Population-specific F_{ST} and Pairwise F_{ST}: History and Environmental Pressure

Shuichi Kitada*, Reiichiro Nakamichi†, and Hirohisa Kishino‡

*Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan

†Japan Fisheries Research and Education Agency, Yokohama 236-8648, Japan

‡Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan

Short running title:

Population-specific and pairwise F_{ST}

Keywords:

Adaptation, evolution, migration, multi-dimensional scaling, population structure, species specificity

Corresponding:

Shuichi Kitada,

Tokyo University of Marine Science and Technology, Minato-ku, Tokyo 108-8477, Japan

+81-297-45-5267

kitada@kaiyodai.ac.jp

ABSTRACT

Appropriate estimates of population structure are the basis of population genetics, with applications varying from evolutionary and conservation biology to association mapping and forensic identification. The common procedure is to first compute Wright's F_{ST} over all samples (global F_{ST}) and then routinely estimate between-population F_{ST} values (pairwise F_{ST}). An alternative approach for estimating population differentiation is the use of population-specific F_{ST} measures. Here, we characterize population-specific F_{ST} and pairwise F_{ST} estimators by analyzing publicly available human, Atlantic cod and wild poplar data sets. The bias-corrected moment estimator of population-specific F_{ST} identified the source population and traced the migration and evolutionary history of its derived populations by way of genetic diversity, whereas the bias-corrected moment estimator of pairwise F_{ST} was found to represent current population structure. Generally, the first axis of multi-dimensional scaling for the pairwise F_{ST} distance matrix reflected population history, while subsequent axes indicated migration events, languages and the effect of environment. The relative contributions of these factors were dependent on the ecological characters of the species. Given shrinkage towards mean allele frequencies, maximum likelihood and Bayesian estimators of locus-specific global F_{ST} improved the power to detect genes under environmental selection. In contrast, bias-corrected moment estimators of global F_{ST} measured species divergence and enabled reliable interpretation of population structure. The genomic data highlight the usefulness of the bias-corrected moment estimators of F_{ST}. The R package FinePop2_ver.0.2 for computing these F_{ST} estimators is available at CRAN.

Quantifying genetic relationships among populations is of substantial interest in population biology, ecology and human genetics (Weir and Hill 2002). Appropriate estimates of population structure are the basis of population genetics, with applications varying from evolutionary and conservation studies to association mapping and forensic identification (Weir and Hill 2002; Weir and Goudet 2017). For such objectives, Wright's F_{ST} (Wright 1951) is commonly used to quantify the genetic divergence of populations (reviewed by Excoffier 2001; Rousset 2001; Balloux and Lugon-Moulin 2002; Weir and Hill 2002; Rousset 2004; Beaumont 2005; Holsinger and Weir 2009). Wright (1951) defined F_{ST} as "the correlation between randomly sampled gametes relative to the total drawn from the same subpopulation"; that is, $F_{\mathrm{ST}} = \frac{F_{\mathrm{IT}} - F_{\mathrm{IS}}}{1 - F_{\mathrm{IS}}}$, where F_{IS} and F_{IT} are the inbreeding coefficients of individuals relative to their substrains and the total. Nei (1973) proposed the G_{ST} measure as a formulation of Wright's F_{ST}, which is the ratio of between-population heterozygosity to total gene heterozygosity. Nei (1977) defined Wright's F-statistics ($F_{\mathrm{ST}}, F_{\mathrm{IS}}$ and F_{ST}) in terms of heterozygosity. Crow and Aoki (1984) introduced the concept of gene identity and defined G_{ST}, while Slatkin (1991) defined F_{ST} in terms of probabilities of identity. F_{ST} has also been defined in terms of gene identity (Rousset 2001, 2002; Excoffier 2001). G_{ST} is identical to F_{ST} (Nei 1975, 1977; Crow and Aoki 1984). The equality of G_{ST} and F_{ST} has been given for bi-allelic (Nei 1975, 1977; Crow and Aoki 1984) and multi-allelic cases for pairwise F_{ST} at a single locus (Kitada *et al.* 2017). Published definitions of G_{ST} and F_{ST}, the relationship between G_{ST} and F_{ST}, and the F_{ST} estimators used in this study are summarized in Supplemental Note.

Various F_{ST} estimators based on sampling design as well as maximum likelihood and Bayesian framework have been considered (reviewed by Weir 1996; Excoffier 2001; Rousset

2001; 2004, Balloux and Lugon-Moulin 2002; Holsinger and Weir 2009). Nei and Chesser (1983) proposed a moment estimator (hereafter, NC83) that corrects the bias of G_{ST} using unbiased estimators of the numerator (between-population heterozygosity) and denominator (total heterozygosity). NC83 takes into account a sample set of subpopulations taken from a present-day metapopulation but does not consider sample replicates. Weir and Cockerham (1984) proposed a moment estimator of F_{ST} denoted by θ (hereafter, WC84) that uses unbiased estimators of the numerator (between-population variance) and the denominator (total variance). WC84, a coancestry coefficient, considers replicates of samples taken from an ancestral population. This estimator considers observed variance components of allele and heterozygote frequencies between populations, between individuals within populations, and between gametes within an individual, based on the analysis of variance (ANOVA) framework of Cockerham (1969, 1973). The analysis of molecular variance approach (Excoffier *et al.* 1992) used in Arlequin (Excoffier and Lischer 2010) also defines F_{ST} as the proportion of the variance components. Estimators of Wright's F-statistics based on gene identity coefficients with a different weighting scheme have also been defined (Rousset 2008). These moment F_{ST} estimators, which are implemented in either Arlequin or Genepop (Raymond and Rousset 1995; Rousset 2008), produce identical or very similar values to those obtained using WC84 implemented in FSTAT (Goudet 1995).

Wright (1931, 1951) modeled the distribution of gene frequencies in island populations using a beta distribution. As an extension of Wright's approach, maximum likelihood methods for estimating F_{ST} and/or genetic differentiation apply beta-binomial (for two alleles) and/or multinomial-Dirichlet (for multiple alleles) distributions to determine the likelihood of sample allele counts (Balding and Nichols 1995; Lange 1995; Rannala and Hartigan 1996; Kitada *et*

al. 2000; Balding 2003; Kitada and Kishino 2004; Kitakado *et al.* 2006; Kitada *et al.* 2007). Bayesian approaches have also been developed to obtain posterior mean F_{ST} values and/or posterior distributions (Balding *et al.* 1996; Holsinger 1999; Lockwood *et al.* 2001; Holsinger *et al.* 2002; Corander *et al.* 2003; Steele *et al.* 2014). These methods use beta and/or Dirichlet distributions as prior distributions of allele frequencies.

All the above-mentioned F_{ST} estimators were basically developed to estimate mean F_{ST} over loci and over populations based on a set of population samples, often called global F_{ST} (*e.g.*, Pérez-Lezaun *et al.* 1997; Rousset 2004). In empirical studies, global F_{ST} is first estimated and then F_{ST} values between pairs of populations (pairwise F_{ST}) are routinely estimated as implemented in standard population genetics software programs such as Arlequin (Excoffier and Lischer 2010), FSTAT (Goudet 1995) and Genepop (Raymond and Rousset 1995; Rousset 2008). Pairwise differences are tested, similar to one-way ANOVA, when a significant difference in the means is detected. This procedure has been widely adopted to study species ranging from plants to animals (Weir and Hill 2002) and, more recently, the approach has been used in population genomics analyses of various species such as small freshwater fish (Malinsky 2015), herring (Lamichhaney *et al.* 2017), tobacco cutworm (Cheng *et al.* 2017), drosophila (Griffin *et al.* 2017) and humans (Bhatia *et al.* 2013; Lazaridis *et al.* 2016; Anopheles gambiae 1000 Genomes Consortium 2017; Stolarek *et al.* 2018). Bhatia et al. (2013) proposed the bias-corrected F_{ST} estimator of Hudson (1992) for comparing two populations at a series of single nucleotide polymorphisms (SNPs). In a later study, this estimator was applied to measure the difference between ancient and contemporary human populations (Prohaska et al. 2019). An F_{ST} estimator between two populations that has been defined for large sample sizes by Weir and Goudet (2017, Equation 10) is applicable to such

126 situations.

127

128 An alternative approach for estimating population differentiation is the use of population-
129 specific F_{ST} estimators (Balding and Nichols 1995; Nicholson *et al.* 2002; Weir and Hill 2002;
130 Weir *et al.* 2005; reviewed by Gaggiotti and Foll 2010; Weir and Goudet 2017). Model-based
131 (Bayesian) approaches for estimating population-specific F_{ST}, referred to as F-models (Falush
132 *et al.* 2003), have been proposed for bi-allelic (Balding and Nichols 1995; Nicholson *et al.*
133 2002) and multi-allelic (Falush *et al.* 2003) cases. Karhunen and Ovaskainen (2012) proposed
134 an admixture F-model that extends the F-model of Falush *et al.* to account for limited gene
135 flow and small effective population size. Locus-population-specific F_{ST} is applied to identify
136 adaptive genetic divergence at a gene among populations using empirical Bayes (Beaumont
137 and Balding 2004) and full Bayesian methods (BayeScan) (Foll and Gaggiotti 2006, 2008).
138 Bias-corrected moment estimators of population-specific F_{ST} have also been proposed (Weir
139 and Hill 2002; Weir *et al.* 2005; Weir and Goudet 2017). Weir and Goudet (2017) derived
140 unbiased estimators for the components of the coancestry coefficients of the population-
141 specific F_{ST} estimator (hereafter, WG population-specific F_{ST}). The definition of the
142 population-specific F_{ST} is $\beta_{\mathrm{ST}}^{i} = \frac{\theta_S^i - \theta^B}{1-\theta^B}$, where θ_S^i is the average coancestry coefficient
143 within subpopulation i, and θ^B is that "over pairs of populations" (Weir and Goudet 2017).
144 Therefore, θ^B represents the average coancestry in all population samples, and $\beta_{\mathrm{ST}}^{i} = F_{\mathrm{ST}}^{i}$.
145 An important property of the population-specific F_{ST} measure is that "the usual global F_{ST}
146 estimator can be an unweighted average of population-specific F_{ST}"; as defined by Hudson et
147 al. (1992, Equation 3), this measure is described by the equation $\beta_{\mathrm{ST}} = \frac{\theta^S - \theta^B}{1-\theta^B}$ (Weir and
148 Goudet 2017). The population-specific F_{ST} for allele frequency data is $\beta_{\mathrm{WT}}^{i} = \frac{\theta_W^i - \theta^B}{1-\theta^B}$, and its

149 average over populations has been given previously (*e.g.*, Rousset 2004; Karhunen and
150 Ovaskainen 2012), namely, $F_{\mathrm{ST}} = \frac{\theta^W - \theta^B}{1-\theta^B}$, which equals the expectation of WC84 for equal
151 sample sizes (Weir and Goudet 2017). For random mating populations, there will be no need
152 for distinction between β_{ST} and β_{WT} (Weir and Goudet 2017). The moment estimators of
153 WG population-specific F_{ST} and WC84 are "ratio of averages" estimators; their precision
154 becomes higher for larger numbers of markers, and unbiased estimates can be obtained (Weir
155 and Cockerham 1984; Weir and Hill 2002; Bhatia *et al.* 2013). Because "the combined ratio
156 estimate (ratio of averages) is much less subject to the risk of bias than the separate estimate
157 (average of ratios)" (Cochran 1977), the population-specific F_{ST} measure is expected to
158 accurately reflect population history and provide more informative results than analyses using
159 pairwise F_{ST} values (Weir and Goudet 2017). Compared with these substantial efforts and
160 progress related to methodological development, however, empirical examples using
161 population-specific F_{ST} measures, except for limited applications to human populations (*e.g.,*
162 Nicholson *et al.* 2002; Weir *et al.* 2005; Foll and Gaggiotti 2006; Buckleton *et al.* 2013), have
163 been sparse. Furthermore, no comparison of population-specific F_{ST} and pairwise F_{ST} using
164 real data has been reported.
165
166 In this study, we characterized current population-specific F_{ST} estimators to infer population
167 histories and structures using publicly available human, marine fish and plant data sets. We
168 also compared our results obtained using pairwise F_{ST} and global F_{ST} estimates. Multi-
169 dimensional scaling (MDS) was applied to pairwise F_{ST} distance matrices to understand
170 causal mechanisms, such as environmental adaptation of current populations. The human
171 population data consisted of microsatellite genotypes collected from 51 world populations,
172 which may be a good example for characterizing these F_{ST} estimators because their

evolutionary history, migration and population structure has been extensively studied and is well understood (*e.g.*, Diamond 1997; Rosenberg *et al.* 2002; Ramachandran *et al.* 2005; Liu *et al.* 2006; Rutherford 2016; Nielsen *et al.* 2017). The SNP data sets were from a commercially important marine fish, Atlantic cod (*Gadus morhua*) in the North Atlantic, and from a tree, wild poplar (*Populus trichocarpa*) in the American Pacific Northwest. The Atlantic cod genotype data included historical samples collected 50–80 years ago as well as contemporary samples from the northern range margin of the species in Greenland, Norway and the Baltic Sea. The inclusion of both types of data might facilitate the detection of the effects of global warming on population structure. The wild poplar samples were collected under different environmental conditions over an area of 2,500 km near the Canadian–US border along with various environmental data and are thus possibly useful for the detection of environmental effects on population structure.

Materials and Methods

Population-specific F_{ST}

We applied WG bias-corrected population-specific F_{ST} moment estimators (Weir and Goudet 2017) in our data analyses:

$$\mathrm{ps}\hat{F}^{i}_{\mathrm{ST}} = \frac{\sum_{l=1}^{L}(\widetilde{M}^{i}_{W,l} - \widetilde{M}^{B}_{l})}{\sum_{l=1}^{L}\left(1 - \widetilde{M}^{B}_{l}\right)}, \qquad (1)$$

where $\widetilde{M}^{i}_{W}$ is the unbiased within-population matching of two distinct alleles of population i, and $\widetilde{M}^{B}$ is the between-population-pair matching average over pairs of populations i, i'. We derived the asymptotic variance of $\mathrm{ps}\hat{F}^{i}_{\mathrm{ST}}$ over all loci (Supplemental Note).

We also used empirical (Beaumont and Balding 2004) and full Bayesian (Foll and Gaggiotti

2006) population-specific F_{ST} estimators. Beaumont and Balding (2004) maximized the Dirichlet-multinomial marginal likelihood in their Equation 1 and estimated θ_{li}:

$$L_{li}\left(\theta_{li}\middle|n_{li1},\dots,n_{lim_l}\right)=\frac{\Gamma(\theta_{li})}{\Gamma(N_{li}+\theta_{li})}\prod_{u=1}^{m_l}\frac{\Gamma(n_{liu}+\theta_{li}\bar{p}_{lu})}{\Gamma(\theta_{li}\bar{p}_{lu})}. \qquad (2)$$

Here, θ_{li} is the scale parameter of the Dirichlet prior distribution for locus *l* and population *i*, $\tilde{p}_{lu}$ is the observed frequency of allele *u* at locus *l*, n_{liu} is the observed allele count in population *i*, and N_{li} is the total number of alleles. Importantly, $\bar{p}_{lu}$ is the mean allele frequency over all subpopulations, while $\theta_{li}\bar{p}_{lu}=\alpha_{liu}$, where $\theta_{li}=\sum_{u=1}^{m_l}\alpha_{liu}$. The parametrization reduces the number of parameters to be estimated ($\theta_{li}, l=1,\dots L; i=1,\dots,r$). Equation 2 is modeled to estimate locus-population-specific F_{ST} but is basically the same as the likelihood functions for estimating genetic differentiation and/or global F_{ST} previously given in Lange (1995), Rannala and Hartigan (1996), Kitada *et al.* (2000), Balding (2003) and Corander *et al.* (2003). The Dirichlet-multinomial marginal likelihood was also used for estimating global F_{ST} and linkage disequilibrium (Kitada and Kishino 2004), for detecting loci with outlier F_{ST} values (Foll and Gaggiotti 2006, 2008), for estimating global F_{ST} when the number of sampled population is small (Kitakado *et al.* 2006), and for estimating locus-specific global F_{ST} and posterior distributions of pairwise F_{ST} (Kitada *et al.* 2007). Based on a Dirichlet and/or a beta scale parameter, population-specific F_{ST} values are estimated for each locus using the following function of $\hat{\theta}_{li}$:

$$\mathrm{ps}\tilde{F}^{i}_{\mathrm{ST},l}=\frac{1}{\hat{\theta}_{li}+1}. \qquad (3)$$

Pairwise F_{ST}

We used Nei and Chesser's (1983) bias-corrected G_{ST} estimator (NC83) for estimating

218 pairwise F_{ST} over loci in our analysis:

219 $$\mathrm{pw}\hat{F}_{\mathrm{ST}} = \frac{\sum_{l=1}^{L}\left(\hat{H}_{T,l} - \hat{H}_{S,l}\right)}{\sum_{l=1}^{L}\hat{H}_{T,l}}, \qquad (4)$$

220 where $\hat{H}_T$ and $\hat{H}_S$ are the unbiased estimators of total and within-population heterozygosity,
221 respectively (Supplemental Note), with each variance component obtained from its moments.
222 G_{ST} (Nei 1973) is defined "by using the gene frequencies at the present population, so that no
223 assumption is required about the pedigrees of individuals, selection, and migration in the
224 past" (Nei 1977). G_{ST} (Nei 1973) assumes no evolutionary history (Holsinger and Weir 2009),
225 while NC83 does not consider any population replicates (Weir and Cockerham 1984). Our
226 pairwise F_{ST} values obtained from NC83 therefore measured current population structures
227 based on a fixed set of samples of subpopulations.

228

229 ***Genome-wide and locus-specific global F_{ST}***

230 We used Weir and Cockerham's (1984) θ (WC84) for estimating global F_{ST} over all loci
231 (genome-wide F_{ST}) as given by Equation 10 in the original study:

232 $$\theta_{WC} = \frac{\sum_{l=1}^{L}\sum_{u=1}^{m} a_{lu}}{\sum_{l=1}^{L}\sum_{u=1}^{m}(a_{lu} + b_{lu} + c_{lu})}. \qquad (5)$$

233 This equation is the ratio of the observed variance components for an allele: *a* for between
234 subpopulations, *b* for between individuals within subpopulations, and *c* for between gametes
235 within individuals. Each variance component is an unbiased estimator obtained with the
236 corresponding method of moment estimation (Supplemental Note). The WC84 moment
237 estimator of F_{ST} assumes that populations of the same size have descended separately from a
238 single "ancestral population" that was in both Hardy–Weinberg and linkage equilibrium (Weir
239 and Cockerham 1984). The statistical model of WC84 regards the sampled set of
240 subpopulations as having been taken from an ancestral population and considers replicates of

241 samples. This model also assumes that each population has the same population-specific F_{ST}
242 derived from the ancestral population (Weir and Hill 2002; Weir and Goudet 2017); it
243 therefore estimates the mean F_{ST} over subpopulations in terms of the coancestry coefficient
244 (global F_{ST}). The asymptotic variance of θ_{WC} over all loci was derived in a similar fashion
245 as that of WG population-specific F_{ST} (ps$\hat{F}_{\mathrm{ST}}^{i}$) as given in Supplemental Note.
246
247 A maximum likelihood estimator of a Dirichlet or beta scale parameter θ_l was used to
248 estimate locus-specific global F_{ST} values using Equation 1 in Kitada et al. (2007):

249 $$\mathrm{global}\hat{F}_{\mathrm{ST},l} = \frac{1}{\hat{\theta}_l + 1}. \qquad (6)$$

250 Throughout this paper, we use notations consistent with those of Weir and Hill (2002): i for
251 populations ($i = 1, \ldots, r$), u for alleles ($u = 1, \ldots, m$) and l for loci ($l = 1, \ldots, L$).
252
253 ***Empirical data***
254 ***Human microsatellite data*:** The data in Rosenberg *et al.* (2002) were retrieved from
255 https://web.stanford.edu/group/rosenberglab/index.html. We removed the Surui sample
256 (Brazil) from the data because that population was reduced to 34 individuals in 1961 as a
257 result of introduced diseases (Liu *et al.* 2006). We retained genotype data (n = 1,035) of 377
258 microsatellite loci from 51 populations categorized into six groups as in the original study: 6
259 populations from Africa, 12 from the Middle East and Europe, 9 from Central/South Asia, 18
260 from East Asia, 2 from Oceania and 4 from America. Longitudes and latitudes of the sampling
261 sites were obtained from Cann *et al.* (2002) (Supplemental Data).
262
263 ***Atlantic cod SNP data*:** The genotype data of 924 markers common to 29 populations
264 reported in Therkildsen *et al*. (2013a, b) and 12 populations in Hemmer-Hansen *et al*. (2013a,

265 b) were combined. We compared genotypes associated with each marker in samples that were
266 identical between the two studies, namely, CAN08 and Western_Atlantic_2008, ISO02 and
267 Iceland_migratory_2002, and ISC02 and Iceland_stationary_2002, and standardized the gene
268 codes. We removed cgpGmo.S1035, whose genotypes were inconsistent between the two
269 studies. We also removed cgpGmo.S1408 and cgpGmo.S893, for which genotypes were
270 missing in several population samples in Therkildsen *et al*. (2013b). Temporal replicates in
271 Norway migratory, Norway stationary, North Sea and Baltic Sea samples were removed for
272 simplicity. The final data set consisted of genotype data (n = 1,065) at 921 SNPs from 34
273 populations: 3 from Iceland, 25 from Greenland, 3 from Norway and 1 each from Canada, the
274 North Sea and the Baltic Sea. Two ecotypes (migratory and stationary) that were able to
275 interbreed but were genetically differentiated (Hemmer-Hansen *et al*. 2013a; Berg *et al*. 2016)
276 were included in the Norway and Iceland samples. All individuals in the samples were adults,
277 and most were mature (Therkildsen *et al*. 2013a). The longitudes and latitudes of the sampling
278 sites in Hemmer-Hansen *et al*. (2013a) were used. For the data from Therkildsen *et al*.
279 (2013a), approximate sampling points were estimated from the map of the original study, and
280 longitudes and latitudes were recorded (Supplemental Data).
281
282 ***Wild poplar SNP data*:** Environmental/geographical data and genotype data were retrieved
283 from the original studies of McKown *et al*. (2014a, b). The genotype data contained 29,355
284 SNPs of 3,518 genes of wild poplar (n = 441) collected from 25 drainage areas (McKown *et*
285 *al*. 2014c). Details of array development and selection of SNPs are provided in Geraldes et al.
286 (2011, 2013). The samples covered various regions over a range of 2,500 km near the
287 Canadian–US border at altitudes between 0 and 800 m (Supplemental Data). A breakdown of
288 the 25 drainages (hereafter, subpopulations) is as follows: 9 in northern British Colombia

289 (NBC), 2 in inland British Colombia (IBC), 12 in southern British Colombia (SBC) and 2 in
290 Oregon (ORE) (Geraldes *et al.* 2014). The original names of clusters and population numbers
291 were combined and used for our population labels (NBC1, NBC3,…, ORE30). Each sampling
292 location was associated with 11 environmental/geographical parameters: latitude (lat),
293 longitude (lon), altitude (alt), longest day length (DAY), frost-free days (FFD), mean annual
294 temperature (MAT), mean warmest month temperature (MWMT), mean annual precipitation
295 (MAP), mean summer precipitation (MSP), annual heat-moisture index (AHM) and summer
296 heat-moisture index (SHM) (Supplemental Data). The annual heat-moisture index was
297 calculated in the original study as (MAT+10)/(MAP/1000). A large heat-moisture index
298 indicates strong drying.
299

300 ***Data analysis***

301 ***Implementation of F_{ST} estimators:*** We converted the genotype data into Genepop format
302 (Raymond and Rousset 1995; Rousset 2008) for implementation in the R package
303 FinePop2_ver.0.2 (Nakamichi *et al.* 2020). Expected heterozygosity was calculated for each
304 population with the `read.GENEPOP` function. WG population-specific F_{ST} (Equation 1)
305 values over all loci were computed using the `pop_specificFST` function. Because WG
306 population-specific F_{ST} is a linear function of $\widehat{H}_{Si}$ with an intercept of 1 given $\widehat{H}_T$ (Equation
307 7 in Discussion), we examined the linear relationship between expected heterozygosity (*He*)
308 and population-specific F_{ST} estimates using the `lm` function in R. In addition to the "ratio of
309 averages" (Weir and Cockerham 1984; Weir and Hill 2002) used in the F_{ST} function, we
310 computed the "average of ratios" (Bhatia *et al.* 2013) of the WG population-specific F_{ST} of
311 the human data for comparison. We maximized Equation 2 and estimated Beaumont and
312 Balding's population-specific F_{ST} at each locus according to Equation 3. We then averaged

these values over all loci. For the full Bayesian model (Foll and Gaggiotti 2006), GESTE_ver. 2.0 was used to compute population-specific F_{ST} values. Pairwise F_{ST} values (NC83, Equation 4) were computed using the `pop_pairwiseFST` function. Global F_{ST} values over all loci (genome-wide global F_{ST}) were computed using the `globalFST` function (WC84, Equation 5). In addition, maximum-likelihood locus-specific global F_{ST} values (Equation 6) were estimated using the `locus_specificFST` function.

Visualization of population structure: All analyses were performed in R. We drew dendrograms based on pairwise F_{ST} values using the `hclust` function. Sampling points based on longitudes and latitudes were located on maps using the `sf` package. The size of each sampling point was drawn to be proportional to the expected heterozygosity, and the same colors were used as those of the clusters in the dendrograms. Sampling points with pairwise F_{ST} values smaller than a given threshold were connected by lines to visualize gene flow between populations. Multi-dimensional scaling (MDS) was applied on pairwise F_{ST} distance matrices to translate the information into a set of coordinates (axes) using the `cmdscale` function. As an explained variation measure, we used the cumulative contribution ratio up to the *k*th axis $(j = 1, \dots, k, \dots, K)$, which was computed in the function as $C_k = \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{K} \lambda_j$, where λ_j is the eigenvalue and $\lambda_j = 0 \; if \; \lambda_j < 0$. The signs of MDS values of coordinates were reversed to be consistent with WG population-specific F_{ST} values. Population-specific F_{ST} values for each sampling point were visualized by color gradients based on rgb $(1 - F_{ST,0}, 0, F_{ST,0})$, where $F_{ST,0} = (F_{ST} - \min F_{ST})/(\max F_{ST} - \min F_{ST})$. This conversion represents the magnitude of a population-specific F_{ST} value at the sampling point in colors between blue (for the largest F_{ST}) and red (smallest F_{ST} = closest to the coancestral population).

Effect of environment on population structure: We inferred the effect of environmental and geographical variables on population-specific F_{ST} and MDS axes of the pairwise F_{ST} values using a linear regression with the `lm` function. This analysis was only performed on the wild poplar data sets, for which 11 environmental/geographical parameters were associated with each sampling location.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, and supplemental information. Supplemental material available at figshare: ###. The R package for computing the F_{ST} estimators used in this paper is available in the FinePop2_ver.0.2 package at CRAN (https://CRAN.R-project.org/package=FinePop).

Results

Genome-wide and locus-specific global F_{ST}

The global F_{ST} estimate ± standard error (SE) was unexpectedly similar for the three cases. The estimate was 0.0488 ± 0.0012 for humans with a coefficient of variation (CV) of 0.025. The value for Atlantic cod, 0.0424 ± 0.0026 (CV=0.061), was slightly lower than that of human populations. The lowest global F_{ST} estimate was for wild poplar, 0.0415 ± 0.0002 (CV = 0.005), which was slightly lower than the estimate for Atlantic cod. The maximum likelihood estimator of locus-specific global F_{ST} generated positive F_{ST} values for all loci, and the mean of locus-specific global F_{ST} values ± SE was similar to global F_{ST} values: 0.0423 ± 0.0010 (CV = 0.024) for humans, 0.0390 ± 0.0018 (CV = 0.047) for Atlantic cod and 0.0431

361 ± 0.0002 (CV = 0.004) for wild poplar.

362

363 ***Human***

364 ***Population-specific F_{ST}*:** WG population-specific F_{ST} values were smallest in Africa (Figure
365 1A; Supplemental Data). Interestingly, Bantu Kenyans were associated with the smallest
366 value in Africa, while the San population, located in southern Africa, had the largest (Figure
367 S1). In Central/South Asia, the F_{ST} value closest to Africa was that of the Makrani, followed
368 by Sindhi, Balochi, Pathan and Brahui. Three samples from Uyghur, Hazara and Burusho
369 populations, which are the nearest to East Asia in terms of location, had the largest F_{ST} values.
370 The Kalash were isolated from other Central/South Asia populations. In the Middle East, the
371 closest population to Africa was Mozabite, followed by Palestinian and Bedouin. In Europe,
372 Adygean, Tuscan, Russian and French populations had similar F_{ST} values, and the largest
373 were those of Italians, Sardinians, Druze, Orcadians and Basques. In East Asia, the Xibo had
374 the smallest F_{ST} value, while She and Lahu populations were closest to Papuans and
375 Melanesians and had the largest F_{ST} values. In America, Mayans possessed the smallest F_{ST}
376 and the Karitiana the largest. Expected heterozygosity was highest in Africa and lowest in
377 South America (Figure 1A). The linear regression of population-specific F_{ST} on expected
378 heterozygosity was highly significant: $y = -0.1895x + 0.8908$ ($R^2 = 0.91, F =$
379 501.8 (1 , 49DF), $P < 2.2 \times 10^{-16}$).

380

381 As shown on the map in Figure 2A, visualization of human population history based on the
382 WG population-specific F_{ST} estimator indicated that populations in Africa had the smallest
383 F_{ST} values (shown in red), followed by the Middle East, Central/South Asia, Europe and East
384 Asia. The pattern in Oceania was similar to East Asia, but America was much different. As

385 illustrated by sampling point radii, heterozygosity was high in Africa, the Middle East,
386 Central/South Asia, Europe and East Asia but relatively small in Oceania and America. The
387 Kalash were less heterozygous than other populations in Central/South Asia. The Karitiana in
388 Brazil had the lowest heterozygosity. Bayesian population-specific F_{ST} values estimated using
389 the methods of Beaumont and Balding (2004) and Foll and Gaggiotti (2006) were nearly
390 identical, but in African populations they were higher than WG population-specific F_{ST}
391 (Figure S2). The distributions of F_{ST} values obtained from the two Bayesian methods were
392 very similar, with the smallest F_{ST} values observed in the Middle East, Europe and
393 Central/South Asia (Figure 2B, C; Supplemental Data). The "ratio of averages" and "average
394 of ratios" of the WG population-specific F_{ST} estimator were almost identical in all populations
395 for this data set (Figure S3).
396
397 ***Pairwise F_{ST}*:** On the basis of pairwise F_{ST} values, the populations were divided into five
398 clusters: 1) Africa, 2) the Middle East, Europe and Central/South Asia, 3) East Asia, 4)
399 Oceania and 5) America (Figure 1B). As indicated by sampling points with F_{ST} values below
400 the 0.02 threshold (connected by yellow lines in Figure 1C), gene flow from Africa was low.
401 Gene flow was substantial within Eurasia but was much smaller than that inferred from that
402 continent to Oceania and America (Figure 1C).
403
404 The first axis of MDS of pairwise F_{ST} exhibited a similar pattern to that of population-specific
405 F_{ST} values, with populations divided into five clusters (Figure 2D) as in the dendrogram
406 (Figure 1B). The second axis identified Caucasian and Mongoloid populations and indicated
407 close relationships between East Asia and Oceania (Figure 2E). The third axis uncovered
408 similarities among Europe, the Middle East, Central/South Asia, Central America and Oceania

409 and between Africa and South America (the Karitiana in Brazil) (Figure 2F). The contribution
410 of the first axis was 44% ($C_1 = 0.44$), and 72% of the variation in the pairwise F_{ST} distance
411 matrix was explained by the first, second and third axes of pairwise F_{ST} ($C_3 = 0.72$) (Figure
412 S4A; Supplemental Data). The first axis of pairwise F_{ST} was significantly positively
413 correlated with the population-specific F_{ST} values ($r = 0.85, t = 11.4, P = 2.22 \times 10^{-15}$),
414 indicating the distance from Africa for each population, but no other axes exhibited any
415 correlation (Figures 3A and S4B). The second axis divided populations in East Asia and
416 Oceania from the others, while the third axis characterized three groups: 1) Africa and the
417 Karitiana, 2) East Asia and 3) the Middle East, Europe, Central/South Asia and Oceania
418 (Figure 3B). The Kalash population was still isolated.
419

420 ***Atlantic cod***

421 ***Population-specific F_{ST}*:** The lowest WG population-specific F_{ST} value was in Canada
422 (Figures 4A and S5; Supplemental Data). Greenland west-coast populations (in green in
423 Figure 4A) generally had small F_{ST} values. Fjord populations (in deep purple) had relatively
424 higher F_{ST} values. Population PAA08 on the southwestern coast of Greenland, INC02 and
425 TAS10 in Iceland, and the stationary types ICEsta02 and NORsta09 in Norway had larger F_{ST}
426 values (in magenta). Population-specific F_{ST} values were much higher for offshore samples
427 OSO10/OEA10 and migratory types ICEmig02, NORmig (feed)09 and NORmig (spawn)09
428 (in orange). The F_{ST} value for NOS07 from the North Sea was smaller than those of the
429 migratory type, the highest of which was for BAS0607 from the Baltic Sea (in cyan).
430 Expected heterozygosity was the highest in Canada and the lowest in the Baltic Sea (Figure
431 4A). The linear regression of population-specific F_{ST} on expected heterozygosity was highly
432 significant: $y = -3.397x + 1.004$ ($R^2 = 0.998, F = 20250$ (1 , 32DF), $P < 2.2 \times 10^{-16}$).

The evolutionary history of Atlantic cod populations was clearly visualized on a map using WG population-specific F_{ST} values (Figure 5A). Heterozygosity (indicated by circle radii) was very high in Canada and Greenland, low in other areas and lowest in the Baltic Sea.

***Pairwise F_{ST}*:** The populations were divided according to pairwise F_{ST} values into four large clusters: 1) Canada, 2) Greenland west coast, 3) Greenland east coast, Iceland and Norway, and 4) North and Baltic seas. Fjord populations formed a sub-cluster within the Greenland west coast, and migratory and stationary ecotypes also formed a sub-cluster (Figure 4B). On the basis of pairwise F_{ST} values between sampling points (< 0.02 threshold), substantial gene flow was detected between Greenland, Iceland and Norway. In contrast, gene flow was low from Canada and to North and Baltic seas (Figure 4C).

The first axis of pairwise F_{ST} ($C_1 = 0.72$) was significantly correlated with population-specific F_{ST} values ($r = 0.86, t = 9.4, P = 9.55 \times 10^{-11}$) (Figure S6; Supplemental Data) and revealed patterns of population structure that were very similar to those inferred from population-specific F_{ST} values (Figure 5B). The first axis indicated the distance of each population from Canada (Figure 6A). The second axis of pairwise F_{ST} values, which exhibited a weak but significant correlation with population-specific F_{ST} values ($r = 0.35, t = 2.2, P = 0.04$), revealed different patterns of population differentiation (Figure 5C). This axis recognized the migratory ecotypes and also separated out southern Canadian, North Sea and Baltic Sea populations (Figure 6B). The cumulative contribution of the first and second axes was 94% ($C_2 = 0.92$). Other axes of pairwise F_{ST} were uncorrelated with population-specific F_{ST} values (Figure S6).

457 ***Wild poplar***

458 ***Population-specific F_{ST}:*** Wild-poplar population-specific F_{ST} values were lowest in SBC27,
459 SBC26, IBC15, IBC16 and ORE29 (Figure 7A; Supplemental Data). Samples collected from
460 areas close to the SBC coast had higher population-specific F_{ST} values than other SBC
461 samples. Samples SBC23, SBC24 and especially SBC22 had much higher population-specific
462 F_{ST} values. NBC samples had population-specific F_{ST} values similar to those of the SBC ones.
463 Among NBC samples, NBC8 had the smallest population-specific F_{ST}, and NBC5 had the
464 highest value, followed by NBC6 and NBC7. The population represented by sample ORE30
465 was isolated from ORE2. Expected heterozygosity was highest in SBC27 and lowest in
466 NBC5. The linear regression of population-specific F_{ST} on expected heterozygosity was
467 highly significant: $y = -1.872x + 0.629$ ($R^2 = 0.82, F = 108$ (1 ,23DF), $P <$
468 3.5×10^{-10}), but the variation in population-specific F_{ST} values was much larger than those
469 observed in humans and Atlantic cod. WG population-specific F_{ST}-based map visualization of
470 wild-poplar evolutionary history (Figure 8A) revealed that IBC15, IBC16 SBC27, SBC26 and
471 ORE29 had relatively small F_{ST} values and large heterozygosities, while NBC5, NBC6 and
472 SBC22 had the largest F_{ST} values and lowest heterozygosities.

473

474 ***Pairwise F_{ST}***: On the basis of pairwise F_{ST} values, the populations were divided into three
475 large clusters: 1) SBC, 2) NBC5, 6 and 7, and 3) NBC and IBC (Figure 7B). ORE samples,
476 however, were nested in the third cluster; in this cluster, the ORE29 sample was closely tied
477 to IBC15 and IBC15, while ORE30 was isolated. As inferred by pairwise F_{ST} values between
478 sampling points connected by yellow lines (< 0.02 threshold) in Figure 7C, substantial gene
479 flow was observed among populations (Figure 7C).

480

481 The first axis of pairwise F_{ST} uncovered slightly different patterns of population
482 differentiation (Figure 8B), which highlighted the dissimilarity between NBC5 and NBC6 (in
483 blue) and SBC22 (in red). The contribution of the first axis of pairwise F_{ST} was 47% ($C_1 =$
484 0.47) and had no significant correlation with population-specific F_{ST} values ($r = 0.33, t =$
485 $1.7, P = 0.108$) (Figure S7; Supplemental Data). The contribution of the second axis of
486 pairwise F_{ST} was 23% ($C_2 = 0.71$), and no correlation was likewise detected with population-
487 specific F_{ST} values ($r = 0.36, t = 1.8, P = 0.08$). The second axis strongly differentiated
488 ORE30 (in red) and SBC22 (in blue) (Figure 8C). The relationship between population-
489 specific F_{ST} and the first axis of pairwise F_{ST} was not linear, with expansion indicated in two
490 directions from inner areas: to the coast and to northern areas (Figure 9A). In the plot of the
491 second axis vs. the first axis of pairwise F_{ST}, ORE30 and SBC22 were distinct and located in
492 opposite regions of the graph (Figure 9B). The first axis of pairwise F_{ST} was positively
493 correlated with DAY (Figure 9C). The second axis, however, was negatively correlated with
494 SHM (Figure 9D), with SBC19 and SBC22 experiencing the wettest environment and ORE30
495 subjected to the driest conditions in the summer.
496
497 ***Effect of environment on population structure*:** To avoid multicollinearity, we excluded 7 out
498 of 11 environmental variables that were significantly correlated with each other, namely, lon,
499 alt, FFD, MWMT, MSP and AHM. Linear regression of population-specific F_{ST} values on the
500 four environmental variables (DAY, MAT, MAP and SHM) indicated that DAY, MAP and
501 SHM were significant (Table 1). DAY was also positive and significant on the first axis of
502 pairwise F_{ST}, and SHM was significantly negative on the second axis of pairwise F_{ST}. No
503 significant variable was found on the third axis.
504

505 ### *CPU times*

506 Using a laptop computer with an Intel Core i7-8650U CPU, only 89.8 s of CPU time were
507 required to compute WG population-specific F_{ST} estimates and SEs of wild poplar (29,355
508 SNPs; 25 populations, n = 441). Only 13.3 s were needed to compute maximum-likelihood
509 global F_{ST}, with 120.7 s required to obtain pairwise F_{ST} (NC83) between all pairs and 178.5
510 seconds for genome-wide global F_{ST} with SE (WC84).

511

512 ## Discussion

513 ### *Global F_{ST} measures species divergence*

514 Interestingly, the global F_{ST} estimate of WC84 was similar for the three species: 0.0488 ±
515 0.0012 (CV = 0.025), 0.0424 ± 0.0026 (CV = 0.061) and 0.0415 ± 0.0002 (CV = 0.005) for
516 human (377 microsatellite loci), Atlantic cod (921 SNPs) and wild poplar (29,355 SNPs)
517 populations, respectively. The highest global F_{ST} was that of humans. The rate of human gene
518 flow ($\theta = \frac{1}{F_{ST}} - 1$, Supplemental Note) was estimated to be 20, thus indicating that 20
519 effective individuals migrated per generation between subpopulations. Because this global F_{ST}
520 estimate was inferred using neutral microsatellite markers, it should reflect the random mating
521 history of humans within and between populations, such as from migration events (Diamond
522 1997; Rutherford 2016; Nielsen *et al.* 2017). The lowest global F_{ST} was that of wild poplar.
523 The species had an estimated gene flow rate of 23, which might be due to wind pollination
524 and/or fluffy seeds being carried by the wind. This gene flow rate was almost the same as that
525 of Atlantic cod, a migratory marine fish in which long-distance natal homing (>1,000 km)
526 over 60 years has been documented in the North Atlantic (Bonanomi *et al.* 2016).

527

528 The WC84 moment estimator of global F_{ST} is an unweighted average of population-specific
529 F_{ST} values (Weir and Goudet 2017). Therefore, genome-wide global F_{ST} would be a measure
530 of a species' genetic divergence, which reflects its evolutionary history. Atlantic cod may have
531 colonized the waters around Iceland and Norway following the last glacial maximum (LGM;
532 21 thousand years [kyr] ago) (Kettle *et al.* 2011; Hemmer-Hansen *et al*. 2013a). Wild poplar,
533 inferred to have been extensively distributed in coastal areas from southern California to
534 northern Alaska in the last interglacial (LIG; 135 kyr ago), was reduced to British Columbia,
535 Washington, Oregon and California in the LGM and then expanded to its current distribution
536 range of southern California to northern Alaska (Levsen *et al.* 2012). This scenario suggests
537 that our global F_{ST} value for wild poplar based on samples taken from British Columbia and
538 Oregon reflects population history after LGM, which coincides with the colonization date of
539 Atlantic cod in Iceland and Norway. For humans, the HapMap unweighted average of
540 population-specific F_{ST} values over all populations, estimated from 599,356 SNPs, was 0.13
541 (Weir *et al.* 2005). The earliest-known fossils of anatomically modern humans provide
542 evidence that modern humans originated in Ethiopia approximately 150–190 kyr ago and
543 appeared approximately 100 kyr ago in the Middle East and approximately 80 kyr ago in
544 southern China (reviewed by Nielsen et al. 2017). The genome-wide global F_{ST} value for
545 humans was three times greater than those of Atlantic cod and wild poplar, thus implying that
546 humans originated in Africa approximately 63 kyr ago ($= 21 \times 3$). This suggested timing is in
547 agreement with the best estimate of Liu *et al.* (2006), who inferred that humans initially
548 expanded 56 kyr ago from a small founding population of 1,000 effective individuals.
549

550 ***Population-specific F_{ST} traces population history as reflected by genetic***
551 ***diversity***

552 A linear relationship between heterozygosity and WG population-specific F_{ST} was evident in
553 our case studies (Figures 1A, 4A, 7A). The coefficient of determination, R^2, was 0.91 for 51
554 human populations (n = 1,035), 0.993 for 34 Atlantic cod populations (n = 1,065) and 0.82 for
555 25 wild poplar populations (n = 441). The goodness of fit to the linear function should depend
556 on population sample size (number of individuals). Our results from the three case studies
557 demonstrate that the population-specific F_{ST} estimator traces population history by way of
558 population genetic diversity.
559
560 In our analysis, WG population-specific F_{ST} clearly indicated that humans originated in
561 Africa, expanded from the Middle East into Europe and from Central/South Asia into East
562 Asia, and then possibly migrated to Oceania and America (Figures 1, 2). These results are in
563 good agreement with the highest levels of genetic diversity being detected in Africa
564 (Rosenberg *et al.* 2002), the relationship uncovered between genetic and geographic distance
565 (Ramachandran *et al.* 2005), the shortest colonization route from East Africa (Liu *et al.* 2006)
566 and major migrations inferred from genomic data (Nielsen *et al.* 2017). Our estimates of WG
567 population-specific F_{ST} values are consistent with results obtained from 24 forensic STR
568 markers (Buckleton *et al.* 2016) and successfully illustrate human evolutionary history.
569
570 The evolutionary history of Atlantic cod was also clearly visualized using WG population-
571 specific F_{ST} values (Figures 4A,5A). Our analysis indicated that Atlantic cod originated in
572 Canada (CAN08) and first expanded to the west coast of Greenland before spreading to
573 Iceland, the North Sea, Norway and the Baltic Sea. The migratory ecotypes may have played
574 an important role in this expansion. In the original Atlantic cod study (Therkildsen *et al.*
575 2013a), strong differentiation of CAN08 was found at neutral markers, which prompted the

576 authors to suggest that Greenland populations were the result of colonization from Iceland
577 rather than from refugial populations in southern North America. In our study, CAN08 had the
578 highest expected heterozygosity, which was lower in Iceland than in Greenland (Figure 4A);
579 this result implies that Icelandic populations were the descendants of colonists from
580 Greenland, which in turn originated in Canada. The BAS0607 sample from the Baltic Sea had
581 the highest population-specific F_{ST} and the lowest heterozygosity values, which suggests that
582 Baltic cod is the newest population. This result agrees with the findings of a previous study,
583 which identified Baltic cod as an example of a species subject to ongoing selection for
584 reproductive success in a low salinity environment (Berg et al. 2015).
585
586 The samples used in this study did not cover the whole distribution range of wild poplar,
587 which extends from southern California to northern Alaska. Population-specific F_{ST} values
588 suggested that wild poplar trees in southern British Colombia (SBC26, 27), inland British
589 Colombia (IBC15, 16) and Oregon (ORE29) are the closest to the ancestral population, which
590 later expanded in three directions: to coastal British Colombia (SBC; rainy summers),
591 southern Oregon (ORE30; mostly dry summers) and northern British Colombia (NBC; long
592 periods of daylight) (Figures 7A, 8A). The fact that the largest population-specific F_{ST} value
593 was found in the population with the smallest heterozygosity, SBC22, may be due to a
594 bottleneck (Geraldes *et al.* 2014).
595
596 ***The first axis of pairwise F_{ST} reflects population history***
597 Our results reveal that the first axis of pairwise F_{ST} reflects population history. The population
598 structure estimated for humans (Figure 1B) is in good agreement with that of the original
599 study (Rosenberg *et al.* 2002). MDS of the pairwise F_{ST} matrix decomposed the current

600 population structure into several independent axes. The first axis of pairwise F_{ST} was
601 significantly correlated with population-specific F_{ST} values (r = 0.85), with 72% ($R^2 = 0.72$)
602 of the variation in the current differentiation of the human population explained by their
603 evolutionary history (Figures 2D, 3A).
604
605 The first axis of pairwise F_{ST} of Atlantic cod was also significantly correlated with
606 population-specific F_{ST} values (r = 0.86), with 74% of the variation in the current
607 differentiation explained by the evolutionary history (Figures 5B, 6A). In contrast, the first
608 axis of pairwise F_{ST} of wild poplar was not significantly correlated with population-specific
609 F_{ST} (r = 0.33), and population expansion in two directions was detected from inner areas: to
610 the coast and to northern areas (Figure 9A). The first axis of pairwise F_{ST} was related to DAY,
611 the primary evolutionary factor in wild poplar. This result is consistent with the F_{ST} outlier
612 test of the original study (Geraldes et al. 2014), in which Bayescan (Foll and Gaggiotti 2008)
613 revealed that genes involved in circadian rhythm and response to red/far-red light had high
614 locus-specific global F_{ST} values. Moreover, the first principal component of SNP allele
615 frequencies was significantly correlated with daylength, and a previous enrichment analysis
616 for population structuring uncovered genes related to circadian rhythm and photoperiod
617 (McKown *et al*. 2014a). Our regression analysis of wild poplar revealed that long daylight,
618 abundant rainfall and dry summer conditions are the key environmental factors influencing
619 the evolution of this species (Table 1), thereby supporting the results obtained from
620 population-specific and pairwise F_{ST} values. Among our three case studies, wild poplar had
621 the lowest global F_{ST} and thus the highest gene flow. Since plants cannot move, this situation
622 may be due to wind pollination and/or wind transport of seeds, whose fates depend on the
623 environment in which they land.

624

625 ***Subsequent axes of pairwise F_{ST} reflect migration, languages, environmental***
626 ***effects and species specificity***

627 In humans, the second axis distinguished Caucasian and Mongoloid populations (Figure 2E);
628 it also revealed close relationships between populations in East Asia and Oceania, consistent
629 with an expansion from Asia into Polynesia and Micronesia (Diamond 1997). Descendants of
630 Chinese agriculturalists first spread from the islands of New Guinea to the east 3,600 years
631 ago and became the ancestors of modern Polynesians (Diamond 1997). The third axis
632 uncovered similarities among populations in Europe, the Middle East, Central/South Asia,
633 Central America and Oceania and between Africa and South America (Figure 2F). East Asian
634 populations were distinct from other populations. The Kalash population, separated in the
635 third axis (Figure 3B), lives in isolation in the highlands of northwestern Pakistan and
636 comprises approximately 4,000 individuals (Rutherford 2016) speaking an Indo-European
637 language (Rosenberg *et al.* 2002). The Khoisan-speaking San population, located at the
638 opposite end of the third axis in the plot, was separated from other African populations, who
639 speak Niger–Congo languages (Diamond 1997). Our result agrees with the previously
640 uncovered population genetic structure that is consistent with language classifications in
641 Africa (Tishkoff *et al.* 2009). Papuan and Melanesian populations from Papua New Guinea,
642 where the official language is English, had values similar to those of European populations.
643 Portuguese is the major language in Brazil (Karitiana), whereas Spanish predominates in
644 Colombia (Colombian) and Mexico (Maya and Pima). The third axis should reflect languages,
645 which is a consequence of migration events in the early modern period, such as during the
646 Age of Discovery (*e.g.*, Diamond 1997; Rutherford 2016).

647

648 In Atlantic cod, the second axis of pairwise F_{ST} had a weak but significant correlation with
649 population-specific F_{ST} values ($r = 0.35, t = 2.2, P = 0.04$) and exhibited different patterns
650 of population differentiation (Figure 5C). This axis separated out the migratory ecotype and
651 separated southern Canadian, North Sea and Baltic Sea populations at the opposite end
652 (Figure 6B). This placement of southern Canadian, North Sea and Baltic Sea populations
653 apart from the northern populations (Figure 6B) suggests that central distribution areas arose
654 to the north because of global warming, with the southern populations then becoming
655 isolated. Our result is in agreement with the results of an earlier study that identified parallel
656 temperature-associated clines in the allele frequencies of 40 of 1,641 gene-associated SNPs of
657 Atlantic cod in the eastern and western North Atlantic (Bradbury *et al.* 2010).
658
659 In wild poplar, the second axis of pairwise F_{ST} values was related to SHM gradients (summer
660 humidity) (Figure 9D), which suggests that precipitation, in addition to daylight detected by
661 the first axis, is a key factor in the current population structuring of wild poplar. In a previous
662 study, genes involved in drought response were identified as F_{ST} outliers along with other
663 genes related to transcriptional regulation and nutrient uptake (Geraldes et al. 2014), a finding
664 consistent with the results of our regression analysis (Table 1).
665

666 ### *Properties of F_{ST} estimators*

667 ***Ratio of averages*:** In regard to the WG population-specific F_{ST} estimator, similar results were
668 obtained for the human data using either the "ratio of averages" or the "average of ratios"
669 (Figure S3). These similar outcomes may have been due to the relatively small variation in the
670 locus-specific global F_{ST} values of human microsatellites. The highest precision, which was
671 dependent on the number of markers, was obtained for wild poplar. These results are in

agreement with previous studies that have suggested or indicated that the "ratio of averages" works better than the "average of ratios" (Cochran 1977; Weir and Cockerham 1984; Weir and Hill 2002; Bhatia et al. 2013).

To show the underlying mechanism, we use the observed heterozygosity of population *i* as derived in Nei and Chesser (1983) (Supplemental Note). When the number of loci (L) increases, the average observed heterozygosity over all loci converges to its expected value according to the law of large numbers as

$$\frac{1}{L}\sum_{l=1}^{L}\left(1-\sum\nolimits_{u=1}^{m}\tilde{p}_{iu}^{2}\right)\rightarrow\frac{1}{L}\sum_{l=1}^{L}\left(1-E\left[\sum\nolimits_{u=1}^{m}\tilde{p}_{iu}^{2}\right]\right).$$

The observed gene diversity thus converges to the expected value:

$$\widehat{H}_{Si}=\widehat{H}_{Si}\left(1-\frac{1}{n_i}\right)+\frac{\widehat{H}_{0i}}{2n_i}\ \rightarrow H_{Si}\left(1-\frac{1}{n_i}\right)+\frac{H_{0i}}{2n_i}.$$

In the same way, $\widehat{H}_S$ and $\widehat{H}_T$ converge to their expected values. This example indicates that the numerators and denominators of bias-corrected F_{ST} moment estimators, whether global, pairwise or population-specific, converge to their true means and provide unbiased estimates of F_{ST} in population genomics analyses with large numbers of SNPs. Our analyses demonstrate that genomic data highlight the usefulness of the bias-corrected moment estimators of F_{ST} developed in the early1980s (Nei and Chesser 1983; Weir and Cockerham 1984).

***Global F_{ST}*:** We used the WC84 moment estimator to estimate global F_{ST}. This bias-corrected moment estimator considers replicates of a set of population samples and is an unweighted average of population-specific F_{ST} values (Weir and Goudet 2017). Given a large number of loci in a genome, bias-corrected moment estimators of genome-wide F_{ST} ("ratio of averages")

695 enable reliable estimation of current population structure underpinned with evolutionary
696 history. For estimating locus-specific global F_{ST} values, maximum likelihood and Bayesian
697 estimators improve the power to detect genes under environmental selection because of the
698 shrinkage of allele frequencies toward the mean.
699
700 ***Pairwise F_{ST}*:** For estimation of pairwise F_{ST}, our previous coalescent simulations based on
701 ms (Hudson 2002) showed that NC83 performs best, among present F_{ST} estimators, for cases
702 with 10,000 SNPs (Kitada *et al.* 2017). Other F_{ST} moment estimators within an ANOVA
703 framework produce values approximately double those of true values. NC83 considers a fixed
704 set of population samples; in contrast, the other moment F_{ST} estimators consider replicates of
705 a set of populations (Weir and Cockerham 1984; Holsinger and Weir 2009), which causes the
706 over-estimation of pairwise F_{ST} (Kitada *et al.* 2017). Our empirical Bayes pairwise F_{ST}
707 estimator (EBF_{ST}; Kitada *et al.* 2007), which is also based on Equation 2, suffers from a
708 shrinkage effect similar to that of Bayesian population-specific F_{ST} estimators. EBF_{ST} is only
709 useful in cases involving a relatively small number of polymorphic marker loci, such as
710 microsatellites; it performs best by averaging the large sampling variation in allele
711 frequencies of populations with small sample sizes, particularly in high gene flow scenarios
712 (Kitada *et al.* 2017). We note, however, that the shrinkage effect on allele frequencies
713 enhances the bias of EBF_{ST} and other Bayesian F_{ST} estimators, particularly in genome
714 analyses (SNPs).
715
716 ***Population-specific F_{ST}***: The WG population-specific F_{ST} moment estimator measures
717 population genetic diversity under the framework of relatedness of individuals and identifies
718 the population with the largest genetic diversity as the ancestral population. This estimator

thus works to infer evolutionary history through genetic diversity. The WG population-specific F_{ST} estimator is based on allele matching probabilities, where within-population observed heterozygosity can be written as $1-\widetilde{M}_W^i$. When Hardy–Weinberg equilibrium is assumed ($\widehat{H}_{0i}=\widehat{H}_{Si}$), the preceding formula is equivalent to the NC83 unbiased estimator of the gene diversity of population i $\left(\widehat{H}_{Si}\right)$ (Supplemental Note):

$$1-\widetilde{M}_W^i=\frac{2n_i}{2n_i-1}\left(1-\sum_{u=1}^{m}\tilde{p}_{iu}^2\right)=\widehat{H}_{Si}\,.$$

Another variable, $\widetilde{M}^B$, is total homozygosity in terms of paired matching of alleles over all populations. The definition of WG population-specific F_{ST} is

$$\mathrm{ps}F_{\mathrm{ST}}^i=\beta_{\mathrm{WT}}^i=\frac{\theta_W^i-\theta^B}{1-\theta^B}$$

and the estimator is

$$\mathrm{ps}\hat{F}_{\mathrm{ST}}^i=\hat{\beta}_{\mathrm{WT}}^i=\frac{\widetilde{M}_w^i-\widetilde{M}^B}{1-\widetilde{M}^B}\,.$$

Weir and Goudet (2017) showed that $E[\hat{\beta}_{\mathrm{WT}}^i]=\hat{\beta}_{\mathrm{WT}}^i$ and that θ^B is the average of identical-by-descent (ibd) probabilities of alleles from different populations. We may therefore write $1-\widetilde{M}^B=\widehat{H}_T$. When working with allele frequencies, the population-specific F_{ST} estimator can be written in terms of Nei's gene diversity as

$$\mathrm{ps}\hat{F}_{\mathrm{ST}}^i=\frac{\widetilde{M}_w^i-\widetilde{M}^B}{1-\widetilde{M}^B}=\frac{\widehat{H}_T-\widehat{H}_{Si}}{\widehat{H}_T}=1-\frac{\widehat{H}_{Si}}{\widehat{H}_T}\,. \quad (7)$$

This formulation is reasonable, since WG population-specific F_{ST} uses "allele matching, equivalent to homozygosity and complementary to heterozygosity as used by Nei, rather than components of variance (Weir and Cockerham 1984)" (Weir and Goudet 2017).

In our study, the empirical Bayesian (Beaumont and Balding 2004) and full Bayesian (Foll and Gaggiotti 2006) population-specific F_{ST} estimators consistently indicated that the Middle

East, Europe and Central/South Asia were centers of human origin. The results obtained with the empirical Bayesian estimator were a consequence of Equation 2, which uses the mean allele frequency over subpopulations ($\bar{p}_{lu}$) to reduce the number of parameters to be estimated. The locations of the 51 human populations were as follows: 21 from the Middle East, Europe and Central/South Asia, 18 from East Asia, 6 from Africa, 2 from Oceania and 4 from America. The mean allele frequency ($\bar{p}_{lu}$) reflected the weight of samples from the Middle East, Europe and Central/South Asia, thereby resulting in these areas being identified as centers of origin. Instead of $\bar{p}_{lu}$, the full Bayesian method uses allele frequencies in the ancestral population, p_{lu}, which are generated from a noninformative Dirichlet prior, $p_{lu} \sim Dir\ (1, \ldots, 1)$. Our result suggests that not enough information is available to estimate allele frequencies in the ancestral population assumed in the models. The shrinkage effect on allele frequencies in Bayesian inference (Stein 1956) may shift population-specific F_{ST} values toward the average of the whole population. Indeed, Bayesian population-specific F_{ST} values were higher for African populations than WG population-specific F_{ST} ones and close to those for East Asia (Figures S2B).

Conclusions

The moment estimator of WG population-specific F_{ST} identifies the source population and traces migration events and the evolutionary history of its derived populations by way of genetic diversity. In contrast, NC83 pairwise F_{ST} represents the current population structure. Generally, the first axis of MDS of the pairwise F_{ST} distance matrix reflects population history, while subsequent axes reflect migration events, languages and the effect of environment. The relative contributions of these factors depend on the ecological characteristics of the species. Because of shrinkage towards mean allele frequencies,

maximum likelihood and Bayesian estimators of locus-specific global F_{ST} improve the power to detect genes under environmental selection. In contrast, the WC84 bias-corrected moment estimator of global F_{ST} enables reliable estimation of current population structure, reflecting evolutionary history. Given a large number of loci, bias-corrected F_{ST} moment estimators, whether global, pairwise or population-specific, provide unbiased estimates of F_{ST} supported by the law of large numbers. Genomic data highlight the usefulness of the bias-corrected moment estimators of F_{ST}. All F_{ST} estimators described in this paper have reasonable CPU times.

Acknowledgements

This study was supported by Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research KAKENHI nos. 16H02788 and 19H04070 to HK and 18K0578116 to SK. We thank B. Goodson from Edanz Group for editing the English text of a draft of this manuscript.

Literature Cited

Anopheles gambiae 1000 Genomes Consortium, 2017 Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature 552 (7683): 96. https://doi:10.1038/nature24995

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. 63: 221–230. https://doi.org/10.1016/S0040-5809(03)00007-8

Balding, D. J., and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12. https://doi.org/10.1007/BF01441146

Balding, D. J., M. Greenhalgh, and R. A. Nichols, 1996. Population genetics of STR loci in

Caucasians. Int. J. Legal Med. 108: 300–305. https://doi.org/10.1007/BF02432124

Balloux, F., and N. Lugon-Moulin, 2002 The estimation of population differentiation with microsatellite markers. Mol. Ecol. 11: 155–165. https://doi.org/10.1046/j.0962-1083.2001.01436.x

Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Molec. Ecol. 13: 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x

Beaumont, M. A., 2005 Adaptation and speciation: what can F_{ST} tell us? Trends Ecol. Evol. 20: 435–440. https://doi.org/10.1016/j.tree.2005.05.017

Berg, P. R., S. Jentoft, B. Star, K. H. Ring, H. Knutsen *et al.* 2015 Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.). Genome Biol. Evol. 7: 1644–1663. https://doi.org/10.1093/gbe/evv093

Berg, P. R., B. Star, C. Pampoulie, M. Sodeland, J. M. Barth *et al.* 2016 Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. Sci. Rep. 6: 23246. https://doi.org/10.1038/srep23246

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price, 2013 Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. 23: 1514–1521. http://www.genome.org/cgi/doi/10.1101/gr.154831.113

Bonanomi, S., N. Overgaard Therkildsen, A. Retzel, R. Berg Hedeholm, M. W. Pedersen, *et al.*, 2016 Historical DNA documents long-distance natal homing in marine fish. Molec. Ecol. 25: 2727–2734. https://doi.org/10.1111/mec.13580

Bradbury, I. R., S. Hubert, B. Higgins, T. Borza, S. Bowman, *et al.*, 2010 Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. Proc. Royal Soc. B: Biol. Sci. 277: 3725–3734.

https://doi.org/10.1098/rspb.2010.0985

Browning, S. R., and B. S. Weir, 2010 Population structure with localized haplotype clusters. Genetics, 185: 1337–1344. https://doi.org/10.1534/genetics.110.116681

Cann, H. M., C. De Toma, L. Cazes, M. F. Legrand, V. Morel, *et al.*, 2002 A human genome diversity cell line panel. Science, 296: 261–262. DOI: 10.1126/science.296.5566.261b

Cheng, T., J. Wu, Y. Wu, R. V. Chilukuri, L. Huang *et al*., 2017 Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. Nat. Ecol. Evol. 1, 1747. http://doi:10.1038/s41559-017-0314-4

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution 23: 72–84. https://doi.org/10.1111/j.1558-5646.1969.tb03496.x

Cockerham, C. C., 1973 Analyses of gene frequencies. Genetics 74: 679–700. PubMed 17248636

Cochran, W. G. 1977 *Sampling Techniques*. John Wiley & Sons, New York.

Corander, J., P. Waldmann, and M. J. Sillanpää, 2003 Bayesian analysis of genetic differentiation between populations. Genetics 163: 367–374. PubMed 12586722

Crow, J. F., and K. Aoki, 1984 Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. 81: 6073–6077. https://doi.org/10.1073/pnas.81.19.6073

Diamond, J. 1997 *Guns, Germs and Steel: The Fates of Human Societies.* Random House, London.

Excoffier, L., 2001 Analysis of population subdivision, pp. 271–307 in Handbook of Statistical Genetics, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

837 Excoffier, L., and H.E. L. Lischer (2010) Arlequin suite ver 3.5: A new series of programs to
838 perform population genetics analyses under Linux and Windows. Molec. Ecol. Res. 10:
839 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x
840 Foll, M., and O.Gaggiotti, 2006 Identifying the environmental factors that determine the
841 genetic structure of populations. Genetics 174: 875–891.
842 https://doi.org/10.1534/genetics.106.059451
843 Foll, M., and O.Gaggiotti, 2008. A genome-scan method to identify selected loci appropriate
844 for both dominant and codominant markers: a Bayesian perspective. Genetics 180:
845 977–993. https://doi.org/10.1534/genetics.108.092221
846 Geraldes, A., J. Pang, N. Thiessen, T. Cezard, R. Moore *et al.* 2011 SNP discovery in black
847 cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. Molec.
848 Ecol. Resour. 11: 81–92. https://doi.org/10.1111/j.1755-0998.2010.02960.x
849 Geraldes, A., S. P. Difazio, G. T. Slavov, P. Ranjan, W. Muchero *et al.*, 2013 A 34K SNP
850 genotyping array for *Populus trichocarpa*: design, application to the study of natural
851 populations and transferability to other *Populus* species. Molec. Ecol. Resour. 13: 306–
852 323. https://doi.org/10.1111/1755-0998.12056
853 Geraldes, A., N. Farzaneh, C. J. Grassa, A. D. McKown, R. D. Guy *et al.*, 2014 Landscape
854 genomics of *Populus trichocarpa* the role of hybridization limited gene flow and
855 natural selection in shaping patterns of population structure. Evolution 68: 3260–3280.
856 https://doi.org/10.1111/evo.12497
857 Goudet, J. 1995 FSTAT (version 1.2): a computer program to calculate *F*-statistics. J. Hered.
858 86: 485–486.
859 Griffin, P. C., S. B. Hangartner, A. Fournier-Level, and A. A. Hoffmann, 2017 Genomic
860 trajectories to desiccation resistance: convergence and divergence among replicate

selected Drosophila lines. Genetics 205: 871–890. https://doi.org/10.1534/genetics.116.187104

Hemmer‐Hansen, J., E. E. Nielsen, N. O. Therkildsen, M. I. Taylor, R. Ogden *et al.*, 2013a A genomic island linked to ecotype divergence in Atlantic cod. Molec. Ecol. 22: 2653–2667. https://doi.org/10.1111/mec.12284

Hemmer‐Hansen, J., E. E. Nielsen, N. O. Therkildsen, M. I. Taylor, R. Ogden *et al.*, 2013b Data from: A genomic island linked to ecotype divergence in Atlantic cod, Dryad, Dataset, https://doi.org/10.5061/dryad.9gf10

Holsinger, K. E., 1999 Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. Hereditas 130: 245–255. https://doi.org/10.1111/j.1601-5223.1999.00245.x

Holsinger, K. E., P. O. Lewis, and D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol. Ecol. 11: 1157–1164. https://doi.org/10.1046/j.1365-294X.2002.01512.x

Holsinger, K.E., and B. S. Weir, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat. Rev. Genet. 9: 639–650. https://doi.org/10.1038/nrg2611

Hudson, R. R. 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Kettle AJ, Morales-Muniz A, Rosello-Izquierdo E, Heinrich D, Vollestad LA (2011) Refugia of marine fish in the northeast Atlantic during the last glacial maximum: concordant assessment from archaeozoology and palaeotemperature reconstructions.

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

https://doi.org/10.1534/genetics.103.023044

Kitada, S., T. Kitakado, and H. Kishino, 2007. Empirical Bayes inference of pairwise F_{ST} and its distribution in the genome. Genetics 177: 861–873. https://doi.org/10.1534/genetics.107.077263

Kitada, S., R., Nakamichi, and H. Kishino, 2017 The empirical Bayes estimators of fine-scale population structure in high gene flow species. Molec. Ecol. Resour.17: 1210–1222. https://doi.org/10.1111/1755-0998.12663

Kitakado, T., S. Kitada, H. Kishino, and H. J. Skaug, 2006 An integrated-likelihood method for estimating genetic differentiation between populations. Genetics 173: 2073–2082. https://doi.org/10.1534/genetics.106.055350

Lamichhaney, S., A. P. Fuentes-Pardo, N. Rafati, N. Ryman, G. R. McCracken *et al*., 2017 Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. Proc. Natl. Acad. Sci. USA 114: E3452–E3461. https://doi.org/10.1073/pnas.1617728114

Lange, K., 1995 Application of the Dirichlet distribution to forensic match probabilities. Genetica 96: 107–117. https://doi.org/10.1007/BF01441156

Lazaridis, I., D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland *et al.*, 2016 Genomic insights into the origin of farming in the ancient Near East. Nature, 536(7617), 419. https://doi:10.1038/nature19310

Levsen, N. D., P. Tiffin, and M. S. Olson, 2012 Pleistocene speciation in the genus *Populus* (Salicaceae). System. Biol. 61: 401. https://doi.org/10.1093/sysbio/syr120

Liu, H., F. Prugnolle, A. Manica, and F. Balloux, 2006 A geographically explicit genetic model of worldwide human-settlement history. Am. Hum. Genet. 79: 230–237. https://doi.org/10.1086/505436

Lockwood, J. R., K. Roeder, and B. Devlin, 2001 A Bayesian hierarchical model for allele frequencies. Genet. Epidemiol. 20: 17–33. https://doi.org/10.1002/1098-2272(200101)20:1<17::AID-GEPI3>3.0.CO;2-Q

McKown, A. D., R. D. Guy, J. Klápště, A. Geraldes, M. Friedmann *et al.*, 2014a Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. New Phytol. 201, 1263–1276. https://doi.org/10.1111/nph.12601

McKown, A. D., J. Klápště, R. D. Guy, A. Geraldes, I. Porth *et al.*, 2014b Genome‐wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. New Phytol. 203: 535–553. https://doi.org/10.1111/nph.12815

McKown, A. D., R. D. Guy, L. Quamme, J. Klápště, J. La Mantia *et al.*, 2014c Association genetics, geography and ecophysiology link stomatal patterning in *Populus trichocarpa* with carbon gain and disease resistance trade‐offs. Molec, Ecol. 23: 5771‐5790. https://doi.org/10.1111/mec.12969

Malinsky, M., R. J. Challis, A. M. Tyers, S. Schiffels, Y. Terai *et al.*, 2015 Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science 350: 1493-1498. DOI: 10.1126/science.aac9927

Nakamichi, R., H. Kishino, S. Kitada, 2020 Fine-Scale Population Analysis 2. CRAN (https://CRAN.R-project.org/package=FinePop).

Nei, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70: 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Nei, M., 1975 *Molecular population genetics and evolution*. North-Holland. Amsterdam.

Nei, M., 1977 *F*‐statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41: 225–233. https://doi.org/10.1111/j.1469-1809.1977.tb01918.x

Nei, M., and R. K. Chesser, 1983 Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47: 253–259. https://doi.org/10.1111/j.1469-1809.1983.tb00993.x

Nicholson, G., A. V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson *et al.*, 2002 Assessing population differentiation and isolation from single‐nucleotide polymorphism data. J. Roy. Stat. Soc. B. Stat. Method. 64: 695‐715. https://doi.org/10.1111/1467-9868.00357

Nielsen, R., J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, and E. Willerslev, 2017 Tracing the peopling of the world through genomics. Nature 541: 302–310. doi:10.1038/nature21347

Pérez-Lezaun, A., F. Calafell, E. Mateu, D. Comas, R. Ruiz-Pacheco *et al.,* 1997 Microsatellite variation and the differentiation of modern humans. Human Genet. 99: 1–7. https://doi.org/10.1007/s004390050299

Prohaska, A., F. Racimo, A. J. Schork, M. Sikora, A. J. Stern *et al.*, 2019 Human disease variation in the light of population genomics. Cell 177: 115–131. https://doi.org/10.1016/j.cell.2019.01.052

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. 102: 15942–15947. https://doi.org/10.1073/pnas.0507611102

Rannala, B., and J. A. Hartigan, 1996 Estimating gene flow in island populations. Genet. Res. 67: 147–158. https://doi.org/10.1017/S0016672300033607

Raymond, M., and F., Rousset,1995 GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. 86: 248–249.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, 2002 Genetic structure of human populations. Science 298: 2381–

2385. DOI: 10.1126/science.1078311

Rousset, F., 2001 Inferences from spatial population genetics, pp. 239–269 in Handbook of Statistical Genetics, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Rousset, F., 2002 Inbreeding and relatedness coefficients: what do they measure? Heredity 88: 371–380.

Rousset, F., 2004 *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Rousset, F., 2008 Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. Mol. Ecol. Resour. 8: 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Rutherford, A. 2016 *A Brief History of Everyone Who Ever Lived: The Human Story Retold Through Our Genes.* The Experiment, NY.

Slatkin, M. 1991 Inbreeding coefficients and coalescence times. Genet. Res. 58: 167–175. https://doi.org/10.1017/S0016672300029827

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

Steele, C. D., D. S. Court, and D. J. Balding, 2014 Worldwide estimates relative to five continental‐scale populations. Ann. Hum. Genet. 78: 468–477. https://doi.org/10.1111/ahg.12081

Stolarek, I., A. Juras, L. Handschuh, M. Marcinkowska-Swojak, A. Philips *et al*., 2018 A mosaic genetic structure of the human population living in the South Baltic region

during the Iron Age. Sci. Rep. 8(1), 2455. https://doi.org/10.1038/s41598-018-20705-6

Therkildsen, N. O., J. Hemmer‐Hansen, R. B. Hedeholm, M. S. Wisz, C. Pampoulie *et al.* 2013a Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*. Evol. Appl. 6: 690–705. https://doi.org/10.1111/eva.12055

Therkildsen, N. O., J. Hemmer‐Hansen, R. B. Hedeholm, M. S. Wisz, C. Pampoulie *et al.* 2013b Data from: Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*, v2, Dryad, Dataset, https://doi.org/10.5061/dryad.rd250

Therkildsen, N. O., J. Hemmer‐Hansen, T. D. Als, D. P. Swain, M. J. Morgan *et al.*, 2013c Microevolution in time and space: SNP analysis of historical DNA reveals dynamic signatures of selection in Atlantic cod. Molec. Ecol. 22: 2424–2440. https://doi.org/10.1111/mec.12260

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science 324: 1035–1044. DOI: 10.1126/science.1172257

Weir, B. S., 1996 *Genetic data analysis II*. Sinauer, Sunderland.

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

Weir, B. S., and W. G. Hill, 2002 Estimating *F*-statistics. Annu. Rev. Genet. 36: 721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill, 2005 Measures of human population structure show heterogeneity among genomic regions. Genome Res. 15: 1468–1476. doi: 10.1101/gr.4398405

Weir, B. S., and J. Goudet, 2017 A unified characterization of population structure and relatedness. Genetics 206: 2085–2103. https://doi.org/10.1534/genetics.116.198424

Wright, S., 1931 Evolution in mendelian populations. Genetics 16: 97–158. PMID: 17246615

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

1010 ## Figure legends

1011 **Figure 1** Population structure of 51 human populations (*n* = 1,035; 377
1012 microsatellites). (A) *He* vs. population-specific F_{ST} (Weir and Goudet 2017). (B)
1013 Population structure and (C) gene flow based on pairwise F_{ST}. Populations
1014 connected by yellow lines are those with pairwise $F_{ST} < 0.01$. The radius of each
1015 sampling point is proportional to the level of *He* as visualized by H_e^{12}.

1016

1017 **Figure 2** Population-specific F_{ST} visualization of the population structure of 51 human
1018 populations. (A) WG (Weir and Goudet 2017). (B) BB (Beaumont and Balding 2004).
1019 (C) FG (Foll and Gaggiotti 2006). (D–F) First to third MDS axes of pairwise F_{ST} with
1020 72% goodness of fit. Data from Rosenberg *et al.* (2002).

1021

1022 **Figure 3** Relationships between population-specific F_{ST} and MDS axes of pairwise
1023 F_{ST} for 51 human populations. (A) First axis of pairwise F_{ST} vs. population-specific
1024 F_{ST} (Weir and Goudet 2017). (B) Second vs. third axes of pairwise F_{ST}. Data from
1025 Rosenberg *et al.* (2002).

1026

1027 **Figure 4** Population structure of 34 geographical samples of wild Atlantic cod (*n* =
1028 1,065; 921 SNPs). (A) *He* vs. population-specific F_{ST} (Weir and Goudet 2017). (B)
1029 Population structure and (C) gene flow based on pairwise F_{ST}. Populations
1030 connected by yellow lines are those with pairwise $F_{ST} < 0.04$. The radius of each
1031 sampling point is proportional to the level of heterozygosity (*He*) as visualized by
1032 H_e^{100}. Data are combined data from Therkildsen *et al.* (2013) and Hemmer-Hansen *et*
1033 *al.* (2013).

1034

1035 **Figure 5** Population-specific F_{ST} visualization of the population structure of 34
1036 samples of Atlantic cod. (A) Population-specific F_{ST} (Weir and Goudet 2017). (B) First
1037 and (C) second axes of pairwise F_{ST} with 97% goodness of fit. Data are combined
1038 data from Therkildsen *et al.* (2013) and Hemmer-Hansen *et al.* (2013).

1039

1040 **Figure 6** Relationships between population-specific F_{ST} and MDS axes of pairwise
1041 F_{ST} for 34 Atlantic cod populations. (A) First axis of pairwise F_{ST} vs. population-
1042 specific F_{ST} (Weir and Goudet 2017). (B) First vs. second axes of pairwise F_{ST}.
1043 Combined data from Therkildsen *et al.* (2013) and Hemmer-Hansen *et al.* (2013).

1044

1045 **Figure 7** Population structure of 25 geographical samples of wild poplar (n = 441;
1046 29,355 SNPs). (A) *He* vs. population-specific F_{ST} (Weir and Goudet 2017). (B)
1047 Population structure and (C) gene flow based on pairwise F_{ST}. Populations
1048 connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The radius of each
1049 sampling point is proportional to the level of heterozygosity (*He*) as visualized by
1050 H_e^{100}. Data from McKown *et al.* (2014b).

1051

1052 **Figure 8** Population-specific F_{ST} visualization of the population structure of 25
1053 samples of wild poplar. (A) Population-specific F_{ST} (Weir and Goudet 2017). (B) First
1054 and (C) second axes of pairwise F_{ST} with 71% goodness of fit. Data from McKown *et*
1055 *al.* (2014b).

1056

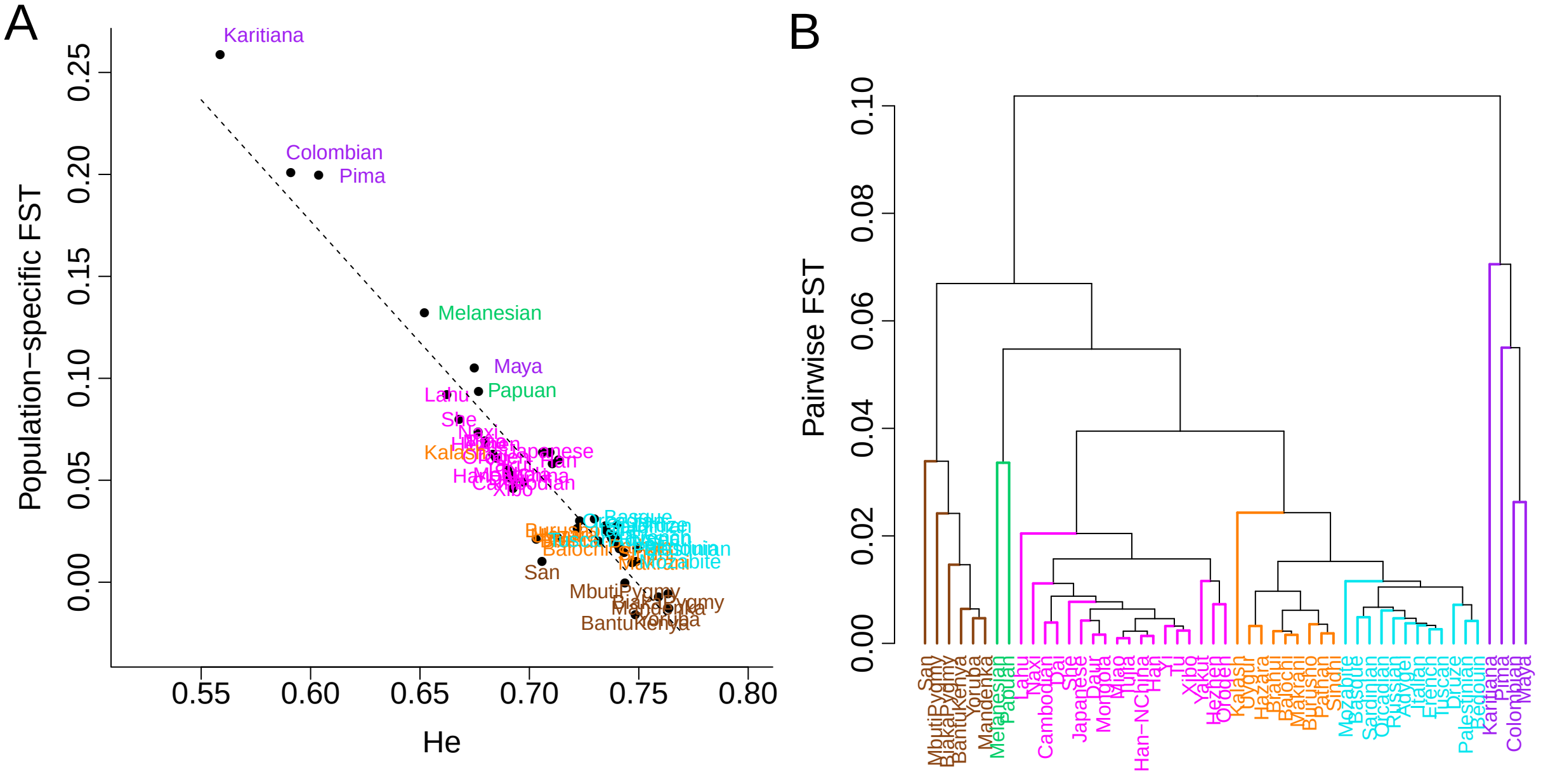
1057 **Figure 9** Relationships between population-specific F_{ST}, MDS axes of pairwise F_{ST}

1058 and environmental variables for 25 wild poplar samples. (A) First axis of pairwise F_{ST}
1059 vs. population-specific F_{ST} (Weir and Goudet 2017). (B) First vs. second axes of
1060 pairwise F_{ST}. (C) Longest day length (h) vs. first axis of pairwise F_{ST}. (D) Summer
1061 heat-moisture index vs. second axis of pairwise F_{ST}. Data from McKown *et al.*
1062 (2014a,b).

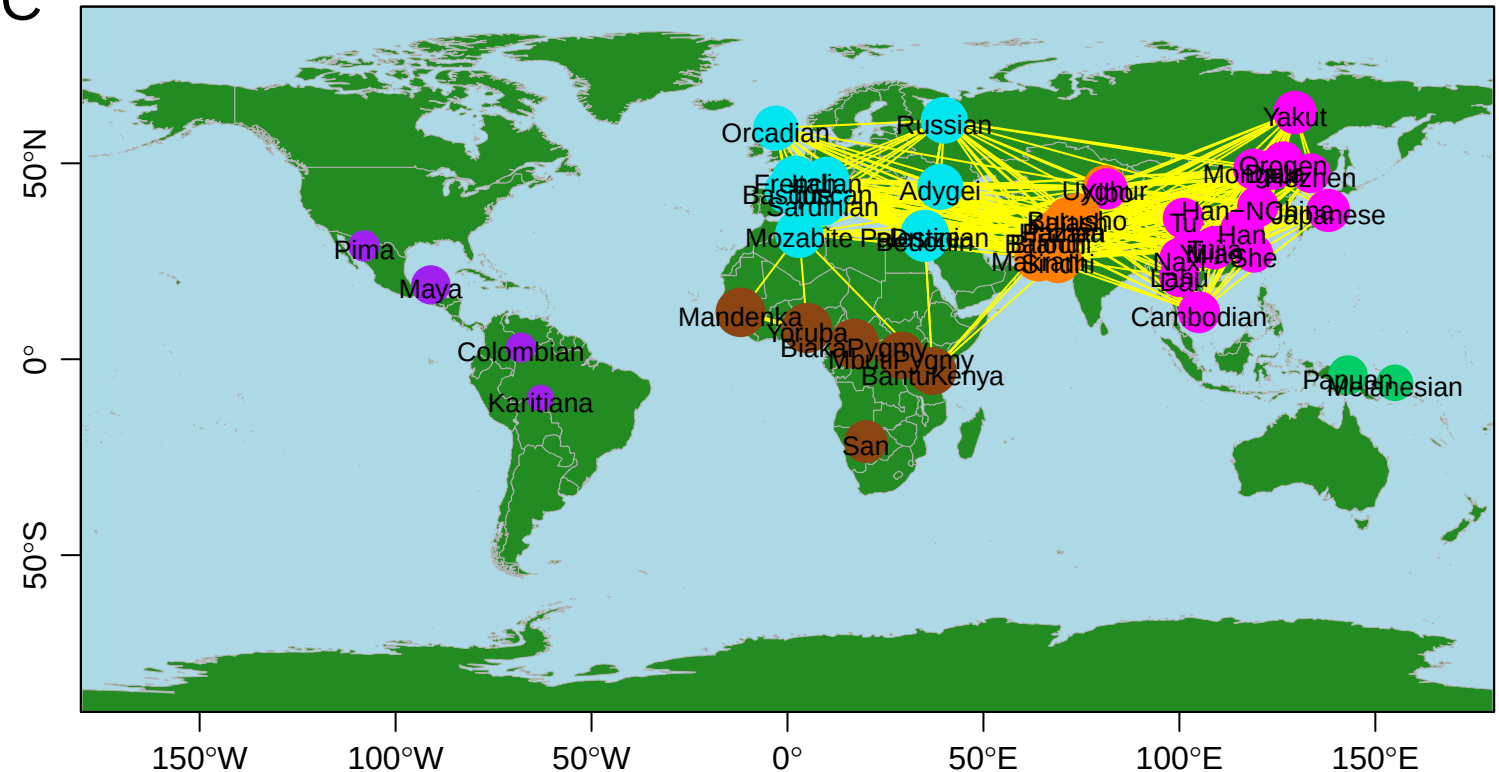
Table 1 Multiple regression of environmental variables of wild poplar on population-specific F_{ST} values of 25 populations

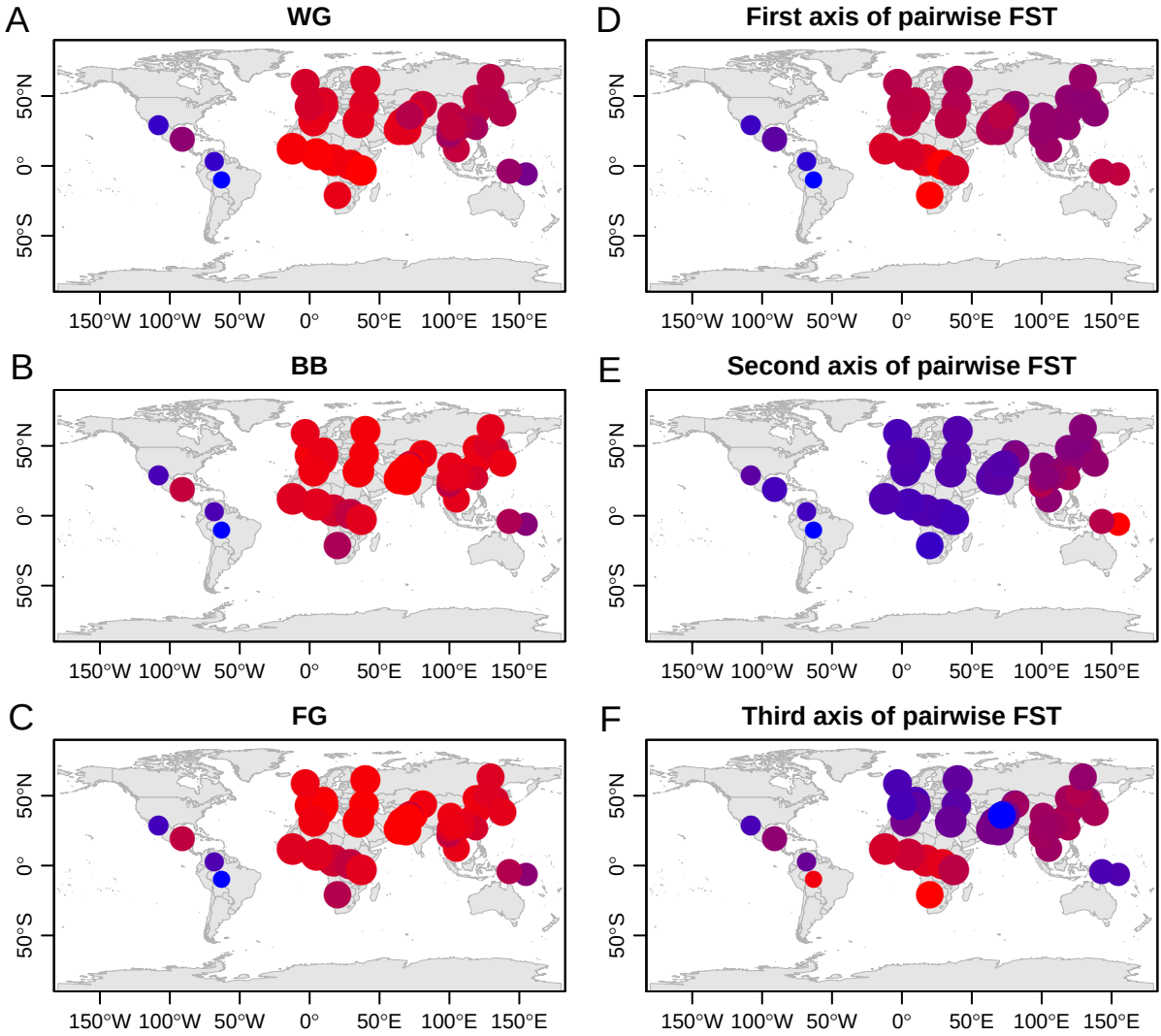
Explanatory	Population-specific F_{ST}				First axis of pairwise F_{ST}				Second axis of pairwise F_{ST}			
variables	Estimate	SE	*t*	*p*	Estimate	SE	*t*	*p*	Estimate	SE	t	*p*
Intercept	-0.7992	0.3229	-2.468	0.0224*	-0.2790	0.0858	-3.255	0.0040**	0.0341	0.0615	0.555	0.5851
DAY	0.0392	0.0165	2.354	0.0289*	0.0159	0.0044	3.600	0.0018**	-0.0004	0.0032	-0.134	0.8948
MAT	-0.0165	0.0104	-1.578	0.1302	-0.0019	0.0028	-0.678	0.5053	0.0008	0.0020	0.380	0.7082
MAP	0.0001	0.0000	2.793	0.0112*	0.0000	0.0001	0.222	0.8268	0.0000	0.0000	-0.579	0.5690
SHM	0.0028	0.0011	2.444	0.0239*	0.0004	0.0003	1.241	0.2290	-0.0005	0.0002	-2.448	0.0237*

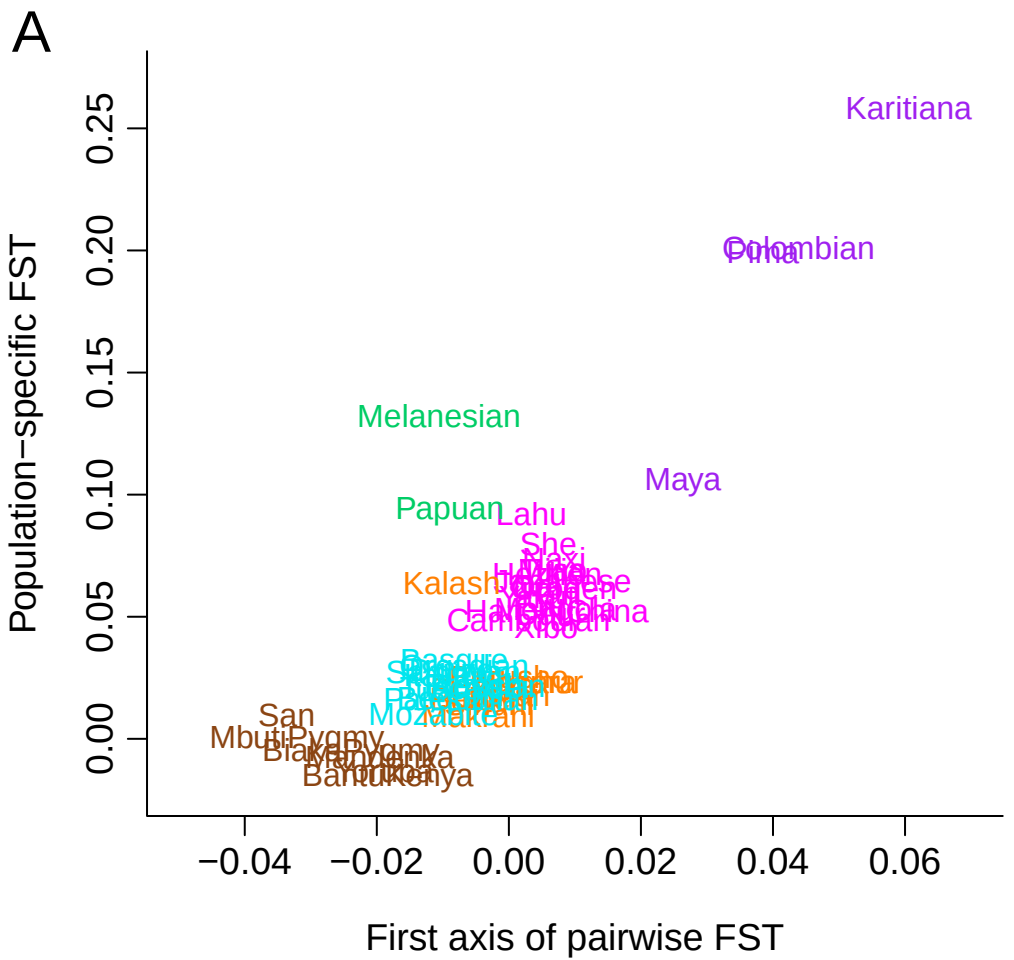
DAY; longest day length (hours), MAT; mean annual temperature (℃), MAP; mean annual precipitation (mm), SHM; summer heat-moisture index, *$p < 0.05$ and **$p < 0.01$

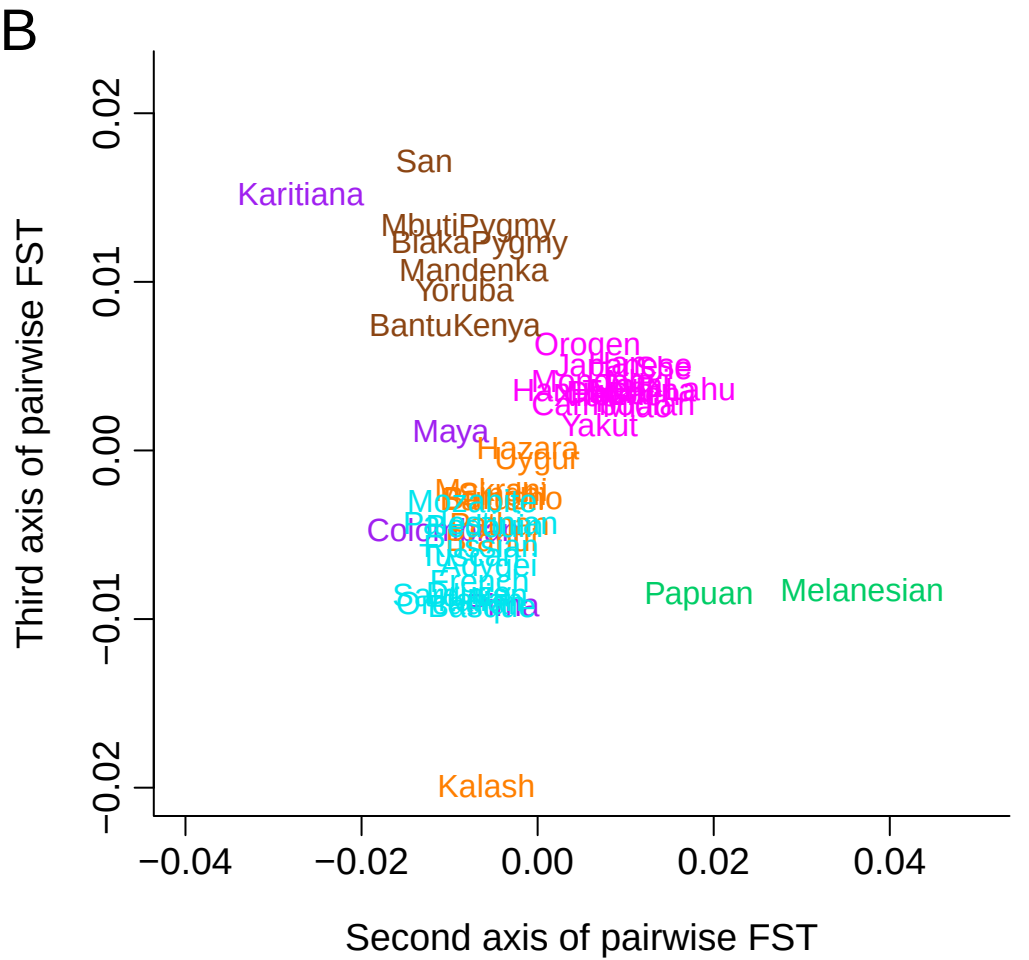

A
Population−specific FST
He
0.00
0.05
0.10
0.15
0.20
0.25
0.55
0.60
0.65
0.70
0.75
0.80
Karitiana
Colombian
Pima
Melanesian
Maya
Papuan
Lahu
Xibo
San
MbutiPygmy
BantuKenya
B
Pairwise FST
0.00
0.02
0.04
0.06
0.08
0.10
San
MbutiPygmy
BiakaPygmy
BantuKenya
Yoruba
Mandenka
Melanesian
Papuan
Lahu
Naxi
Cambodian
Dai
She
Japanese
Daur
Mongola
Miao
Tujia
Han−NChina
Han
Yi
Tu
Xibo
Yakut
Hezhen
Oroden
Kalash
Uyour
Hazara
Brahui
Balochi
Makrani
Burusho
Pathan
Sindhi
Mozabite
Basque
Sardinian
Orcadian
Russian
Adygei
Italian
French
Tuscan
Druze
Palestinian
Bedouin
Karitiana
Pima
Colombian
Maya

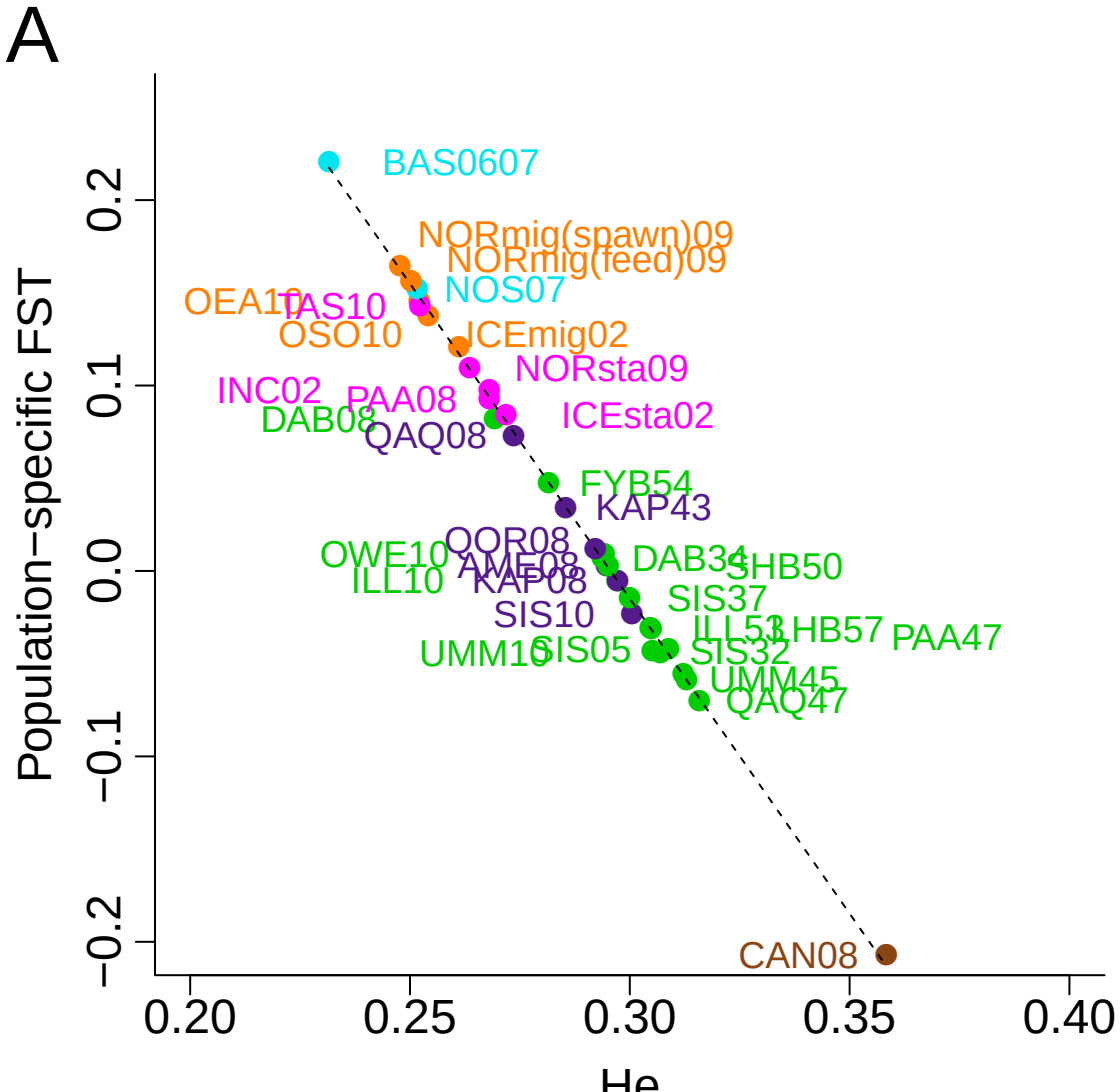
C

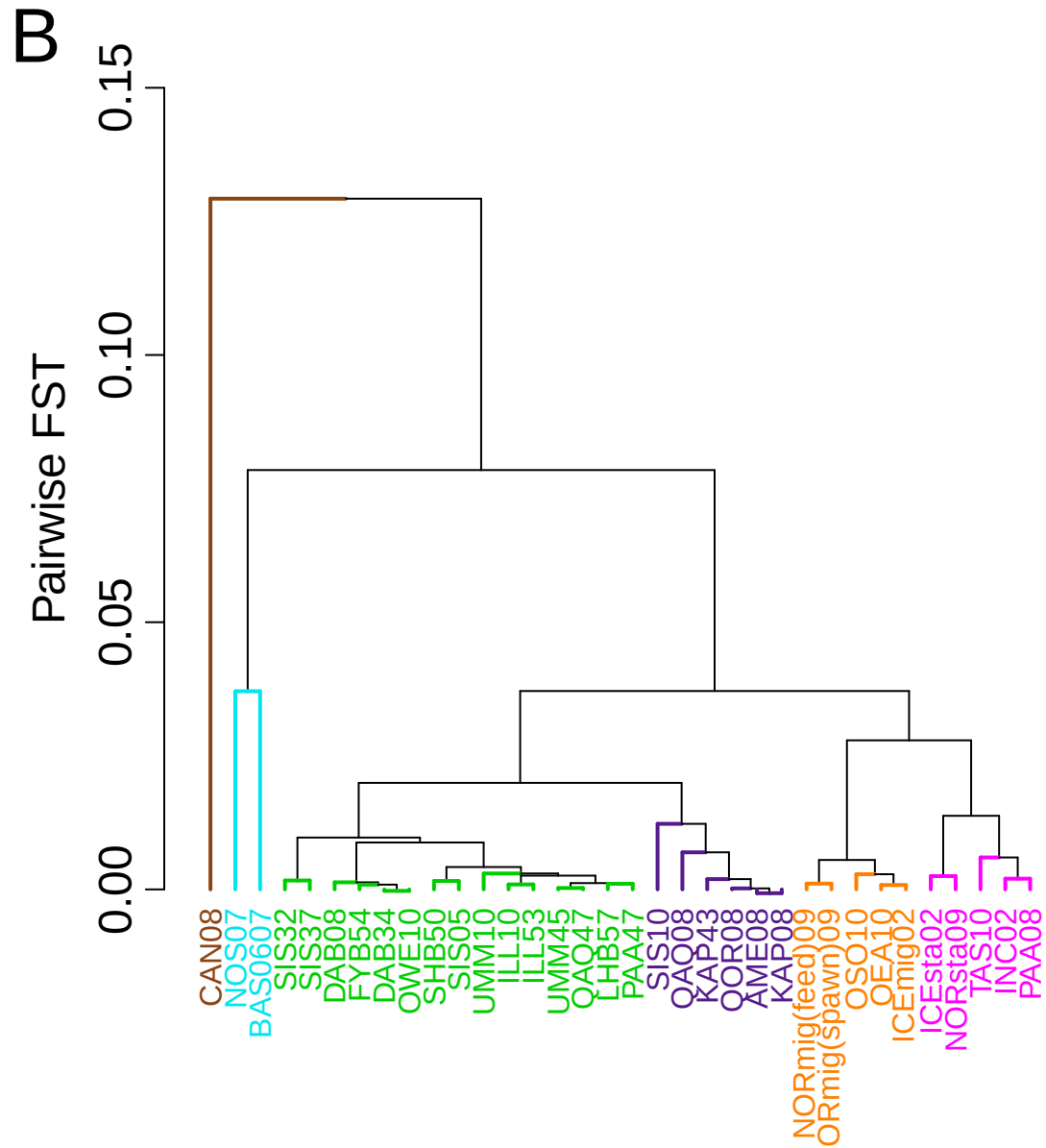

Orcadian
Russian
Yakut
Adygei
Sardinian
Mozabite
Pima
Maya
Colombian
Karitiana
Mandenka
Yoruba
Biaka
Pygmy
Mbuti
Pygmy
Bantu
Kenya
San
Cambodian
Papuan
Melanesian
50°N
0°
50°S
150°W
100°W
50°W
0°
50°E
100°E
150°E

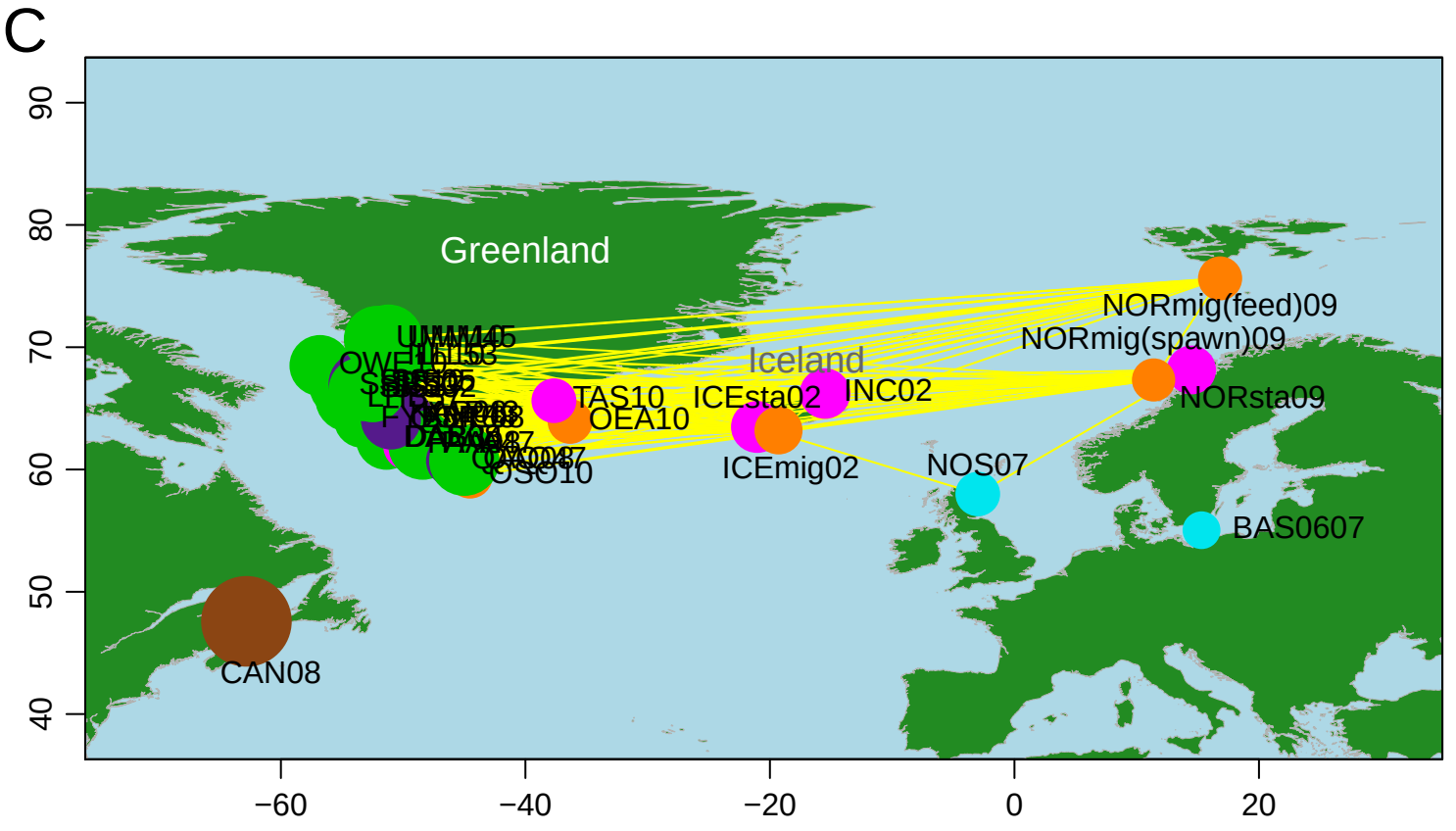

A
WG
B
BB
C
FG
D
First axis of pairwise FST
E
Second axis of pairwise FST
F
Third axis of pairwise FST
50°N
0°
50°S
150°W
100°W
50°W
0°
50°E
100°E
150°E

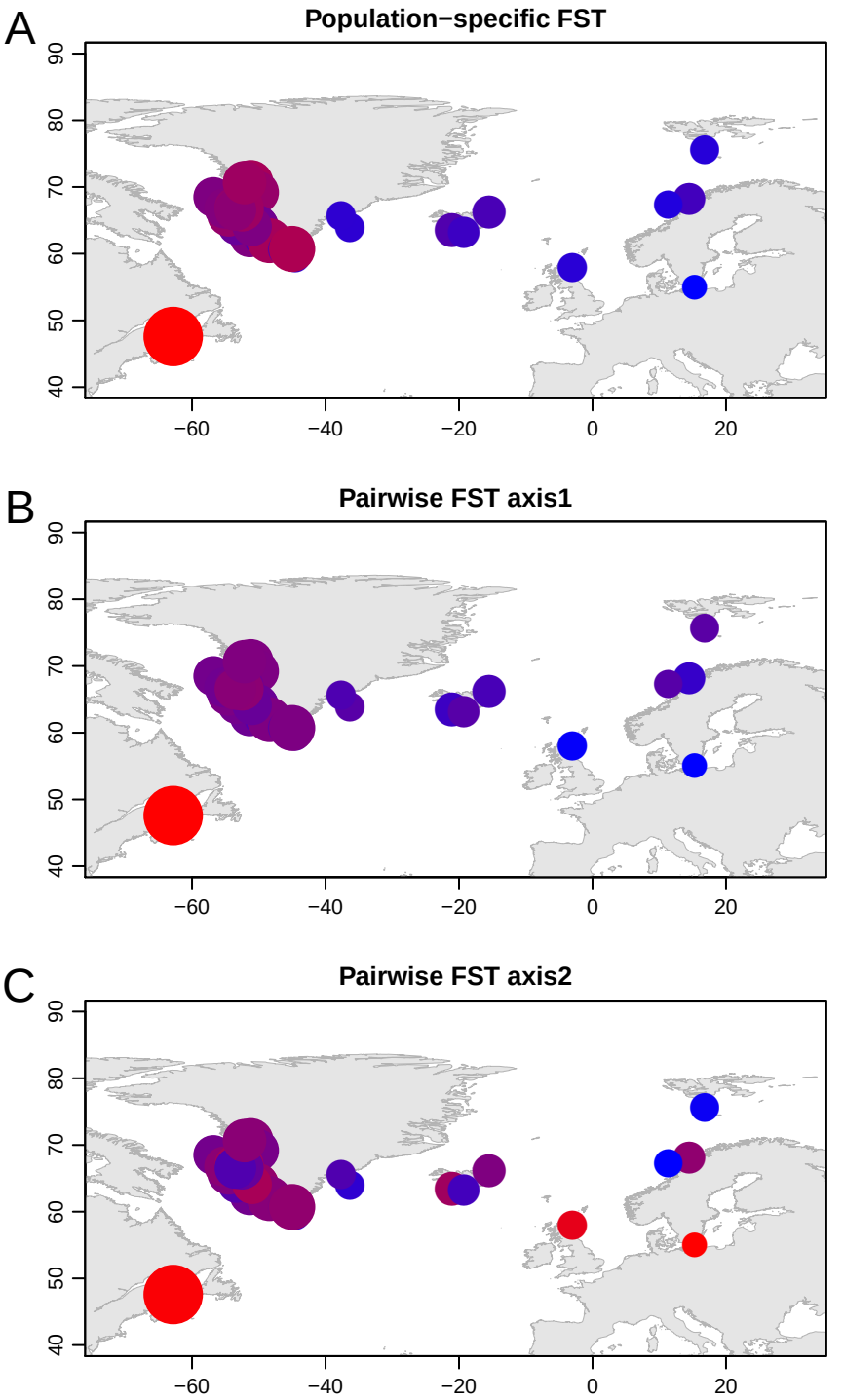

A
Population−specific FST
First axis of pairwise FST
0.00
0.05
0.10
0.15
0.20
0.25
−0.04
−0.02
0.00
0.02
0.04
0.06
Karitiana
Colombian
Pima
Melanesian
Maya
Papuan
Lahu
She
Kalash
Xibo
San
MbutiPygmy
BantuKenya

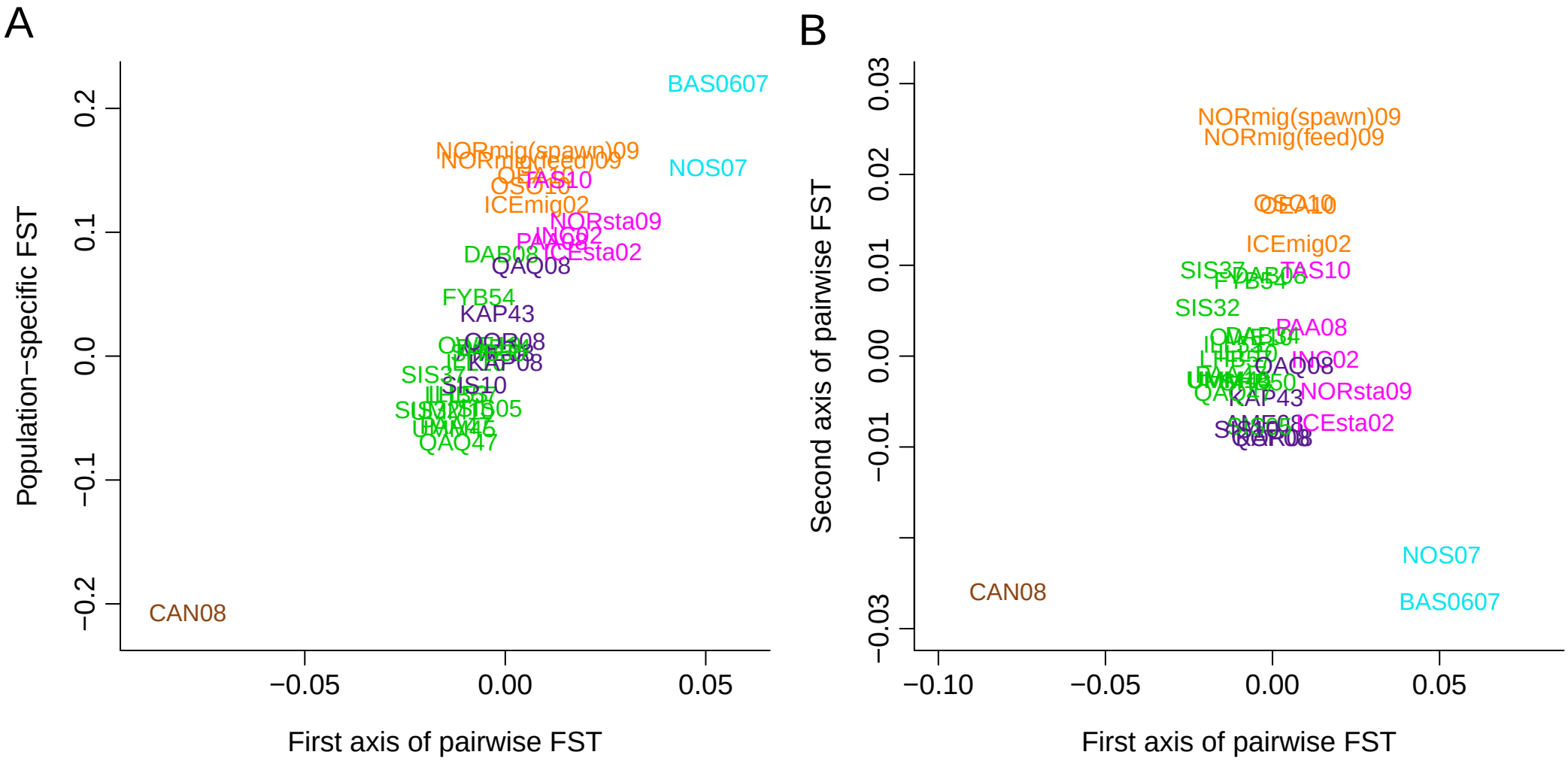

B
Third axis of pairwise FST
Second axis of pairwise FST
0.02
0.01
0.00
−0.01
−0.02
−0.04
−0.02
0.00
0.02
0.04
San
Karitiana
MbutiPygmy
BiakaPygmy
Mandenka
Yoruba
BantuKenya
Orogen
Yakut
Maya
Hazara
Uygur
Papuan
Melanesian
Kalash

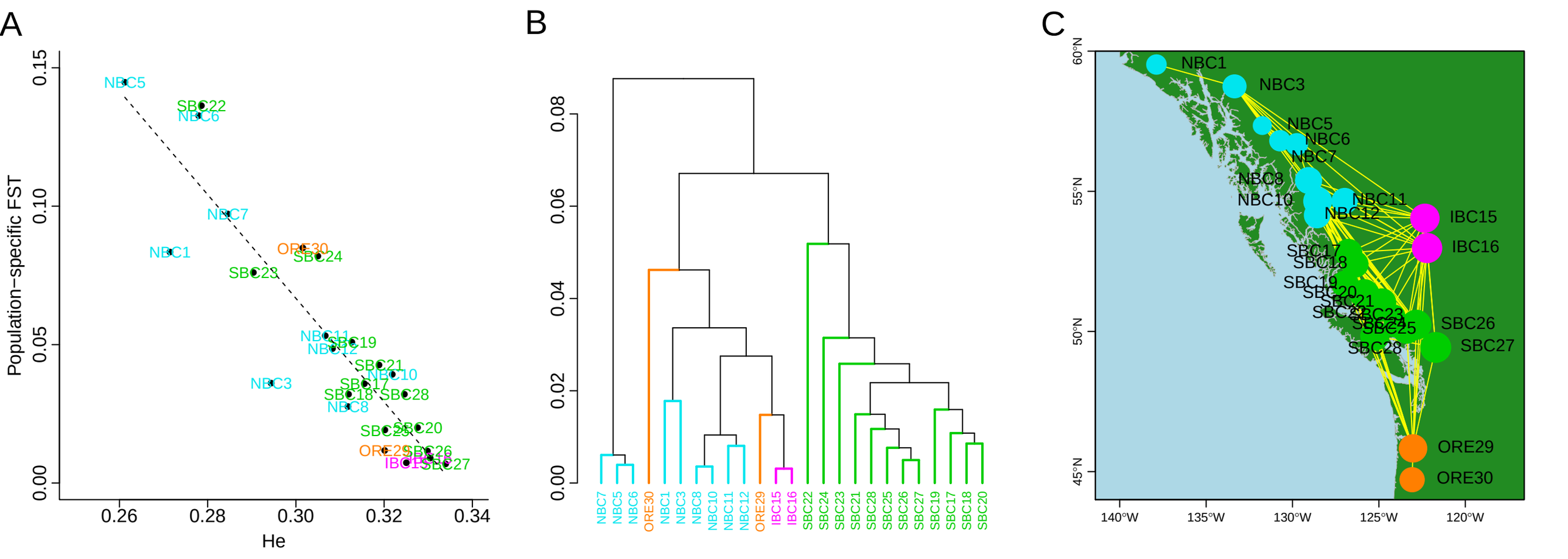

A
Population−specific FST
He
0.20
0.25
0.30
0.35
0.40
−0.2
−0.1
0.0
0.1
0.2
BAS0607
NORmig(spawn)09
NORmig(feed)09
NOS07
OEA10
TAS10
OSO10
ICEmig02
NORsta09
INC02
PAA08
DAB08
QAQ08
ICEsta02
FYB54
KAP43
QOR08
OWE10
AME08
DAB34
SHB50
ILL10
KAP08
SIS37
SIS10
ILL53
LHB57
PAA47
UMM10
SIS05
SIS32
UMM45
QAQ47
CAN08

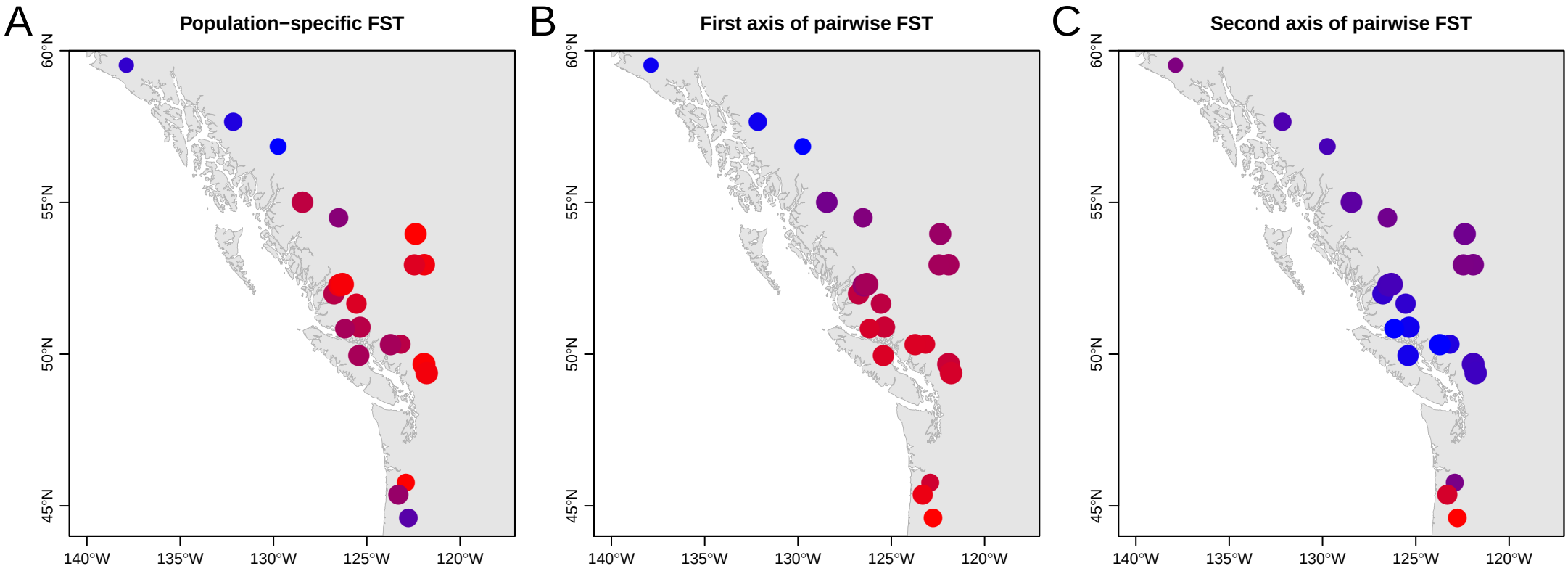

B
Pairwise FST
0.00
0.05
0.10
0.15
CAN08
NOS07
BAS0607
SIS32
SIS37
DAB08
FYB54
DAB34
OWE10
SHB50
SIS05
UMM10
ILL10
ILL53
UMM45
QAQ47
LHB57
PAA47
SIS10
QAQ08
KAP43
QOR08
AME08
KAP08
NORmig(feed)09
NORmig(spawn)09
OSO10
OEA10
ICEmig02
ICEsta02
NORsta09
TAS10
INC02
PAA08

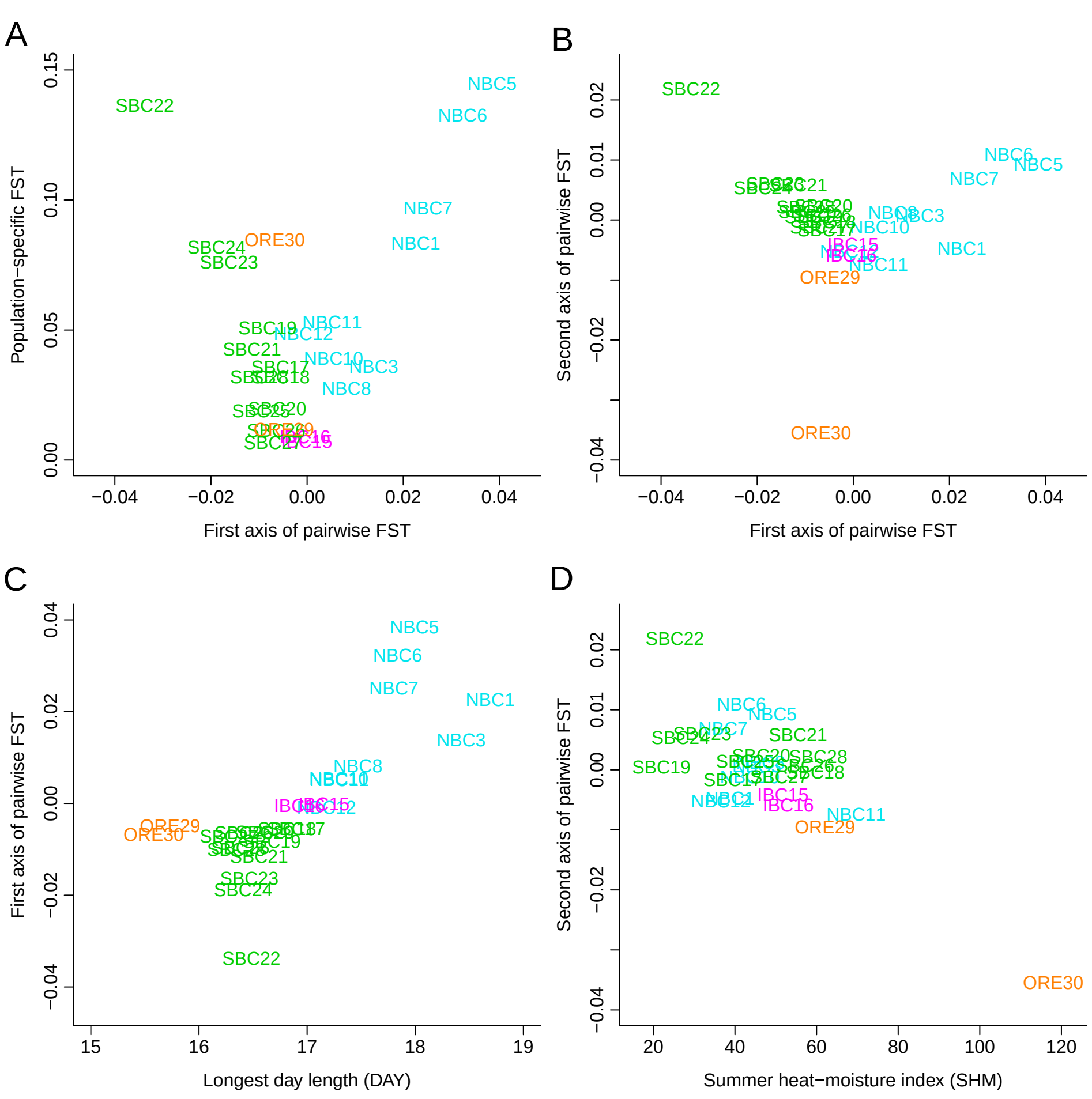

C
Greenland
Iceland
NORmig(feed)09
NORmig(spawn)09
NORsta09
TAS10
OEA10
ICEsta02
INC02
ICEmig02
NOS07
BAS0607
CAN08
90
80
70
60
50
40
−60
−40
−20
0
20


A
Population−specific FST
B
Pairwise FST axis1
C
Pairwise FST axis2
90
80
70
60
50
40
−60
−40
−20
0
20


A
Population−specific FST
First axis of pairwise FST
0.2
0.1
0.0
−0.1
−0.2
−0.05
0.00
0.05
BAS0607
NOS07
NORmig(spawn)09
NORmig(feed)09
OSO10
TAS10
ICEmig02
NORsta09
ING02
PAA08
ICEsta02
DAB08
QAQ08
FYB54
KAP43
KAP08
SIS37
SIS10
QAQ47
CAN08
B
Second axis of pairwise FST
First axis of pairwise FST
0.03
0.02
0.01
0.00
−0.01
−0.03
−0.10
−0.05
0.00
0.05
NORmig(spawn)09
NORmig(feed)09
ICEmig02
TAS10
SIS37
FYB54
SIS32
PAA08
ING02
QAQ08
NORsta09
KAP43
ICEsta02
NOS07
CAN08
BAS0607


A
Population−specific FST
He
0.00
0.05
0.10
0.15
0.26
0.28
0.30
0.32
0.34
NBC5
SBC22
NBC6
NBC7
NBC1
ORE30
SBC24
SBC23
NBC11
SBC19
NBC12
SBC21
NBC10
NBC3
SBC17
SBC18
SBC28
NBC8
SBC20
SBC26
SBC27
B
0.00
0.02
0.04
0.06
0.08
NBC7
NBC5
NBC6
ORE30
NBC1
NBC3
NBC8
NBC10
NBC11
NBC12
ORE29
IBC15
IBC16
SBC22
SBC24
SBC23
SBC21
SBC28
SBC25
SBC26
SBC27
SBC19
SBC17
SBC18
SBC20
C
60°N
55°N
50°N
45°N
140°W
135°W
130°W
125°W
120°W
NBC1
NBC3
NBC5
NBC6
NBC7
NBC8
NBC10
NBC11
NBC12
IBC15
IBC16
SBC17
SBC18
SBC19
SBC20
SBC21
SBC23
SBC25
SBC26
SBC27
SBC28
ORE29
ORE30


A
Population−specific FST
B
First axis of pairwise FST
C
Second axis of pairwise FST
60°N
55°N
50°N
45°N
140°W
135°W
130°W
125°W
120°W


A
Population−specific FST
First axis of pairwise FST
B
Second axis of pairwise FST
First axis of pairwise FST
C
First axis of pairwise FST
Longest day length (DAY)
D
Second axis of pairwise FST
Summer heat−moisture index (SHM)

Supplemental Note

1. Definitions of F_{ST} in the literature
2. G_{ST} is identical to F_{ST}
3. Bias-corrected moment estimators of F_{ST}
3.1 Nei and Chesser's F_{ST} estimator (NC83)
3.2 Weir and Cockerham's F_{ST} estimator (WC84)
3.3 Weir and Goudet's population-specific F_{ST} estimator (WG)

1. Definitions of F_{ST} in the literature

Wright defined F_{ST} as "the correlation between random gametes, drawn from the same subpopulation, relative to the total" and mentioned that "the inbreeding coefficient from matching random lines of ancestry is the type of F_{ST}" (Wright 1951, p. 328). Wright defined F_{ST} in terms of variance as

$$F_{ST} = \frac{\sigma_{ST}^2}{\bar{p}(1-\bar{p})}, \quad \text{(Eq. S1)}$$

where σ_{ST}^2 is the variance of the distribution of allele frequencies in a meta-population described as a beta distribution (two alleles), namely, $\text{Beta}\,(4Nm\bar{p}, 4Nm(1-\bar{p}))$, where $\bar{p}$ is the mean allele frequency and $\sigma_{ST}^2 = \bar{p}(1-\bar{p})/(4Nm+1)$ (Wright 1931, 1965; Rousett 2004). Substituting the latter expression into Equation S1 gives

$$F_{ST} = \frac{1}{4Nm+1} = \frac{1}{\theta+1}$$

as expressed by the scale parameter of the beta distribution ($\theta = \alpha + \beta$) (Wright 1951). F_{ST} has been defined as the ratio of the between-population variance to the total variance of allele frequencies (Cockerham 1973; Weir and Cockerham 1984; Excoffier 2001; Balloux and Lugon-Moulin 2002; Holsinger and Weir 2009), namely,

$$F_{\mathrm{ST}} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})},$$

which is of the same form as Equation S1.

Nei (1973) proposed the G_{ST} measure to explicitly formulate Wright's F-statistics using gene diversities allowing multiple alleles. G_{ST} is defined as the ratio of between-population heterozygosity to total heterozygosity, that is,

$$G_{\mathrm{ST}} = \frac{H_T - H_S}{H_T}, \qquad \text{(Eq. S2)}$$

where H_T is total heterozygosity and H_S is within-population heterozygosity. Nei (1977) also defined Wright's fixation indices as follows:

$$F_{\mathrm{IS}} = \frac{H_S - H_0}{H_S}$$

$$F_{\mathrm{IT}} = \frac{H_T - H_0}{H_T}$$

$$F_{\mathrm{ST}} = \frac{H_T - H_S}{H_T}.$$

In these equations, $H_0 = 1 - \sum_{u=1}^{m} P_{uu}$, $H_S = 1 - \sum_{u=1}^{m} \overline{p_u^2}$, and $H_T = 1 - \sum_{u=1}^{m} \bar{p}_u^2$. Here, $P_{uu} = \sum_{i=1}^{r} w_i P_{iuu}$, $\overline{p_u^2} = \sum_{i=1}^{r} w_i p_{iu}^2$ and $\bar{p}_u^2 = (\sum_{i=1}^{r} w_i p_u)^2$, where w_i is the relative size of population i with $\sum_{i=1}^{r} w_i = 1$. Since the relative size of a population is not known in most cases, Nei and Chesser (1983) suggested an equal weight: $w_i = \frac{1}{r}$. In this setting, $P_{uu} = \frac{1}{r}\sum_{i=1}^{r} P_{iuu}$, $\overline{p_u^2} = \frac{1}{r}\sum_{i=1}^{r} p_{iu}^2$ and $\bar{p}_u^2 = \left(\frac{1}{r}\sum_{i=1}^{r} p_u\right)^2$. Thus, apparently,

$$G_{\mathrm{ST}} \equiv F_{\mathrm{ST}}.$$

Crow and Aoki (1984) introduced the concept of gene identity and defined G_{ST} as

$$G_{\mathrm{ST}} = \frac{F_0 - \bar{F}}{1 - \bar{F}}.$$

Here, F_0 is "the probability that two homologous genes drawn at random from a group are the same (identical in state)," and $\bar{F}$ is "the probability for genes drawn at random from the entire population." Slatkin (1991) defined F_{ST} in terms of probabilities of identity:

$$F_{\mathrm{ST}} = \frac{f_0 - \bar{f}}{1 - \bar{f}} ,$$

where "f_0 is the probability of identity of two genes sampled from the same subpopulation and $\bar{f}$ is the probability of identity of two genes sampled at random from the collection of subpopulations being considered."

Rousset (2001) defined F_{ST} as follows:

$$F_{\mathrm{ST}} = \frac{Q_{\mathrm{w}} - Q_{\mathrm{b}}}{1 - Q_{\mathrm{b}}} . \qquad \text{(Eq. S3)}$$

Here, Q_{w} is "the probabilities of identity of genes from adults sampled within a deme, and Q_{b} is the probabilities of identity of genes from different demes." Rousset (2004) noted that "$H_i = 1 - Q.$ is known as gene diversity or heterozygosity." Excoffier (2001) also formulated the definition

$$F_{\mathrm{ST}} = \frac{Q_{\mathrm{S}} - Q_{\mathrm{T}}}{1 - Q_{\mathrm{T}}} ,$$

where Q_{S} and Q_{T} are "the probabilities that two genes within subpopulations and within the total population are identical, respectively." Although definitions differ, Q_{b} is "the probability of identity between genes in two different subpopulations" (Rousset 2002). Q_{b} is calculated for all combinations of population pairs; therefore, Q_{b} would apply to the total population (see p. 2089 in Weir and Goudet 2017). Indeed, $1 - Q_{\mathrm{b}}$ represents the total variance component (Rousset 2008). Therefore, $Q_{\mathrm{b}} = Q_{\mathrm{T}}$. Karhunen and Ovaskainen (2012) also defined a coancestry coefficient in line with the coalescent-based definition of F_{ST} (Rousset 2004):

$$F_{\mathrm{ST}} = \frac{\theta^W - \theta^B}{1 - \theta^B}.$$

In the above equation, θ^W is the average coancestry within subpopulations, and θ^B is that between subpopulation pairs. The definitions of F_{ST} and G_{ST} are thus identical, and

$$F_{\mathrm{ST}} = G_{\mathrm{ST}} = \frac{\text{between heterozygosity}}{\text{total heterozygosity}}.$$

The values of gene identity (*e.g.*, Q_{S} and Q_{T}) can differ, however, depending on spatial and mutation models and weighting schemes (*e.g.*, Crow and Aoki 1984; Slatkin 1991; Cockerham and Weir 1993; Rousset 2001, 2002, 2004, 2008), as stated by Excoffier (2001).

2. G_{ST} is identical to F_{ST}

In this section, we revisit Kitada *et al.* (2017) and describe the relationship between the definitions of Wright's F_{ST} and Nei's G_{ST} using notations consistent with those of Weir and Hill (2002): i for populations ($i = 1, \ldots, r$), u for alleles ($u = 1, \ldots, m$) and l for loci ($l = 1, \ldots, L$).

Nei (1973) first defined the coefficient of gene differentiation known as G_{ST}, which is a multi-allelic analog of F_{ST} among a finite number of subpopulations (Balloux and Lugon-Moulin 2002). Nei's G_{ST} formula is the ratio of between-population heterozygosity to total heterozygosity (Equation S2). At a single locus, H_{T} (total heterozygosity) and H_{S} (within-population heterozygosity) are

$$H_S = 1 - \frac{1}{r}\sum_{i=1}^{r}\sum_{u=1}^{m} p_{iu}^2 \qquad \text{(Eq. S4)}$$

$$H_T = 1 - \sum_{u=1}^{m}\left(\frac{1}{r}\sum_{i=1}^{r} p_{iu}\right)^2. \quad \text{(Eq. S5)}$$

The F_{ST} over a meta-population is often called global F_{ST}. When estimating population structure, F_{ST} values between pairs of population samples (pairwise F_{ST}) are routinely used in addition to global F_{ST}. Here, we focus on F_{ST} between population pairs (pairwise F_{ST}) at a single locus. Let population i ($i = 1, 2$) have the allele frequencies p_{iu}. In such cases, the denominator of G_{ST} is

$$H_T = 1 - \frac{1}{4}\sum_{u=1}^{m}(p_{1u} + p_{2u})^2 = 1 - \sum_{u=1}^{m}\left(\frac{p_{1u} + p_{2u}}{2}\right)^2 = 1 - \sum_{u=1}^{m}\bar{p}_u^2 .$$

This is decomposed as

$$H_T = 1 - \sum_{u=1}^{m}\bar{p}_u^2 = 1 - \left(\sum_{u=1}^{m-1}\bar{p}_u^2 + \left(1 - \sum_{u=1}^{m-1}\bar{p}_u\right)^2\right) = 2\left[\sum_{u=1}^{m-1}\bar{p}_u\,(1 - \bar{p}_u) - \sum_{u<u'}^{m-1}\bar{p}_u\bar{p}_{u'}\right].$$

As the second term can be written as

$$-2\sum_{u<u'}^{m-1}\bar{p}_u\bar{p}_{u'} = \bar{p}_m(1 - \bar{p}_m) - \sum_{u=1}^{m-1}\bar{p}_u(1 - \bar{p}_u),$$

we have

$$2\left[\sum_{u=1}^{m-1}\bar{p}_u\,(1 - \bar{p}_u) - \sum_{u<u'}^{m-1}\bar{p}_u\bar{p}_{u'}\right]$$

$$= 2\sum_{u=1}^{m-1}\bar{p}_u(1 - \bar{p}_u) + \bar{p}_m(1 - \bar{p}_m) - \sum_{u=1}^{m-1}\bar{p}_u(1 - \bar{p}_u) = \sum_{u=1}^{m}\bar{p}_u(1 - \bar{p}_u)\,.$$

Therefore,

$$H_T = 1 - \sum_{u=1}^{m}\bar{p}_u^2 = \sum_{u=1}^{m}\bar{p}_u(1 - \bar{p}_u)\,.$$

The numerator of G_{ST} is

$$H_T - H_S = 1 - \frac{1}{4}\sum_{u=1}^{m}(p_{1u} + p_{2u})^2 - \left[1 - \frac{1}{2}\sum_{u=1}^{m}(p_{1u}^2 + p_{2u}^2)\right] = \frac{1}{2}\sum_{u=1}^{m}(p_{1u}^2 + p_{2u}^2)$$

$$- \frac{1}{4}\sum_{u=1}^{m}(p_{1u} + p_{2u})^2 = \frac{1}{4}\sum_{u=1}^{m}(p_{1u} - p_{2u})^2. \quad \text{(Eq. S6)}$$

For the numerator, we consider the general case of variance of random variables $V[x] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The variance can be written as

$$V[x] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{j=1}^{n}x_i)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}(x_i - x_j)\right)$$

$$= \frac{1}{(n-1)n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(x_i - x_j)(x_i - x_k)$$

$$= \frac{1}{(n-1)n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(x_i^2 - x_i x_k - x_i x_j - x_j x_k)$$

$$= \frac{1}{(n-1)n^2}\left[n^2\sum_{i=1}^{n}x_i^2 - n\left(\sum_{i=1}^{n}x_i^2 + \sum_{i\neq k}^{n}x_i x_k\right) - n\left(\sum_{i=1}^{n}x_i^2 + \sum_{i\neq j}^{n}x_i x_j\right)\right.$$

$$\left. + n\left(\sum_{i=1}^{n}x_i^2 + \sum_{j\neq k}^{n}x_j x_k\right)\right] = \frac{1}{(n-1)n}\left[(n-1)\sum_{i=1}^{n}x_i^2 - \sum_{i\neq j}^{n}x_i x_j\right]$$

$$= \frac{1}{2(n-1)n}\sum_{i\neq j}^{n}(x_i - x_j)^2.$$

In the case of pairwise F_{ST} $(n = 2)$, the numerator is expressed as

$$V[x] = \frac{1}{4}(x_1 - x_2)^2 .$$

The numerator of G_{ST} (Equation S6) is thus the variance of allele frequencies between subpopulations. From Equations S4 and S5, we have

$$G_{\mathrm{ST}} = \frac{\frac{1}{4}\sum_{u=1}^{m}\left(p_{1,u} - p_{2,u}\right)^2}{\sum_{u=1}^{m}\bar{p}_u(1-\bar{p}_u)} = \frac{\sigma_{\mathrm{ST}}^2}{\sum_{u=1}^{m}\bar{p}_u(1-\bar{p}_u)}. \quad \text{(Eq. S7)}$$

For bi-allelic cases, $H_T - H_S = \frac{1}{2}(p_1 - p_2)^2$ and $H_T = 2\bar{p}(1-\bar{p})$, and

$$G_{\mathrm{ST}} = \frac{\frac{1}{4}(p_1 - p_2)^2}{\bar{p}(1-\bar{p})} = \frac{\sigma_{\mathrm{ST}}^2}{\bar{p}(1-\bar{p})}. \quad \text{(Eq. S8)}$$

Equation S8 is equal to Equation S1, and we have

$$G_{\mathrm{ST}} \equiv F_{\mathrm{ST}}.$$

From Equations S6 and S7, the definition of pairwise F_{ST} (pwF_{ST}) at a single locus for multi-allelic cases is

$$\mathrm{pw}F_{\mathrm{ST}} = \frac{\frac{1}{4}\sum_{u=1}^{m}(p_{1u} - p_{2u})^2}{\sum_{u=1}^{m}\bar{p}_u(1-\bar{p}_u)} = \frac{\sum_{u=1}^{m}(p_{1u} - p_{2u})^2}{4\sum_{u=1}^{m}\bar{p}_u(1-\bar{p}_u)}.$$

For bi-allelic cases ($m = 2$), it is

$$\mathrm{pw}F_{\mathrm{ST}} = \frac{\frac{1}{2}(p_1 - p_2)^2}{2\bar{p}(1-\bar{p})} = \frac{(p_1 - p_2)^2}{4\bar{p}(1-\bar{p})}.$$

3. Bias-corrected moment estimators of F_{ST}

3.1 Nei and Chesser's F_{ST} estimator (NC83)

F_{ST} is the ratio of between to total variance. The estimator is therefore a ratio estimator and consequently biased. To correct the bias, Nei and Chesser (1983) derived the unbiased estimators of H_S and H_T in diploid populations. This estimator (hereafter, NC83) assumes that samples (n_i individuals) are randomly chosen from a set of fixed subpopulations ($i = 1, \ldots, r$). The gene diversities in population i are written based on the

homozygote genotype frequencies P_{iuu} and allele frequencies p_{iu}, respectively given as

$$H_{0i} = 1 - \sum_{u=1}^{m} P_{iuu}$$

$$H_{Si} = 1 - \sum_{u=1}^{m} p_{iu}^2 .$$

Because observed genotype frequencies are unbiased, H_{0i} is unbiasedly estimated by

$$\widehat{H}_{0i} = 1 - \sum_{u=1}^{m} \tilde{P}_{iuu} ,$$

where $\tilde{p}_{iuu}$ denotes observed homozygote genotype frequencies. For estimating H_{Si}, Nei and Chesser (1983) took the expectation of $\tilde{p}_{iu}^2$, since $\tilde{p}_{iu}^2$ is biased, as

$$E[\tilde{p}_{iu}^2] = p_{iu}^2 + \frac{P_{iuu}}{n_i} + \sum_{u \neq u}^{m} \frac{P_{iuu'}}{4n_i} - \frac{p_{iu}^2}{n_i} .$$

Those authors then calculated the expectation of observed gene diversity in population *i*: since $p_{iu} = P_{uu} + \frac{\sum_{u \neq u}^{m} P_{uu'}}{2}$,

$$1 - E\left[\sum\nolimits_{u=1}^{m} \tilde{p}_{iu}^2\right] = 1 - \sum_{u=1}^{m} p_{iu}^2 - \frac{1}{n_i}\sum_{u=1}^{m} P_{iuu} - \frac{1}{2n_i}\sum_{u=1}^{m}(p_{iu} - P_{uu}) + \frac{1}{n_i}\sum_{u=1}^{m} p_{iu}^2$$

$$= 1 - \sum_{u=1}^{m} p_{iu}^2 - \frac{1}{n_i}\sum_{u=1}^{m} P_{iuu} - \frac{1}{2n_i} + \frac{1}{2n_i}\sum_{u=1}^{m} P_{uu} + \frac{1}{n_i}\sum_{u=1}^{m} p_{iu}^2$$

$$= H_{Si}\left(1 - \frac{1}{n_i}\right) + \frac{H_{0i}}{2n_i} . \qquad \text{(NC83 Eq. 6)}$$

Using the method of moments, they obtained the unbiased estimator of gene diversity in population i (H_{Si}):

$$\widehat{H}_{Si} = \frac{n_i}{n_i - 1}\left(1 - \sum\nolimits_{u=1}^{m} \tilde{p}_{iu}^2 - \frac{\widehat{H}_{0i}}{2n_i}\right) . \qquad \text{(NC83 Eq. 7)}$$

Because observed genotype frequencies are unbiased, H_{0i} was unbiasedly estimated by $\widehat{H}_{0i} = 1 - \sum_{u=1}^{m} \tilde{P}_{iuu}$. Thus, they unintentionally derived population-specific gene diversity. Under the assumption that all subpopulations are in Hardy–Weinberg equilibrium, $\widehat{H}_{Si} = \widehat{H}_{0i}$, and we have

$$\widehat{H}_{Si} = \frac{2n_i}{2n_i - 1}\left(1 - \sum_{u=1}^{m} \tilde{p}_{iu}^2\right). \qquad \text{(Eq. S9)}$$

The expectation of the average of observed gene diversity over all populations was similarly derived as

$$1 - E\left[\sum_{u=1}^{m} \overline{\tilde{p}_u^2}\right] = \frac{1}{r}\left[\sum_{i=1}^{r} \left\{H_{Si}\left(1 - \frac{1}{n_i}\right) + \frac{H_{0i}}{2n_i}\right\}\right] = H_S\left(1 - \frac{1}{\tilde{n}}\right) + \frac{H_{0i}}{2\tilde{n}}, \qquad \text{(NC Eq. 8)}$$

where $\overline{\tilde{p}_u^2} = \frac{1}{r}\sum_{i=1}^{r} \tilde{p}_{iu}^2$, and $\tilde{n}$ is the harmonic mean of n_i, namely, $\tilde{n} = \frac{r}{\sum_{i=1}^{r}\frac{1}{n_i}}$. The unbiased moment estimator of gene diversity over all populations was derived as

$$\widehat{H}_S = \frac{\tilde{n}}{\tilde{n} - 1}\left(1 - \sum_{u=1}^{m} \overline{\tilde{p}_u^2} - \frac{\widehat{H}_0}{2\tilde{n}}\right). \qquad \text{(NC83 Eq. 9)}$$

Here, $\widehat{H}_0$ is the unbiased estimator of observed gene diversity based on the homozygote genotype frequencies over all populations:

$$\widehat{H}_0 = 1 - \sum_{u=1}^{m} \bar{\tilde{P}}_{uu}, \qquad \text{(NC83 Eq. 5)}$$

where $\bar{\tilde{P}}_{uu} = \frac{1}{r}\sum_{i=1}^{r} \tilde{P}_{uu}$. To obtain the unbiased estimator of H_{T}, they derived the expectation of observed mean gene diversity over all populations:

$$1 - E\left[\sum_{u=1}^{m} \bar{\tilde{p}}_u^2\right] = H_T - \frac{H_S}{\tilde{n}r} + \frac{H_0}{2\tilde{n}r}, \qquad \text{(NC83 Eq. 10)}$$

where $\bar{\tilde{p}}_u = \frac{1}{r}\sum_{i=1}^{r} \tilde{p}_{iu}$. From this equation, they obtained the unbiased moment estimator of H_T as

$$\widehat{H}_T = 1 - \sum_{u=1}^{m} \bar{\tilde{p}}_u^2 + \frac{\widehat{H}_S}{\tilde{n}r} - \frac{\widehat{H}_0}{2\tilde{n}r}. \quad \text{(NC83 Eq. 11)}$$

Under the assumption that all subpopulations are in Hardy–Weinberg equilibrium, $H_0 = H_S$. In this situation, the estimators are a function of allele frequencies only at a single locus:

$$\widehat{H}_S = \frac{2\tilde{n}}{2\tilde{n}-1}\left(1 - \sum_{u=1}^{m} \overline{\tilde{p}_u^2}\right) \quad \text{(NC83 Eq. 15)}$$

$$\widehat{H}_T = 1 - \sum_{u=1}^{m} \bar{\tilde{p}}_u^2 + \frac{\widehat{H}_S}{2\tilde{n}r}. \quad \text{(NC83 Eq. 16)}$$

3.2 Weir and Cockerham's F_{ST} estimator (WC84)

Weir and Cockerham (1984) derived the moment estimator of F_{ST} as a coancestry coefficient denoted by θ, which we call θ_{WC} according to Weir and Goudet (2017):

$$\theta_{WC} = \frac{a}{a+b+c}. \quad \text{(WC84 Eq. 1)}$$

This moment estimator of F_{ST} is the ratio of observed unbiased variance components for an allele: a for between-subpopulation components, b for those between individuals within subpopulations and c for those between gametes within individuals, given as

$$a = \frac{\bar{n}}{n_c}\left\{s^2 - \frac{1}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{1}{4}\bar{h}\right]\right\} \quad \text{(WC84 Eq. 2)}$$

$$b = \frac{\bar{n}}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{2\bar{n}-1}{4\bar{n}}\bar{h}\right] \quad \text{(WC84 Eq. 3)}$$

$$c = \frac{1}{2}\bar{h}, \quad \text{(WC84 Eq. 4)}$$

where $\bar{n} = \frac{1}{r}\sum_{i=1}^{r} n_i$, $n_c = \frac{r\bar{n} - \frac{1}{r\bar{n}}\sum_{i=1}^{r} n_i^2}{r-1}$, $\bar{p} = \frac{1}{r\bar{n}}\sum_{i=1}^{r} n_i\tilde{p}_i$, $s^2 = \frac{1}{r-1}\sum_{i=1}^{r}\frac{n_i}{\bar{n}}(\tilde{p}_i - \bar{p})^2$, $\bar{h} =$

$\frac{1}{r}\sum_{i=1}^{r}\frac{n_i}{\bar{n}}\tilde{h}_i$ and $\tilde{h}_i$ is the observed heterozygote frequency in a diploid population. For multiple alleles at a locus, the combined ratio estimator is given in Weir and Cockerham (1984) as

$$\theta_{\mathrm{WC}} = \frac{\sum_{u=1}^{m} a_u}{\sum_{u=1}^{m}(a_u + b_u + c_u)}.$$

The combined ratio estimator over all loci ("ratio of averages") is

$$\theta_{\mathrm{WC}} = \frac{\sum_{l=1}^{L}\sum_{u=1}^{m} a_{lu}}{\sum_{l=1}^{L}\sum_{u=1}^{m}(a_{lu} + b_{lu} + c_{lu})}. \qquad \text{(WC84 Eq. 10)}$$

Here, we formulate the asymptotic variance of θ_{WC}. Let $\bar{y}$ be the numerator and $\bar{x}$ be the denominator of θ_{WC} over loci:

$$\theta_{\mathrm{WC}} = \frac{\frac{1}{L}\sum_{l=1}^{L} a_l}{\frac{1}{L}\sum_{l=1}^{L}(a_l + b_l + c_l)} = \frac{\bar{y}}{\bar{x}}.$$

By Taylor series expansion for the first term, we have the asymptotic variance as

$$V[\theta_{\mathrm{WC}}] \simeq \frac{\bar{y}^2}{\bar{x}^4}V[\bar{x}] + \frac{1}{\bar{x}^2}V[\bar{y}] - 2\frac{\bar{y}}{\bar{x}^3}Cov[\bar{x},\bar{y}], \qquad \text{(Eq. S10)}$$

where the variance and covariance components are calculated by

$V[\bar{x}] = \frac{1}{L(L-1)}\sum_{l=1}^{L}(x_l - \bar{x})^2$, $V[\bar{y}] = \frac{1}{L(L-1)}\sum_{l=1}^{L}(y_l - \bar{y})^2$ and

$Cov[\bar{x},\bar{y}] = \frac{1}{L(L-1)}\sum_{l=1}^{L}(x_l - \bar{x})(y_l - \bar{y})$.

3.3 Weir and Goudet's population-specific F_{ST} estimator (WG)

Weir and Goudet (2017) defined population-specific F_{ST} as

$$\mathrm{ps}F_{\mathrm{ST}}^{i} = \beta_{\mathrm{WT}}^{i} = \frac{\theta_W^i - \theta^B}{1 - \theta^B}, \qquad \text{(Eq. S11)}$$

where θ_W^i is the identical-by-descent (ibd) probability of two alleles drawn from population

i, and θ^B is the average of ibd probabilities for alleles from different populations. The definition refers to the "probability two alleles drawn from population i are ibd, relative to the probability an allele drawn from one population is ibd to an allele drawn from another population." Equation S11 corresponds to allele-based F_{ST}, and the average over subpopulations is the usual "population-average F_{ST}" (Weir and Goudet 2017):

$$F_{\mathrm{ST}} = \beta_{\mathrm{WT}} = \frac{\theta^W - \theta^B}{1 - \theta^B},$$

as given previously (Equation S3) (*e.g.*, Rousset 2004; Karhunen and Ovaskainen 2012). Weir and Goudet's other definition of population-specific F_{ST} is

$$\mathrm{ps}F_{\mathrm{ST}}^i = \beta_{\mathrm{ST}}^i = \frac{\theta_S^i - \theta^B}{1 - \theta^B},$$

which uses θ_S^i instead of θ_W^i. According to Weir and Goudet, the use of θ_S^i "for within-population pairs of alleles indicates that we are referring to genotypes, whereas, if we work only with alleles, we write θ_W^i." Averaging over population, they gave

$$F_{\mathrm{ST}} = \beta_{\mathrm{ST}} = \frac{\theta^S - \theta^B}{1 - \theta^B}.$$

"For random mating populations, there will be no need for distinction between β_{ST} and β_{WT}" (Weir and Goudet 2017).

Weir and Goudet (2017) derived β_{WT}, the bias-corrected moment estimator of population-specific F_{ST} when only allele frequencies are used, as

$$\mathrm{ps}\hat{F}_{\mathrm{ST}}^i = \hat{\beta}_{\mathrm{WT}}^i = \frac{\widetilde{M}_w^i - \widetilde{M}^B}{1 - \widetilde{M}^B}.$$

$\widetilde{M}_W^i$ is the unbiased within-population matching of two distinct alleles of population *i:*

$$\widetilde{M}_W^i = \frac{2n_i}{2n_i - 1}\sum_{u=1}^{m} \tilde{p}_{iu}^2 - \frac{1}{2n_i - 1},$$

where n_i is the sample size (number of individuals) taken from population i, and $\tilde{p}_{iu}$ is the observed frequency of allele u. We note that $1-\widetilde{M}_W^i$ equals to $\widehat{H}_{Si}$ (Equation S9):

$$1-\widetilde{M}_W^i=\frac{2n_i}{2n_i-1}\left(1-\sum\nolimits_{u=1}^{m}\tilde{p}_{iu}^2\right)=\widehat{H}_{Si}\,. \quad \text{(Eq. S11)}$$

$\widetilde{M}^B$ is the between-population-pair matching average over pairs of populations i,i':

$$\widetilde{M}^B=\frac{1}{r(r-1)}\sum_{i=1}^{r}\sum_{i'=1(i'\neq i)}^{r}\widetilde{M}_B^{ii'}.$$

$\widetilde{M}_B^{ii'}$ is the matching of one allele in j,j' individuals taken from each of populations i,i':

$$\widetilde{M}_B^{ii'}=\frac{1}{n_i n_{i'}}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}}\widetilde{M}_{jj'}^{ii'}=\sum_{u=1}^{m}\tilde{p}_{iu}\tilde{p}_{i'u}\,.$$

Average pair-matching of alleles over all populations is therefore calculated from the product of allele frequencies over populations i,i':

$$\widetilde{M}^B=\frac{1}{r(r-1)}\sum_{i=1}^{r}\sum_{i'=1(i'\neq i)}^{r}\sum_{u=1}^{m}\tilde{p}_{iu}\tilde{p}_{i'u}\,.$$

We can write $\mathrm{ps}\widehat{F}_{\mathrm{ST}}^i$ over all loci as

$$\mathrm{ps}\widehat{F}_{\mathrm{ST}}^i=\frac{\frac{1}{L}\sum_{l=1}^{L}(\widetilde{M}_{W,l}^i-\widetilde{M}_l^B)}{\frac{1}{L}\sum_{l=1}^{L}(1-\widetilde{M}_l^B)}=\frac{\bar{y}}{\bar{x}}\,.$$

Equation S10 can be applied for use with the population-specific F_{ST} estimator in the same way, thereby yielding

$$V\left[\mathrm{ps}\widehat{F}_{\mathrm{ST}}^i\right]\simeq\frac{\bar{y}^2}{\bar{x}^2}\left\{\frac{V[\bar{x}]}{\bar{x}^2}+\frac{V[\bar{y}]}{\bar{y}^2}-\frac{2Cov[\bar{x},\bar{y}]}{\bar{x}\bar{y}}\right\}. \quad \text{(Eq. S12)}$$

Supplemental Figures for

Population-specific F_{ST} and Pairwise F_{ST}: History and Environmental Pressure

Shuichi Kitada*, Reiichiro Nakamichi†, and Hirohisa Kishino‡

*Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan

†Japan Fisheries Research and Education Agency, Yokohama 236-8648, Japan

‡Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan

*Correspondence to: kitada@kaiyodai.ac.jp

This PDF file includes Figures S1 to S7.

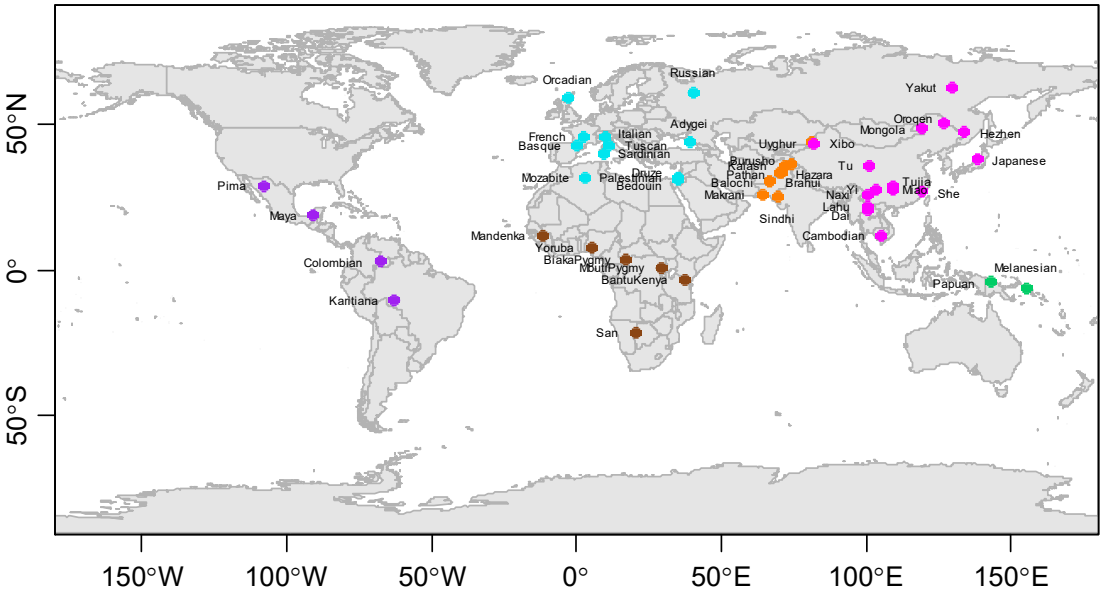


Figure S1 Sampling locations of 51 human populations. Data from Cann *et al.* (2002).

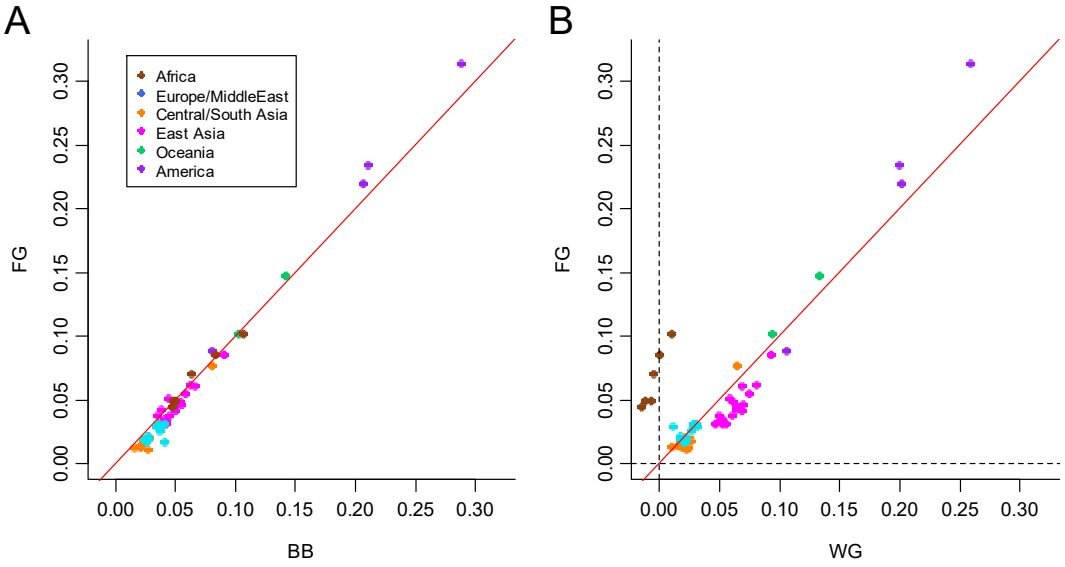


Figure S2 Relationships between different population-specific F_{ST} estimators for 51 human populations. (A) BB (Beaumont and Balding 2004) vs. FG (Foll and Gaggiotti 2006) ($R^2 = 0.9885, p = 2.2 \times 10^{-16}$). (B) WG (Weir and Goudet 2017) vs. FG ($R^2 = 0.8630, p = 2.2 \times 10^{-16}$). Data from Rosenberg *et al.* (2002).

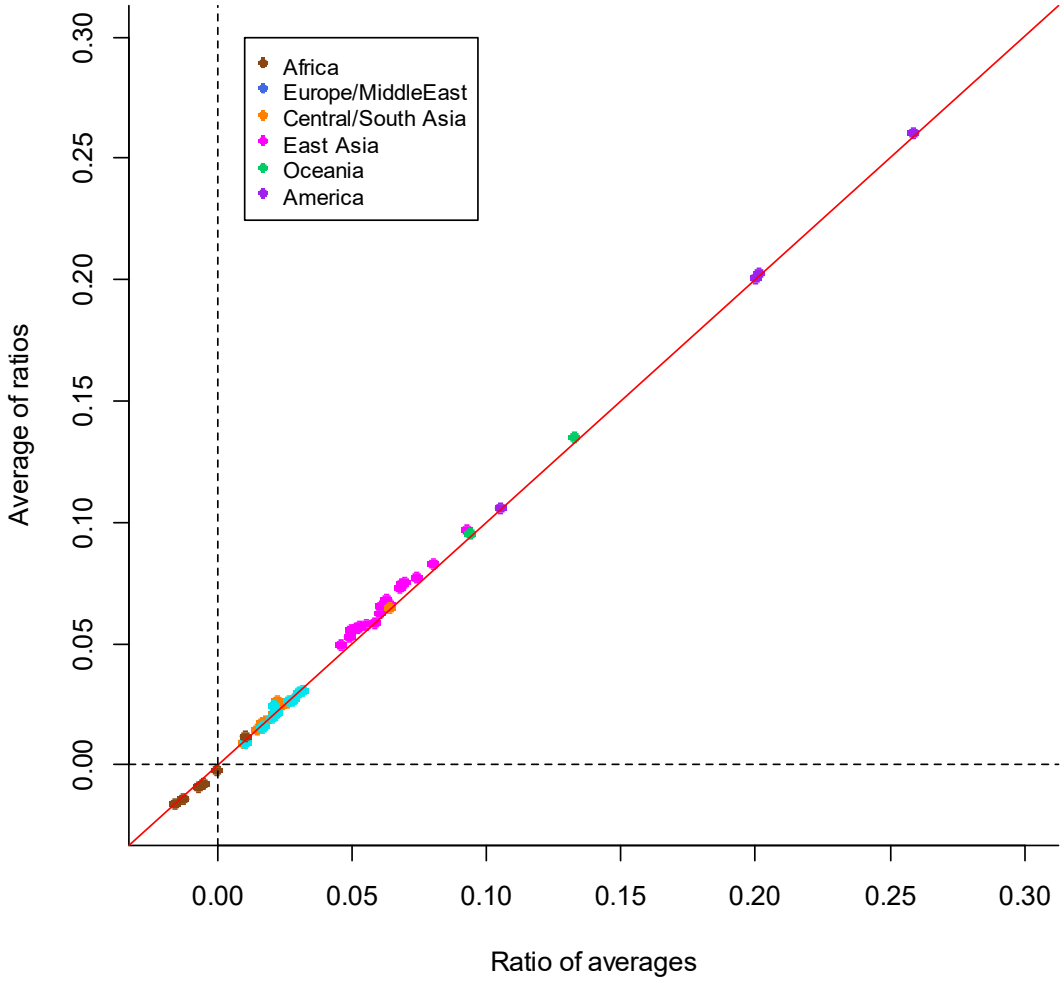


Figure S3 Relationship between the "ratio of averages" and "average of ratios" of population-specific F_{ST} (Weir and Goudet 2017) for 51 human populations ($R^2 = 0.9989, p = 2.2 \times 10^{-16}$). Data from Rosenberg *et al.* (2002).

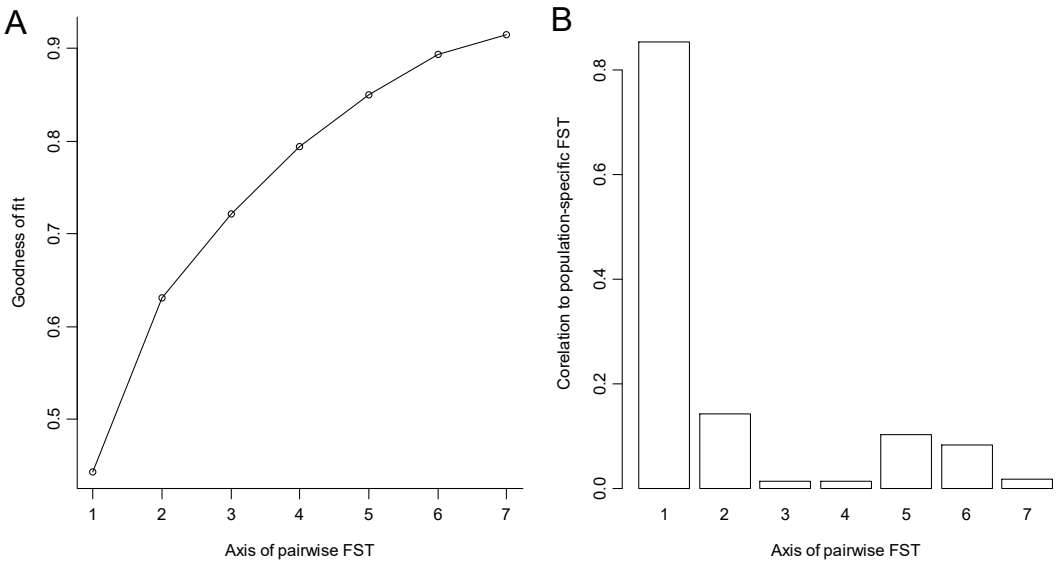


Figure S4 Multi-dimensional scaling (MDS) axes of pairwise F_{ST} for 51 human populations. (A) Goodness of fit. (B) Correlation between the MDS axis of pairwise F_{ST} and population-specific F_{ST} (Weir and Goudet 2017). Data from Rosenberg *et al.* (2002).

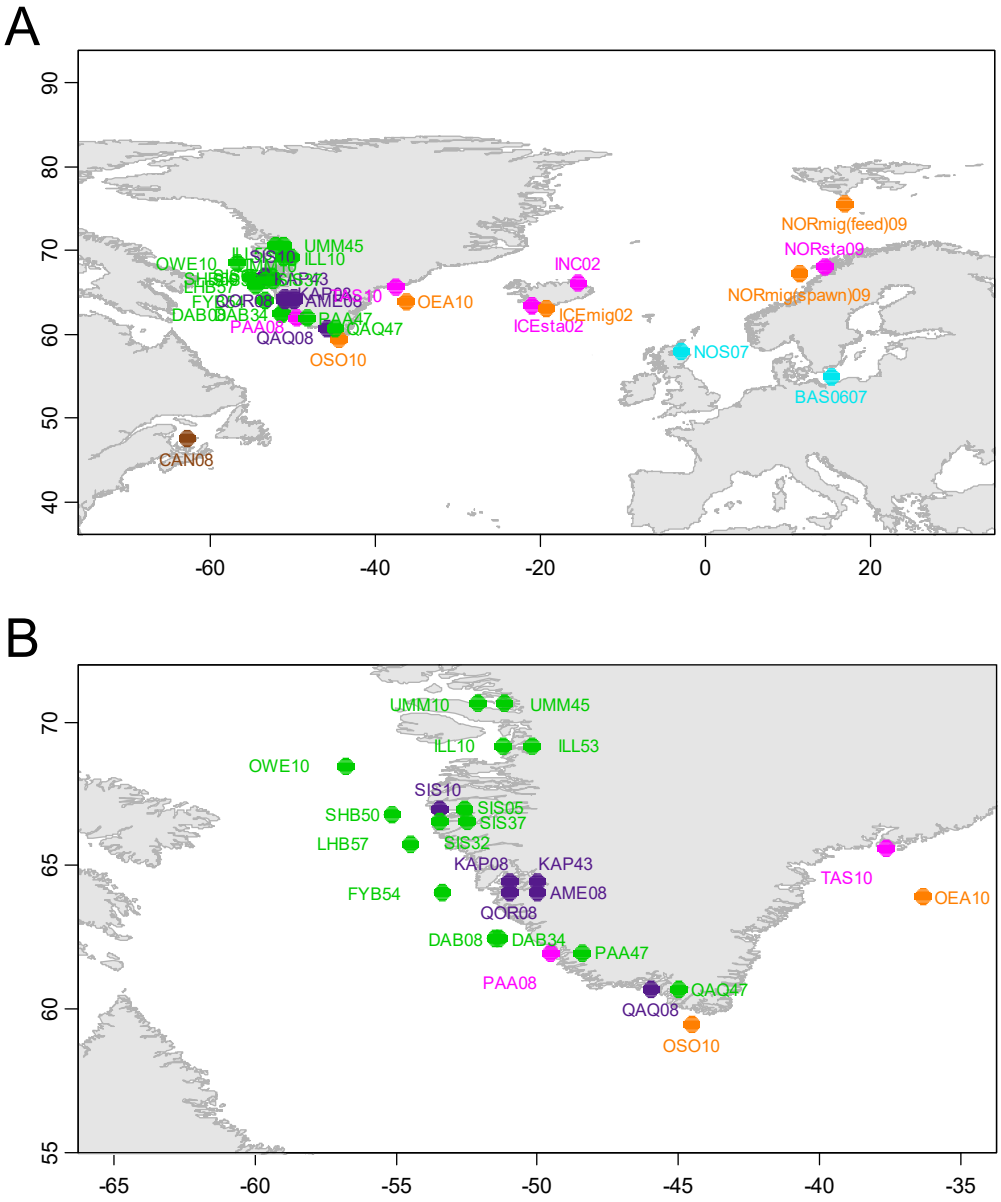


Figure S5 Sampling locations of 37 wild Atlantic cod populations. (A) Entire area. (B) Enlarged view of the area around Greenland. Data from Therkildsen *et al.* (2013) and Hemmer-Hansen (2013).

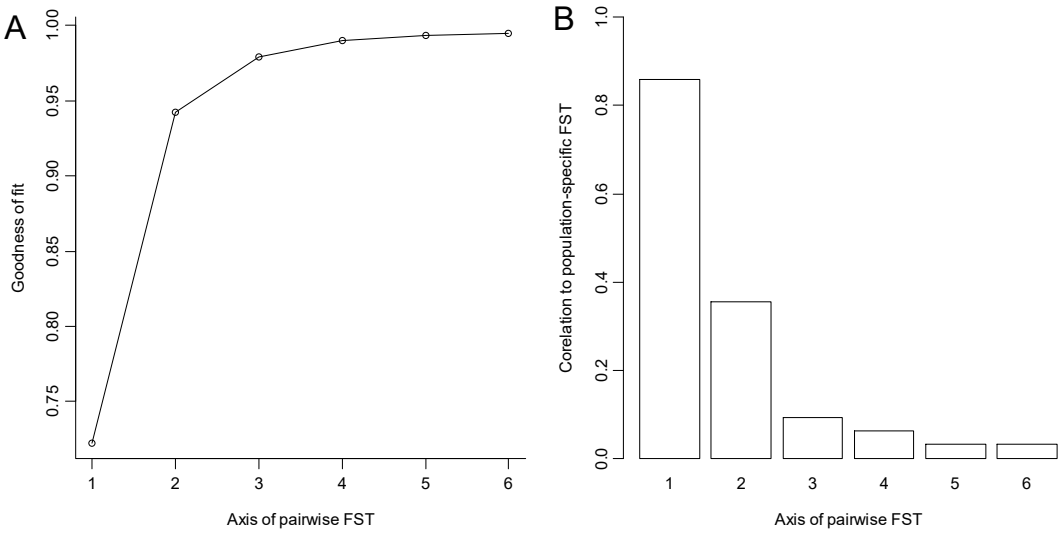


Figure S6 MDS axes of pairwise F_{ST} for 37 wild Atlantic cod populations. (**A**) Goodness of fit. (**B**) Correlation between the MDS axis of pairwise F_{ST} and population-specific F_{ST} (Weir and Goudet 2017). Data from Therkildsen *et al.* (2013) and Hemmer-Hansen (2013).

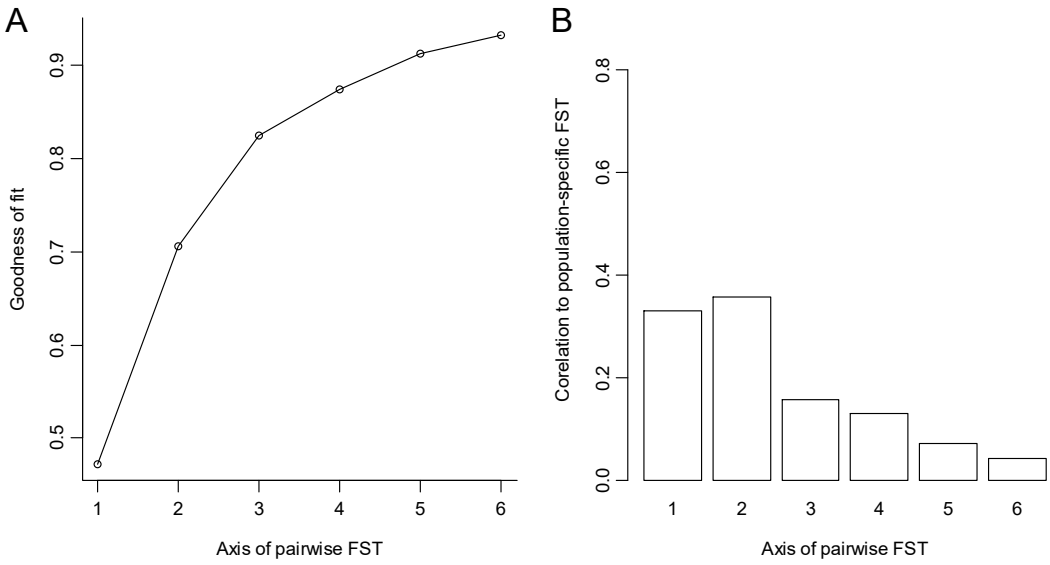


Figure S7 MDS axes of pairwise F_{ST} for 25 wild poplar populations. (**A**) Goodness of fit. (**B**) Correlation between the MDS axis of pairwise F_{ST} and population-specific F_{ST}. Data from McKown et al. (2014b).