square error of $\hat{F}_{ST}$ is decomposed as

$$E\left[\left(\frac{\hat{F}_{ST}-F_{ST}}{F_{ST}}\right)^2\right] \approx E\left[\left(\frac{\hat{F}_{ST}^{Num}-F_{ST}^{Num}}{F_{ST}^{Num}}\right)^2\right] + E\left[\left(\frac{\hat{H}_T-H_T}{H_T}\right)^2\right] - 2E\left[\left(\frac{\hat{F}_{ST}^{Num}-F_{ST}^{Num}}{F_{ST}^{Num}}\right)\left(\frac{\hat{H}_T-H_T}{H_T}\right)\right].$$

The relative mean square error of $\hat{F}_{ST}$ becomes large for higher gene flow.

The F_{ST} estimator numerator F_{ST}^{Num} is expressed as the sum of the square difference in the allele frequencies between the two populations:

$$F_{ST}^{Num} = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{m} p_j^{(i)^2} - \frac{1}{2}\sum_{j=1}^{m} p_j^{(1)}p_j^{(2)} = \frac{1}{4}\sum_{j=1}^{m}\left(p_j^{(1)}-p_j^{(2)}\right)^2.$$

The bias of the estimator $\hat{F}_{ST}^{Num} = \hat{H}_T - \hat{H}_S$ is explicitly given below and is the sum of the variance of the difference in allele frequencies:

$$\begin{aligned} &E\left[\hat{H}_T-\hat{H}_S\right] - (H_T - H_S) \\ &= \frac{1}{4}\sum_{j=1}^{m}\left(E\left[\left(\hat{p}_j^{(1)}-\hat{p}_j^{(2)}\right)^2\right] - \left(p_j^{(1)}-p_j^{(2)}\right)^2\right) \\ &= \frac{1}{4}\sum_{j=1}^{m} V\left[\hat{p}_j^{(1)}-\hat{p}_j^{(2)}\right]. \end{aligned}$$

The $\hat{F}_{ST}^{Num}$ bias becomes larger for a larger number of alleles m. Because $\hat{F}_{ST}^{Num}$ is small in the case of high gene flow, the $\hat{F}_{ST}^{Num}$ relative bias becomes large.

We begin with a two-allele case to derive the explicit formula for the relative square error:

$$\begin{aligned} F_{ST}^{Num}\left(p^{(1)}p^{(2)}\right) &= \frac{1}{4}\left\{\left(p^{(1)}-p^{(2)}\right)^2 + \left(\left(1-p^{(1)}\right)-\left(1-p^{(2)}\right)\right)^2\right\} \\ &= \frac{1}{2}\left(p^{(1)}-p^{(2)}\right)^2. \end{aligned}$$

After Taylor series expansion, we obtain $\hat{F}_{ST}^{Num} = F_{ST}^{Num}\left(\hat{p}^{(1)},\hat{p}^{(2)}\right)$ around the true value:

$$\begin{aligned} &F_{ST}^{Num}\left(\hat{p}^{(1)},\hat{p}^{(2)}\right) - F_{ST}^{Num}\left(p^{(1)},p^{(2)}\right) \\ &\approx \left(p^{(1)}-p^{(2)}\right)\left\{\left(\hat{p}^{(1)}-p^{(1)}\right)-\left(\hat{p}^{(2)}-p^{(2)}\right)\right\}. \end{aligned}$$

Therefore, the relative square error is calculated as:

$$\begin{aligned} &E\left[\left(\frac{F_{ST}^{Num}\left(\hat{p}^{(1)},\hat{p}^{(2)}\right)-F_{ST}^{Num}\left(p^{(1)},p^{(2)}\right)}{F_{ST}^{Num}\left(p^{(1)},p^{(2)}\right)}\right)^2\right] \\ &\approx \frac{\left(p^{(1)}-p^{(2)}\right)^2\left(V\left[\hat{p}^{(1)}\right]+V\left[\hat{p}^{(2)}\right]\right)}{F_{ST}^{Num}\left(p^{(1)},p^{(2)}\right)^2} \\ &= \frac{2}{F_{ST}^{Num}\left(p^{(1)},p^{(2)}\right)}\times\left(\frac{p^{(1)}\left(1-p^{(1)}\right)}{n^{(1)}}+\frac{p^{(2)}\left(1-p^{(2)}\right)}{n^{(2)}}\right). \end{aligned}$$

This equation shows that the relative variation in $\hat{F}_{ST}^{Num}$ becomes larger for higher gene flow.

In general cases with m alleles, the Taylor series expansion is as follows:

$$\begin{aligned} &F_{ST}^{Num}\left(\hat{\mathbf{p}}^{(1)},\hat{\mathbf{p}}^{(2)}\right) - F_{ST}^{Num}\left(\mathbf{p}^{(1)},\mathbf{p}^{(2)}\right) \\ &\approx \frac{1}{2}\sum_{j=1}^{m}\left(p_j^{(1)}-p_j^{(2)}\right)\left\{\left(\hat{p}_j^{(1)}-p_j^{(1)}\right)-\left(\hat{p}_j^{(2)}-p_j^{(2)}\right)\right\}. \end{aligned}$$

The mean square error of the $\hat{F}_{ST}$ denominator is as follows:

$$\begin{aligned} &E\left[\left(F_{ST}^{Num}\left(\hat{\mathbf{p}}^{(1)},\hat{\mathbf{p}}^{(2)}\right)-F_{ST}^{Num}\left(\mathbf{p}^{(1)},\mathbf{p}^{(2)}\right)\right)^2\right] \\ &\approx \frac{1}{4}\sum_{i=1}^{m}\sum_{j'=1}^{m}\left(p_j^{(1)}-p_j^{(2)}\right)\left(p_{j'}^{(1)}-p_{j'}^{(2)}\right) \\ &\left(Cov\left[\hat{p}_j^{(1)},\hat{p}_{j'}^{(1)}\right]+Cov\left[\hat{p}_j^{(2)},\hat{p}_{j'}^{(2)}\right]\right) \\ &= \frac{1}{4}\sum_{j=1}^{m}\sum_{j'=1}^{m}\left(p_j^{(1)}-p_j^{(2)}\right)\left(p_{j'}^{(1)}-p_{j'}^{(2)}\right) \\ &\left(\frac{\delta_{jj'}p_j^{(1)}-p_j^{(1)}p_{j'}^{(1)}}{n^{(1)}}+\frac{\delta_{jj'}p_j^{(2)}-p_j^{(2)}p_{j'}^{(2)}}{n^{(2)}}\right) \\ &= \frac{1}{4}\sum_{j=1}^{m}\left(\frac{p_j^{(1)}}{n^{(1)}}+\frac{p_j^{(2)}}{n^{(2)}}\right)\left(p_j^{(1)}-p_j^{(2)}\right)^2 \\ &-\frac{1}{4}\left(\sum_{j=1}^{m}\left(\frac{p_j^{(1)}}{n^{(1)}}+\frac{p_j^{(2)}}{n^{(2)}}\right)\left(p_j^{(1)}-p_j^{(2)}\right)\right)^2, \end{aligned} \quad \text{(eqn 5)}$$

where $\delta_{jj'}$ is a delta function that takes 1 for $j = j'$ and 0 for $j \neq j'$.

Here, $q_j = \left(\frac{p_j^{(1)}}{n^{(1)}}+\frac{p_j^{(2)}}{n^{(2)}}\right), \bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j, \tilde{q} = \frac{q_j}{m\bar{q}}, d_j = p_j^{(1)} - p_j^{(2)}$, $\bar{d} = \sum_{j=1}^{m}\tilde{q}_j d_j$, and substituting them into Eq. (5), we have the simple form for the variance of the $\hat{F}_{ST}$ denominator:

$$E\left[\left(F_{ST}^{Num}\left(\hat{\mathbf{p}}^{(1)},\hat{\mathbf{p}}^{(2)}\right)-F_{ST}^{Num}\left(\mathbf{p}^{(1)},\mathbf{p}^{(2)}\right)\right)^2\right]$$
$$\approx\frac{1}{4}\left\{m\bar{q}\sum_{j=1}^{m}\tilde{q}_j d_j^2-\left(m\bar{q}\bar{d}\right)^2\right\}=\frac{1}{4}m\bar{q}\sum_{j=1}^{m}\tilde{q}_j\left(d_j-\bar{d}\right)^2.$$

The relative root mean square error is obtained as:

$$\sqrt{\left[\left(\frac{F_{ST}^{Num}\left(\hat{\mathbf{p}}^{(1)},\hat{\mathbf{p}}^{(2)}\right)-F_{ST}^{Num}(\mathbf{p}^{(1)},\mathbf{p}^{(2)})}{F_{ST}^{Num}(\mathbf{p}^{(1)},\mathbf{p}^{(2)})}\right)^2\right]}$$
$$\approx\frac{1}{2}\sqrt{m\bar{q}\sum_{j=1}^{m}\tilde{q}_j\left(d_j-\bar{d}\right)^2}\times\frac{1}{F_{ST}^{Num}(\mathbf{p}^{(1)},\mathbf{p}^{(2)})}$$

This equation clearly shows that the variation in $\hat{F}_{ST}^{Num}$ becomes larger for higher gene flow and a larger number of alleles.

S.K. and H.K. designed the study. All authors analysed the data, wrote the manuscript and developed FINEPOP. R.N. and H.K. wrote the R codes, and R.N. performed simulations.

Data accessibility

The R package FINEPOP, user manual and example data set are available on CRAN (https://CRAN.R-project.org/package=FinePop).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Performance of F_{ST} estimators in estimating pairwise F_{ST} at various levels of F_{ST} (0.001–0.128) based on microsatellite genotypes at 10 and 60 loci generated from coalescent simulations using Hudson's `ms`. We set the number of populations sampled to 30 with the sample size of 50 individuals in each population, and estimated F_{ST} between pairs of populations. This procedure was repeated 10 times, and mean bias (MB) and root mean squared error (RMSE) were calculated (see text).

Table S2 Performance of F_{ST} estimators in estimating pairwise F_{ST} at various levels of F_{ST} (0.001–0.128) based on SNP genotypes at 300 and 10 000 loci generated from coalescent simulations using Hudson's `ms`. We set the number of populations sampled to 30 with a sample size of 25 individuals in each population, and estimated F_{ST} between pairs of populations. The procedure was repeated 10 times, and mean bias (MB) and root mean squared error (RMSE) were calculated (see text).

Table S3 Takeuchi Information Criterion (TIC) values for the linear regression analyses of environmental variables on EBF_{ST} and θ_{WC_F} based on the bootstrap sample of the 281 SNPs in the work by Limborg *et al.* (2012a).

Appendix S1 Pairwise F_{ST} estimates from the 281 Atlantic herring SNPs used for bootstrapping

Appendix S2 Shortest waterways among sampling locations, sea surface temperatures, and salinities for the Atlantic herring.

Appendix S3 R script for the herring case study.