Understanding population structure in an evolutionary context: population-specific F_{ST} and pairwise F_{ST}

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, demographic history, genetic diversity, migration, population structure

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

Populations are shaped by their history. Therefore, it is crucial to interpret population structure in an evolutionary context. Pairwise F_{ST} measures population structure, whereas population-specific F_{ST} measures deviation from the ancestral population. To understand current population structure and a population's history of range expansion, we propose a representation method that overlays population-specific F_{ST} estimates on an unrooted neighbor-joining tree inferred from a pairwise F_{ST} distance matrix and on a map of sampling locations. We examined the usefulness of our procedure by conducting simulations that mimicked population colonization from an ancestral population and analyzing published human, Atlantic cod, and wild poplar genotype data sets. Our results demonstrated that population-specific F_{ST} values identify the source population and trace the evolutionary history of its derived populations based on genetic diversity. In contrast, pairwise F_{ST} values represent the current population structure. By integrating results of both estimators, we obtained a new picture of current population structure that incorporates evolutionary history. The generalized least squares of genome-wide population-specific F_{ST} indicated that the wild poplar population expanded its distribution to the north where it adapted to longer day lengths, to seashores where it adapted to abundant rainfall, and to the south where it adapted to dry summers. Genomic data highlight the power of the bias-corrected moment estimators of F_{ST}. All F_{ST} moment estimators described in this paper have reasonable CPU times and are useful in population genomics studies. The R codes for our representation method and simulations are available in the Supporting Information.

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 our understanding of biology and biological applications, which vary 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 and there are many informative reviews on traditional and population-specific F_{ST} estimators (*e.g.*, Excoffier 2001; Rousset 2001, 2004; Balloux and Lugon-Moulin 2002; Weir and Hill 2002; Beaumont 2005; Holsinger and Weir 2009; Gaggiotti and Foll 2010; Bhatia et al. 2013; Weir and Goudet 2017). Because inferences include methods of moment, maximum likelihood and Bayesian estimation, those reviews tended to focus on theoretical perspectives. Therefore, these issues are well understood, particularly among statistical and theoretical population geneticists.

The traditional F_{ST} estimators were originally developed to estimate F_{ST} over a metapopulation based on a set of population samples. In practice, this is often called global F_{ST} (*e.g.*, Pérez-Lezaun et al. 1997). A moment estimator of F_{ST} (Weir & Cockerham in 1984) and Analysis of molecular variance (Excoffier et al. 1992) have been widely used in various fields. Additionally, F_{ST} values between pairs of population samples (pairwise F_{ST}) are routinely used to estimate population structure in human genetics (Pérez-Lezaun et al. 1997; Ramachandran et al., 2005), conservation biology (Palsbøll et al. 2007; Schwartz et al. 2007), and evolutionary biology (*e.g.*, Hemmer-Hansen et al. 2013a; Therkildsen et al. 2013a; Geraldes et al. 2014; McKown et al. 2014a; Jorde et al. 2015; Rougemont et al. 2020). Divergent selection in an environmental gradient can impact population structure (Nosil et al.

2009; Orsini et al. 2013), and prior studies examined geographic distance and habitat differences between populations as explanatory variables that impact population structure estimated based on genome-wide (average over loci) pairwise F_{ST} values (*e.g.*, Rousset 1997; Bradbury and Bentzen 2007; Petrou et al. 2014; Jorde et al. 2015; Kitada et al. 2017).

An alternative approach for estimating population differentiation is the use of population-specific F_{ST} estimators (reviewed by Gaggiotti and Foll 2010). Model-based Bayesian approaches, based on beta and/or Dirichlet distributions, for estimating population-specific F_{ST} have been proposed (Balding and Nichols 1995; Nicholson *et al.* 2002; Falush *et al.* 2003; Beaumont and Balding 2004). Locus-population-specific F_{ST} was developed to identify adaptive divergence of a gene among populations using empirical Bayes (Beaumont and Balding 2004) and full Bayesian methods (BayeScan) (Foll and Gaggiotti 2006, 2008). These methods have been applied in many cases of various species (*e.g.*, Therkildsen et al. 2013a; Geraldes et al. 2014; Limborg et al. 2012). In addition to the model-based methods, bias-corrected moment estimators of population-specific F_{ST} have been proposed (Weir and Hill 2002; Weir and Goudet 2017), which is expected to accurately reflect population history and have a wide range of applications (Weir and Goudet 2017).

Despite the expected usefulness of population-specific F_{ST}, particularly in evolutionary biology, applications have been sparce to date (*e.g.*, Nicholson *et al.* 2002; Weir *et al.* 2005; Foll and Gaggiotti 2006; Buckleton *et al.* 2013; Rougemont et al. 2020). The main goal of our paper was to demonstrate to empirical population geneticists how the two types of F_{ST} (population-specific and pairwise) can be used in combination to help elucidate population structure in the evolutionary context. In our approach, the current population structure is

represented by an unrooted neighbor-joining (NJ) tree (Saitou and Nei 1987) based on pairwise F_{ST}. The colors on the terminal nodes based on the genome-wide population-specific F_{ST} values enable inference of the historical order of population colonization. Moreover, we present a representation on the geographical map. We first examined the usefulness of our procedure by simulations that mimicked population colonization from a single ancestral population with three scenarios of population range expansion. The three scenarios, one-, two-, and three-directional population range expansion, reflect the empirical datasets of the human, Atlantic cod (*Gadus morhua*), and wild poplar (*Populus trichocarpa*), respectively, and well supported the results. Because the populations close to the ancestral population had more opportunities for mutations than recently founded populations, they are likely to have the highest heterozygosity and low values of population-specific F_{ST}. Consistently, our simulations demonstrated that Weir and Goudet's (2017) (WG) population-specific F_{ST} moment estimator correctly identified the source population and traced the evolutionary history of its derived populations based on genetic diversity. The population structure predicted by the pairwise F_{ST} moment estimator (Nei and Chesser 1983, NC83) was consistent with the current population structure.

All F_{ST} estimators were computed using the R package FinePop2_ver.0.2, which is available at CRAN. The R codes for our representation method exemplified by the human data and simulations of population colonization used in this study are available in the Supporting Information. They can be used for microsatellite and single-nucleotide polymorphism (SNP) genotype data with Genepop format (Raymond and Rousset 1995; Rousset 2008), which has been particularly widely used among biologists.

MATERIALS AND METHODS

Visual representation of population structure and demographic history

We integrated genome-wide population-specific and pairwise F_{ST} estimates on a NJ tree and on a map of sampling locations. We drew the NJ tree based on the distance matrix of genome-wide pairwise F_{ST} value using the `nj` function in the R package ape and superimposed the magnitude of genome-wide population-specific F_{ST} values using a color gradient on the NJ tree based on rgb $(1 - F_{\mathrm{ST},0}, 0, F_{\mathrm{ST},0})$, where $F_{\mathrm{ST},0} = (F_{\mathrm{ST}} - \min F_{\mathrm{ST}})/(\max F_{\mathrm{ST}} - \min F_{\mathrm{ST}})$. This conversion represents the standardized magnitude of a population-specific F_{ST} value at the sampling point, with colors between blue (for the largest F_{ST}) and red (smallest F_{ST}). The F_{ST} maps were drawn using the `sf` package in R, where sampling locations were plotted based on the longitudes and latitudes; they were visualized by the population-specific F_{ST} color gradient, and the size of each sampling point was proportional to the expected heterozygosity (H_e). Genome-wide population-specific F_{ST} values were sorted by decreasing values and plotted with standard errors. Sampling points with genome-wide pairwise F_{ST} values smaller than a given threshold were connected by yellow lines to visualize the connectivity between populations. Under the assumption of Wright's island model at equilibrium between drift, mutation, and migration (Wright 1931), $F_{\mathrm{ST}} \approx \frac{1}{4N_e m+1}$, where N_e is the effective population size and m is the average rate of migration between populations (see, Slatkin 1987). For a threshold of $F_{\mathrm{ST}} = 0.02$, $4N_e m \approx 49$ migrants per generation (also see, Whitlock and Mccauley 1999; Waples and Gaggiotti 2006).

Population colonization simulations

To test the performance of our visual representation, we conducted simulations that mimicked

colonization of populations from a single ancestral population (population 1). We modeled three types of colonization, one-, two-, and three-directional population expansion (Figure 1A–C), with 24 demes (populations 2–25). We set the effective population size of the ancestral population to $N_e = 10^4$ (twice the number of individuals in diploid organisms in a random mating population). At the beginning of colonization, 1% of N_e migrated into the adjacent vacant habitat once every 10 generations. We considered one simulation cycle to be one generation for convenience. The effective population size of the newly derived population increased to the same size of the ancestral population ($N_e = 10^4$) after one generation, and the population exchanged 1% of N_e genes with adjacent population(s) in every generation. Like the ancestral population, 1% of N_e individuals migrated into the adjacent vacant habitat once every 10 generations. We simulated the allele frequencies of SNPs in the ancestral and 24 derived populations.

The initial allele frequencies in the ancestral population, q, at 100,000 neutral SNP loci were generated from the predictive equilibrium distribution, $f(q) \propto q^{-1}(1-q)^{-1}$ (Wright 1931). Additionally, 10 newly derived SNPs were introduced to each existing population in each generation. Therefore, in total, 35,100 SNPs were generated. When a new SNP emerged in a population, we set the initial allele frequency of the newly derived SNP to 0.01 in the population and 0 in the other populations. This mimicked new mutations that survived at the initial phase after their birth. These 100,000 ancestral SNPs and 35,100 newly derived SNPs were considered “unobserved.” The allele frequencies of these SNPs were changed by random drift under a binomial distribution in every generation. Many of the SNPs had reduced frequencies of the derived allele over simulation generations and lost their polymorphism. After 260 simulation generations, SNPs that retained their polymorphism were randomly

selected as "observed" SNPs; in this simulation, we selected 10,000 ancestral SNPs and 500 newly derived SNPs. Then, we generated 50 individuals for each population. Genotypes of these 10,500 SNPs were randomly generated for each individual following the allele frequencies in the population to which each individual belonged. Thus, we obtained "observed" genotypes of 1,250 individuals (= 50 individuals × 25 populations) at 10,500 SNP loci (10,000 ancestral SNPs + 500 newly derived SNPs by mutation). We converted the simulated genotypes into Genepop format (Raymond & Rousset, 1995; Rousset, 2008). We computed genome-wide population-specific and pairwise F_{ST} values between 25 populations and overlaid genome-wide population-specific F_{ST} values on the unrooted NJ tree.

Inferring environmental selection from observed population structure

Hereafter, notations consistent with those of Weir and Hill (2002) were used: i for populations ($i = 1, \ldots, r$), u for alleles ($u = 1, \ldots, m$), and l for loci ($l = 1, \ldots, L$). To infer the geography and environment that were experienced by the population range expansion, we regressed the genome-wide population-specific F_{ST} values on the geographical and environmental variables. Residuals are correlated because of population structure; therefore, the effective sample size is lower than the actual sample size. In such circumstances, ordinary least squares overestimate the precision. To take the correlation into account, we derived the components of the variance–covariance matrix of the population-specific F_{ST} estimator. It can be written as $\mathrm{ps}\hat{F}^{i}_{\mathrm{ST}} = \frac{\bar{y}^{i}}{\bar{x}}$. Using the Taylor series expansion for the first term, we inferred the asymptotic variance as

$$V\left[\mathrm{ps}\hat{F}^{i}_{\mathrm{ST}}\right] \simeq \frac{\bar{y}^{2}}{\bar{x}^{2}}\left\{\frac{V[\bar{x}]}{\bar{x}^{2}} + \frac{V\left[\bar{y}^{i}\right]}{\bar{y}^{i\,2}} - \frac{2Cov\left[\bar{x}, \bar{y}^{i}\right]}{\bar{x}\bar{y}^{i}}\right\} \qquad (1)$$

Similarly, the asymptotic covariances between population-specific F_{ST} values of i, j

populations were obtained by

$$Cov\left[\mathrm{ps}\hat{F}_{\mathrm{ST}}^{i}, \mathrm{ps}\hat{F}_{\mathrm{ST}}^{j}\right] \simeq \frac{\bar{y}^{i}\bar{y}^{j}}{\bar{x}^{4}} V[\bar{x}] - \frac{\bar{y}^{i}}{\bar{x}^{3}} Cov\left[\bar{x}, \bar{y}^{j}\right] - \frac{\bar{y}^{j}}{\bar{x}^{3}} Cov\left[\bar{x}, \bar{y}^{i}\right], \qquad (2)$$

where the variance and covariance components were calculated by

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

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

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

This analysis was performed on the wild poplar data set, for which 11 environmental/geographical parameters were available for each sampling location. We used Equations 1 and 2 for the components of the variance–covariance matrix Ω of the `GLS` function in FinePop2_ver.0.2. Once the key factors that are associated with the population range expansion are identified, it would be interesting to search for the crucial genes that enabled adaptation to the local environment by determining outliers of locus-specific, population-specific F_{ST} values (Foll and Gaggiotti 2008; Coop et al. 2010).

F_{ST} estimators applied in this study

We used WG population-specific F_{ST} moment estimators, because we expected that population-specific F_{ST} values accurately reflect population history (Weir and Goudet 2017). In our analyses, we extended WG population-specific F_{ST} estimator to overall loci (genome-wide population-specific F_{ST}), and the combined ratio estimator (Cochran 1977) for overall loci (Buckleton et al. 2016) was:

$$\mathrm{ps}\hat{F}_{\mathrm{ST}}^{i} = \frac{\sum_{l=1}^{L} (\tilde{M}_{W,l}^{i} - \tilde{M}_{l}^{B})}{\sum_{l=1}^{L} \left(1 - \tilde{M}_{l}^{B}\right)}, \qquad (3)$$

where $\widetilde{M}_W^i$ 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' (Supplemental Note). This is called the "ratio of averages" F_{ST} estimator, which asymptotically converges to unbiased estimates of F_{ST} as the number of independent SNPs increases (Weir and Cockerham 1984; Weir and Hill 2002; Bhatia et al. 2013).

We also applied 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}:

$$L_{li}\left(\theta_{li} \middle| n_{li1}, \ldots, n_{lim_l}\right) = \frac{\Gamma(\theta_{li})}{\Gamma(N_{li} + \theta_{li})} \prod_{u=1}^{m_l} \frac{\Gamma(n_{liu} + \theta_{li}\bar{p}_{lu})}{\Gamma(\theta_{li}\bar{p}_{lu})} . \qquad (4)$$

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 populations, whereas $\theta_{li}\bar{p}_{lu} = \alpha_{liu}$, where $\theta_{li} = \sum_{u=1}^{m_l} \alpha_{liu}$. The parametrization reduces the number of parameters to be estimated. Based on a Dirichlet (multi-allelic) and/or a beta (bi-allelic) scale parameter, population-specific F_{ST} values were estimated for each locus using the following function of $\hat{\theta}_{li}$ (Beaumont and Balding, 2004):

$$\mathrm{ps}\tilde{F}_{\mathrm{ST},l}^{i} = \frac{1}{\hat{\theta}_{li} + 1} . \qquad (5)$$

To estimate genome-wide pairwise F_{ST}, we used the NC83 bias-corrected G_{ST} moment estimator:

$$\mathrm{pw}\hat{F}_{\mathrm{ST}} = \frac{\sum_{l=1}^{L}\left(\hat{H}_{T,l} - \hat{H}_{S,l}\right)}{\sum_{l=1}^{L}\hat{H}_{T,l}}, \qquad (6)$$

where $\hat{H}_T$ and $\hat{H}_S$ are the unbiased estimators of total and within-population heterozygosity, respectively (Supplemental Note). G_{ST} (Nei 1973) is defined "by using the gene frequencies at the present population, so that no assumption is required about the pedigrees of individuals, selection, and migration in the past" (Nei 1977). G_{ST} assumes no evolutionary history (Holsinger and Weir 2009), whereas NC83 does not consider any population replicates (Weir and Cockerham 1984; Excoffier 2001). The pairwise F_{ST} values obtained from NC83 therefore measured genetic differentiation among current populations based on a fixed set of samples of populations, which were derived from an ancestral population. Our previous coalescent simulations demonstrated that NC83 performs the best among F_{ST} estimators when estimating pairwise F_{ST} values, particularly for larger numbers of loci (Kitada et al. 2017).

Computing F_{ST} values

We converted the genotype data into Genepop format (Raymond and Rousset 1995; Rousset 2008) for implementation in the R package FinePop2_ver.0.2. Genome-wide WG population-specific F_{ST} (Equation 3) values were computed using the `pop_specificFST` function. In addition to the "ratio of averages" (Weir and Cockerham 1984; Weir and Hill 2002) used for the F_{ST} functions in FinePop2, we computed the "average of ratios" (Bhatia et al. 2013) of the WG population-specific F_{ST} for human data for comparison by averaging locus-specific, population-specific F_{ST} values over loci. We maximized Equation 4 and estimated the empirical Bayesian population-specific F_{ST} (Beaumont and Balding 2004) at each locus according to Equation 5. We then averaged these values over all loci. For the full Bayesian model, GESTE_ver. 2.0 (Foll and Gaggiotti 2006) was used to compute genome-wide population-specific F_{ST} values. The Bayesian genome-wide population-specific

F_{ST} values were computed for human data to compare the results with that from the WG population-specific F_{ST} moment estimator. F_{ST} is equal to G_{ST} for diploid random mating populations (Excoffier 2001; also see Supplemental Note); therefore, NC83 pairwise F_{ST} values (Equation 6) were computed using the `pop_pairwiseFST` function in FinePop2. Expected heterozygosity was calculated for each population with the `read.GENEPOP` function.

Three empirical data sets

As the first empirical data set, we chose microsatellite genotypes collected from human populations worldwide, because human evolutionary history, migration, and population structure have been particularly well studied (*e.g.*, Diamond 1997; Rosenberg et al. 2002; Ramachandran et al. 2005; Liu et al. 2006; Rutherford 2016; Nielsen et al. 2017). These patterns are well known by statistical/theoretical population geneticists and biologists; therefore, testing our new integrative F_{ST} analysis on this data set could provide a good example of the usefulness of our practical approach. The human microsatellite data in Rosenberg et al. (2002) were retrieved from https://web.stanford.edu/group/rosenberglab/index.html. We converted the genotype data to Genepop format (Raymond and Rousset 1995; Rousset 2008). We removed the Surui sample (Brazil) from the data because that population was reduced to 34 individuals in 1961 as a result of introduced diseases (Liu et al. 2006). We retained genotype data (n = 1,035) of 377 microsatellite loci from 51 populations categorized into six groups as in the original study: 6 populations from Africa, 12 from the Middle East and Europe, 9 from Central/South Asia, 18 from East Asia, 2 from Oceania, and 4 from America. Longitudes and latitudes of the sampling sites were obtained from Cann et al. (2002) (Table S1).

An SNP data set was obtained from a commercially important fish, Atlantic cod, in the North Atlantic. We combined the Atlantic cod genotype data of 924 SNPs common to 29 populations reported in Therkildsen et al. (2013a, b) and 12 populations in Hemmer-Hansen et al. (2013a, b) in Genepop format. We compared genotypes associated with each marker in samples that were identical between the two studies, namely, CAN08 and Western_Atlantic_2008, ISO02 and Iceland_migratory_2002, and ISC02 and Iceland_stationary_2002, and standardized the gene codes. We removed cgpGmo.S1035, whose genotypes were inconsistent between the two studies. We also removed cgpGmo.S1408 and cgpGmo.S893, for which genotypes were missing in several population samples in Therkildsen et al. (2013b). Temporal replicates in Norway migratory, Norway stationary, North Sea, and Baltic Sea samples were removed for simplicity. The final data set consisted of genotype data (n = 1,065) at 921 SNPs from 34 populations: 3 from Iceland, 25 from Greenland, 3 from Norway, and 1 each from Canada, the North Sea, and the Baltic Sea. The genotype data included historical samples collected 50–80 years ago and 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 detection of migration history of this highly migratory marine fish. Two ecotypes (migratory and stationary) that were able to interbreed but were genetically differentiated (Hemmer-Hansen et al., 2013a; Berg et al., 2016) were included in the Norway and Iceland samples. All individuals in the samples were adults, and most were mature (Therkildsen et al., 2013a). The longitudes and latitudes of the sampling sites in Hemmer-Hansen et al. (2013a) were used. For the data from Therkildsen et al. (2013a), approximate sampling points were estimated from the map of the original study, and longitudes and latitudes were recorded (Table S2).

The other SNP data set was from a tree, wild poplar, in the American Pacific Northwest. The SNP genotype data and environmental/geographical data were retrieved from the original studies of McKown et al. (2014a, b). The genotype data contained 29,355 SNPs of 3,518 genes of wild poplar (n = 441) collected from 25 drainage areas (McKown et al., 2014c). Details of array development and selection of SNPs are provided in Geraldes et al. (2011, 2013). We converted the data to Genepop format. The samples covered various regions over a range of 2,500 km near the Canadian–US border at altitudes between 0 and 800 m. A breakdown of the 25 drainages (hereafter, populations) is as follows: 9 in northern British Colombia (NBC), 2 in inland British Colombia (IBC), 12 in southern British Colombia (SBC), and 2 in Oregon (ORE) (Geraldes et al., 2014). The original names of clusters and population numbers were combined and used for our population labels (NBC1, NBC3,…, ORE30). Each sampling location was associated with 11 environmental and geographical parameters: latitude (lat), longitude (lon), altitude (alt), longest day length (DAY), frost-free days (FFD), mean annual temperature (MAT), mean warmest month temperature (MWMT), mean annual precipitation (MAP), mean summer precipitation (MSP), annual heat-moisture index (AHM), and summer heat-moisture index (SHM) (Table S3). The AHM was calculated in the original study as (MAT+10)/(MAP/1000); a large AHM indicates extremely dry conditions.

RESULTS

Simulations of population colonization

In the one-directional simulation, our method correctly identified the ancestral population with the highest genetic diversity, and populations were located in order from 1 to 25 on the NJ tree (Figure 1D). In the two-directional simulation, our method correctly identified the

ancestral population and detected that populations were split at population 9 and expanded in two directions, which was consistent with the simulation scenario (Figure 1E). In the three-directional simulation, the ancestral population was closely located to the adjacent populations 2, 9, and 17, but correctly detected three directions, as in the other simulations (Figure 1F).

Our simulation results demonstrated that our method represents insight into current population structure that incorporates evolutionary history. Our results revealed that WG population-specific F_{ST} values (standardized by a color gradient) identified the source population and traced the evolutionary history of its derived populations based on genetic diversity. The NC83 pairwise F_{ST} estimator correctly estimated the current population structure. Genome-wide WG population-specific F_{ST} values were negative in the ancestral population and adjacent populations, and the phenomenon was particularly significant in the one- and two-directional models, whereas no negative population-specific F_{ST} values were obtained in the ancestral and adjacent populations in the three-directional model (Figures 1G-I). In contrast, H_e values were larger in the ancestral population in the one- and two-directional models than in the three-directional model, though the variation in H_e values was not substantial because of the relatively few generations (260) in the simulation (Figures 1D-F). Our simulation results indicated that, when gene flow from other populations into the source population was limited, relatively large H_e could be maintained, which resulted in substantial negative population-specific F_{ST} values. Very similar results were obtained from different data with more than 20 simulations (Figures not shown).

Humans

The map (Figure 2A) shows integrated information from genome-wide population-specific and pairwise F_{ST} and H_e, which revealed population structure with migration and genetic diversity of human populations. Interestingly, Bantu Kenyans had the smallest F_{ST} value (as shown in red). Figure 2B ordered population-specific F_{ST} values from Africa to Central/South Asia, the Middle East, Europe, East Asia, Oceania, and America (Table S4). As indicated by sampling points connected by yellow lines with pairwise F_{ST} values below the 0.02 threshold (Figure 2A), genetic connectivity from Africa was low. In contrast, migration was substantial within Eurasia but was much smaller than that inferred from Eurasia to Oceania and America. As illustrated by sampling point radii (Figure 2A), H_e was the highest in Africa, followed by the Middle East, Central/South Asia, Europe, and East Asia, but relatively small in Oceania and lowest in South America. The Karitiana in Brazil had the lowest H_e. The NJ tree (Figure 2C) integrated with population-specific F_{ST} values revealed that human populations originated from Bantu Kenyans and expanded to Europe through Mozabite, the Middle East, Central/South Asia, and East Asia. The Kalash was isolated from Europe/Middle East and Central/South Asia populations. Papuans/Melanesians and American populations diverged from between Central/South Asian and East Asian populations. The ordinal NJ tree of pairwise F_{ST} values divided the populations into five clusters: 1) Africa, 2) the Middle East, Europe, and Central/South Asia, 3) East Asia, 4) Oceania, and 5) America (Figure S1).

Bayesian population-specific F_{ST} values estimated using the methods of Beaumont and Balding (2004) (empirical Bayes) and Foll and Gaggiotti (2006) (full Bayes) were nearly identical; however, in African populations, they were higher than WG population-specific F_{ST} values (Figure S2). The distributions of F_{ST} values obtained from the two Bayesian methods were very similar, with the smallest F_{ST} values observed in the Middle East, Europe, and

Central/South Asia (Figure S2, Table S4). The “ratio of averages” and “average of ratios” of the WG population-specific F_{ST} estimator were almost identical in all populations for this data set (Figure S3).

Atlantic cod

The F_{ST} map (Figure 3A) visualized population structure, migration, and genetic diversity of the Atlantic cod populations. The Canadian population had the smallest population-specific F_{ST} value (as shown in red) and the greatest H_e (indicated by circle radii). As shown by sampling point radii (Figure 3A), H_e was greatest in Canada. It was also high in Greenland, low in other areas, and lowest in the Baltic Sea. Figure 3B ordered population-specific F_{ST} values from Canada to the Baltic sea (Table S5). Greenland west coast populations (green in Figure S4) generally had small population-specific F_{ST} values, whereas fjord populations had relatively higher population-specific F_{ST} values. The F_{ST} values were much higher for populations in Iceland, Norway, and the North Sea, and highest in the Baltic Sea. Based on pairwise F_{ST} values between sampling points (< 0.02 threshold) (Figure 3A), substantial migration was detected between Greenland, Iceland, and Norway. In contrast, migration was low from Canada to Greenland and from Iceland and Norway to the North and Baltic Seas. Our integrated NJ tree (Figure 3C) with population-specific F_{ST} values indicated that Atlantic cod originated from Canada, migrated to the west coast of Greenland, and then expanded their distribution to Iceland, Norway, the North Sea, and the Baltic Sea.

Atlantic cod might migrate to find new habitat when they reach carrying capacity, and individuals with genomic variation were able to adapt to changing environments and formed the current local populations. According to the ordinal NJ tree of the pairwise F_{ST} distance

matrix (Figure S4), the populations were divided into four large clusters: 1) Canada, 2) Greenland west coast, 3) Greenland east coast, Iceland, Norway, and 4) North and Baltic Seas. Fjord populations (in purple) formed a sub-cluster within the Greenland west coast, and migratory (orange) and stationary (magenta) ecotypes also formed a sub-cluster.

Wild poplar

The F_{ST} map (Figure 4A) indicated that population-specific F_{ST} values were lowest in SBC27, IBC15, and IBC16 (as shown in red). Sampling points connected by yellow lines (< 0.02 threshold pairwise F_{ST} values) indicated migration between all populations. As shown by sampling point radii (Figure 4A), H_e was highest in SBC27, IBC15, and IBC16, and lowest in NBC5. Figure 4B showed that samples collected from areas close to the SBC coast had higher population-specific F_{ST} values than other SBC samples (Figure S5, Table S6). NBC samples had population-specific F_{ST} values similar to those of SBC. Among NBC samples, NBC8 had the smallest population-specific F_{ST}, and NBC5 had the highest value, followed by NBC6 and NBC7. The pairwise F_{ST} NJ tree integrated with population-specific F_{ST} values (Figure 4C) indicated that wild poplar originated from Inner BC and SBC27, and expanded in three directions, namely, to the BC southern coast, northern BC and south-western Alaska, and Oregon. The ordinal NJ tree based on the pairwise F_{ST} distant matrix divided populations into three large clusters: 1) IBC, 2) SBC, 3) NBC, and 4) ORE (Figure S7). The population represented by sample ORE30 was isolated from ORE29.

To avoid multicollinearity, we excluded seven out of 11 environmental variables that were significantly correlated with each other, namely, lat, lon, alt, FFD, MWMT, MSP, and AHM. Our GLS of genome-wide population-specific F_{ST} values on the four environmental variables

(DAY, MAT, MAP, and SHM) indicated that DAY, MAP, and SHM were significant (Table 1). All estimates were positive, which indicated that higher population-specific F_{ST} values were expected for longer DAY (longer daylight time), higher MAP (abundant rain), and higher SHM (dry summers), and these values reflected directions of population expansion. The scatter plot of DAY and SHM (each population colored by the population-specific F_{ST} value) (Figure 5A) suggested three directions of population range expansion; the wild popular that originated from IBC15 expanded its distribution to NBC, where it adapted to longer DAY; that which originated from SBC27 expanded to SBC seashores, where it adapted to lower SHM, and to ORE29 and ORE30, where it adapted to higher SHM. This was consistent in the scatter plot of DAY and MAP (Figure 5B), which suggested that the expansion in SBC could have been facilitated by adaptation to higher MAP.

CPU times

Using a laptop computer with an Intel Core i7-8650U CPU, only 89.8 s of CPU time was required to compute WG population-specific F_{ST} estimates and SEs of wild poplar (29,355 SNPs; 25 populations, n = 441). Alternatively, 120.7 s was required to obtain pairwise F_{ST} (NC83) between all population pairs. Based on the results, we may need 50 min to compute WG population-specific F_{ST} and 70 min to compute pairwise NC83 F_{ST} estimates for 1 million SNPs using this laptop. This computation could be much faster if we used a workstation.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, a table, and supplemental information.

DISCUSSION

Genome-wide population-specific F_{ST} traced population history as reflected by genetic diversity

Our simulations demonstrated that the WG population-specific F_{ST} estimator was able to identify the source population and trace the evolutionary history of its derived populations based on genetic diversity. We also found that the NC83 pairwise F_{ST} estimator correctly estimated current population structure. Our one-, two-, and three-directional simulations corresponded to Atlantic cod, human, and wild poplar data sets, respectively, which further supported our results. As discussed below, the WG population-specific F_{ST} estimator is a rescaling of expected heterozygosity, and a linear relationship between expected heterozygosity and population F_{ST} is expected. This shows that the population-specific F_{ST} estimator implicitly assumes that populations closest to the ancestral population have the highest heterozygosity. Our simulations evaluated the performance of the WG population-specific F_{ST} estimator for such cases.

In our analysis, genome-wide WG population-specific F_{ST} values successfully illustrated human evolutionary history, and indicated that humans originated in Kenya, expanded from the Middle East into Europe and from Central/South Asia into East Asia, and then possibly migrated to Oceania and America (Figure 2). Kenya is located just below Ethiopia, where the earliest anatomically modern humans were found from fossils (Nielsen et al. 2017). Our results are also in good agreement with the highest levels of genetic diversity being detected in Africa (Rosenberg et al., 2002), the relationship uncovered between genetic and geographic distance (Ramachandran et al., 2005), the shortest colonization route from East Africa (Liu et

al., 2006), and major migrations inferred from genomic data (Nielsen et al., 2017). The genome-wide WG population-specific F_{ST} values are consistent with results obtained from 24 forensic STR markers (Buckleton et al., 2016).

Our analysis indicated that Atlantic cod originated in Canada (CAN08). Figure 3 suggested that the population expansion of Atlantic cod began by minimal gene flow from Canada. They might have first expanded to the west coast of Greenland before spreading to Iceland, the North Sea, Norway, and the Baltic Sea. This result was consistent with genomic evidence that Atlantic cod inhabit both sides of the Atlantic Ocean and evolved from a common evolutionary origin (Berg et al. 2017). The migratory ecotypes characterized by deeper and more offshore habitats and long-distance migrations (Hemmer-Hansen et al., 2013a) may have played an important role in this expansion, as suggested in Figure 3C. In the original Atlantic cod study (Therkildsen et al. 2013a), strong differentiation of CAN08 was found at neutral markers, which prompted the authors to suggest that Greenland populations were the result of colonization from Iceland rather than from refugial populations in southern North America. In our study, CAN08 had the highest H_e, which was lower in Iceland than in Greenland (Figure 3A); this result implies that Icelandic populations were the descendants of colonists from Greenland, which in turn originated in Canada. The BAS0607 sample from the Baltic Sea had the highest population-specific F_{ST} and the lowest H_e, which suggests that Baltic cod is the newest population. This result agrees with the findings of a previous study, which identified Baltic cod as an example of a species subject to ongoing selection for reproductive success in a low salinity environment (Berg et al. 2015).

Although the wild poplar samples used in this study might not cover the whole distribution

range of the species, which extends from southern California to northern Alaska, Montana, and Idaho (Geraldes et al. 2013), genome-wide population-specific F_{ST} values suggested that wild poplar trees in southern British Colombia (SBC27) and inland British Colombia (IBC15, IBC16) are the closest to the ancestral population. Our results showed that the wild popular expanded in three directions: coastal British Colombia, southern Oregon, and northern British Colombia (Figure 5). The largest population-specific F_{ST} value was found in the population with the smallest heterozygosity, SBC22, which may have resulted from a bottleneck (Geraldes et al., 2014).

Genome-wide population-specific F_{ST} detects key environments that promote adaptation

Our GLS of genome-wide population-specific F_{ST} values revealed that long daylight hours, abundant rainfall, and dry summer conditions are the key environmental factors that influenced the evolution of wild poplar (Table 1). Wild poplar could have expanded their distribution by their fluffy seeds being blown away by wind, and individuals that had genomic variation were able to adapt to local environments. The results suggested that wild popular originated from southern British Colombia (SBC27) and inland British Colombia (IBC15, 16) and expanded its distribution to NBC by adapting to longer day lengths, to SBC seashores adapting to the rainy environment, and to southern Oregon (ORE30) adapting to dry summer conditions (Figure 5). A previous study on wild poplar revealed that genes involved in drought response were identified as F_{ST} outliers along with other genes related to transcriptional regulation and nutrient uptake (Geraldes et al. 2014), which is a finding consistent with our GLS results. Our results were also consistent with the F_{ST} outlier test of the original study (Geraldes et al., 2014), in which Bayescan (Foll and Gaggiotti 2008)

revealed that genes involved in circadian rhythm and response to red/far-red light had high locus-specific global F_{ST} values. Moreover, the first principal component of SNP allele frequencies was significantly correlated with day length, and a previous enrichment analysis for population structuring uncovered genes related to circadian rhythm and photoperiod (McKown et al. 2014a). Our results were in agreement with the previous findings, which show the usefulness of using GLS of genome-wide population-specific F_{ST} to infer environmental adaptation and population expansion of species.

Properties of F_{ST} moment estimators

Previous studies have suggested or indicated that the "ratio of averages" works better than the "average of ratios" as the number of independent SNPs increases (Weir and Cockerham 1984; Weir and Hill 2002; Bhatia et al. 2013). In regard to the WG population-specific F_{ST} estimator, similar results were obtained for the 377 human microsatellite loci using either the "ratio of averages" or the "average of ratios" (Figure S3). This similar outcome may have been due to the relatively small variation in the locus-specific global F_{ST} values (Figure not shown) and the relatively large number of alleles (12 ± 4) of human microsatellites.

To explicitly show the underlying mechanism, we used the observed heterozygosity of population i ($\widehat{H}_{Si}$) 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 heterozygosity thus converges to the expected value:

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

Similarly, $\widehat{H}_S$ 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 show that genomic data highlight the usefulness of the bias-corrected moment estimators of traditional F_{ST} developed in the early 1980s (Nei and Chesser 1983; Weir and Cockerham 1984) and population-specific F_{ST} (Weir and Goudet 2017).

To estimate pairwise F_{ST}, our previous coalescent simulations based on ms (Hudson 2002) showed that NC83 performed best among the present F_{ST} estimators for cases with 10,000 SNPs (Kitada et al. 2017). Other F_{ST} moment estimators within an ANOVA framework produce values approximately double those of true values when used to estimate pairwise F_{ST}. NC83 considers a fixed set of population samples; in contrast, the other F_{ST} moment estimators consider replicates of a set of populations (Weir and Cockerham 1984; Holsinger and Weir 2009). The models for replicates of population samples were considered to appropriately estimate global F_{ST} and/or mean ancestral coefficient, but cause over-estimation when used to estimate pairwise F_{ST}.

The WG population-specific F_{ST} moment estimator measures population genetic diversity under the framework of relatedness of individuals and identifies 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):

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

Another variable, $\widetilde{M}^B$, is "average over pairs of populations of between-population-pair matching" (Weir and Goudet 2017). $\widetilde{M}^B$ is the homozygosity over pairs of populations, and we can write observed heterozygosity over pairs of populations as $1 - \widetilde{M}^B = \widehat{H}_B$. $\widehat{H}_B$ is an estimator for the denominator of Hudson et al. (1992). When using only allele frequencies, the population-specific F_{ST} estimator can be written in terms of gene diversity as

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

This formulation is reasonable, because WG population-specific F_{ST} uses "allele matching, equivalent to homozygosity and complementary to heterozygosity as used by Nei (1973), rather than components of variance" (Weir and Goudet 2017). In our three case studies, a linear relationship between H_e of each population ($= H_{Si}$) and $\mathrm{ps}\widehat{F}_{\mathrm{ST}}^i$ was evident (Figure S6). The coefficient of determination, R^2, was 0.91 for 51 human populations (n = 1,035), 0.99 for 34 Atlantic cod populations (n = 1,065), and 0.82 for 25 wild poplar populations (n = 441). The goodness of fit to the linear function should depend on population sample size (number of individuals).

In the Atlantic cod case study, CAN08 had the highest H_e (Figure 4b) and a very large negative population-specific F_{ST} value of −0.21 ± 0.019 compared with the maximum value of 0.22 ± 0.014 in BAS0607 (Figure S6, Table S5). The Atlantic cod data corresponded to the

one-directional model of our simulations, where the WG population-specific F_{ST} value was significantly negative in the ancestral population, whereas H_e was the largest (Figure 1D,G). Our consistent results between the simulations and Atlantic cod case study indicate that, when gene flow from other populations into the source population is limited, relatively large H_e ($\widehat{H}_{Si}$) is maintained in the source population. In such cases with $\widehat{H}_{Si} > \widehat{H}_B$, Equation 7 produces negative values of population-specific F_{ST}.

Bayesian F_{ST} estimators measure the deviation from the average of the sampled populations

Our F_{ST} map based on the empirical Bayesian population-specific F_{ST} values (Beaumont and Balding 2004) indicated that the Middle East, Europe, and Central/South Asia were centers of human origin (Figure 6A), which was consistent with that from the full Bayesian population-specific F_{ST} estimator (Foll and Gaggiotti 2006) (figure not shown). Our integrated NJ tree showed that the Hazara, Pakistan population was genetically closest to the human ancestors (Figure 6B). The results obtained with Bayesian estimators were a consequence of Equation 4, which uses the mean allele frequency over populations ($\bar{p}_{lu}$) to reduce the number of parameters to be estimated. The numbers of sampling locations of the 51 human populations were as follows: 6 from Africa, 12 from the Middle East/Europe, 9 from Central/South Asia, 18 from East Asia, 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 results indicate 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, when we used a subsample of the 37 human populations (3 from Africa, 6 from the Middle East/Europe, 4 from Central/South Asia, 18 from East Asia, 2 from Oceania, and 4 from America; Table S7), the area with the highest population-specific F_{ST} values shifted toward Central/South Asia and East Asia, whereas Bantu Kenyans had the smallest WG population-specific F_{ST} value; this was consistent with the results from the full data set (Figure 7).

In contrast, because of shrinkage toward mean allele frequencies, maximum likelihood and Bayesian estimators of locus-specific global F_{ST} improve the power to detect genes under environmental selection (Beaumont and Balding 2004). Our empirical Bayes pairwise F_{ST} estimator (EBF_{ST}; Kitada et al. 2007), which is based on Equation 4, is also useful in cases involving a relatively small number of polymorphic marker loci, such as microsatellites; it performs best by averaging large sampling variation of allele frequencies in populations with small sample sizes, particularly in high gene flow scenarios (Kitada et al., 2017). However, EBF_{ST} suffers from a shrinkage effect similar to that of Bayesian population-specific F_{ST} estimators. We note that the shrinkage effect on allele frequencies can enhance the bias of Bayesian F_{ST} estimators, particularly in genome analyses where large numbers of SNPs are used.

Conclusions

The genome-wide WG population-specific F_{ST} moment estimator was able to identify the source population and trace the evolutionary history of the derived population's history based

on genetic diversity under the assumption that populations closest to the ancestral population have the highest heterozygosity. In contrast, NC83 genome-wide pairwise F_{ST} moment estimator represents the current population structure. By integrating estimates from both F_{ST} estimators on NJ trees and maps of sampling locations, we obtained a picture of population structure by incorporating evolutionary history. Our GLS analysis of genome-wide population-specific F_{ST}, which takes the correlation between population-specific F_{ST} values into account, provides insights into how a species has adapted to key environments and expanded its distribution. 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}.

ACKNOWLEDGEMENTS

We express our appreciation to editors Mark Beaumont and Nicholas Barton for reviewing this manuscript and providing the opportunity for revision. We thank the reviewers for providing essential comments that significantly improved the manuscript. 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.

LITERATURE CITED

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

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

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

Berg, P. R., B. Star, C. Pampoulie, I. R. Bradbury, P. Bentzen *et al.* 2017 Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, *119*, 418–428. https://doi.org/10.1038/hdy.2017.54

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

Bradbury, I. R., and P. Bentzen, 2007 Non-linear genetic isolation by distance: implications for dispersal estimation in anadromous and marine fish populations. Mar. Ecol. Prog. Ser. 340: 245–257. https://doi:10.3354/meps340245

Buckleton, J., J. Curran, J. Goudet, D. Taylor, A. Thiery *et al.* 2016 Population-specific F_{ST} values for forensic STR markers: A worldwide survey. Foren. Sci. Intnatl. Genetics 23: 91–100. https://doi.org/10.1016/j.fsigen.2016.03.004

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. https://doi:10.1126/science.296.5566.261b

Cockran, W. G., 1977 *Sampling Techniques*. Wiley, New York.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411–1423. https://doi.org/10.1534/genetics.110.114819

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.

Excoffier, L., P. E. Smouse, & J. M. Quattro, 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479-491. PubMed 1644282

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Foll, M., and O. Gaggiotti, 2006 Identifying the environmental factors that determine the genetic structure of populations. Genetics 174: 875–891. https://doi.org/10.1534/genetics.106.059451

Foll, M., and O. E. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977–993. https://doi.org/10.1534/genetics.108.092221

Gaggiotti, O. E., and M. Foll, 2010 Quantifying population structure using the F-model. Molec. Ecol. Resour. 10: 821–830. https://doi.org/10.1111/j.1755-0998.2010.02873.x

Geraldes, A., J. Pang, N. Thiessen, T. Cezard, R. Moore *et al*. 2011 SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. Molec. Ecol. Resour. 11: 81–92. https://doi.org/10.1111/j.1755-0998.2010.02960.x

Geraldes, A., S. P. Difazio, G. T. Slavov, P. Ranjan, W. Muchero *et al.* 2013 A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. Molec. Ecol. Resour. 13: 306–323. https://doi.org/10.1111/1755-0998.12056

Geraldes, A., N. Farzaneh, C. J. Grassa, A. D. McKown, R. D. Guy *et al.* 2014 Landscape genomics of *Populus trichocarpa* the role of hybridization limited gene flow and natural selection in shaping patterns of population structure. Evolution 68: 3260–3280. https://doi.org/10.1111/evo.12497

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., 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

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589. PubMed 1427045

Jorde, P. E., G. Søvik, J. I. Westgaard, J. Albretsen, C. André *et al.* 2015 Genetically distinct populations of northern shrimp, *Pandalus borealis*, in the North Atlantic: adaptation to different temperatures as an isolation factor. Molec. Ecol. 24: 1742–1757. https://doi.org/10.1111/mec.13158

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

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

Limborg, M. T., S. J. Helyar, M. De Bruyn, M. I. Taylor, E. E. Nielsen *et al.* 2012 Environmental selection on transcriptome□derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). Molec, Ecol. 21: 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x

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

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., 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. https://doi.org/10.1038/nature21347

Nosil, P., D. J. Funk, and D. Ortiz-Barrientos, 2009 Divergent selection and heterogeneous genomic divergence. Molec, Ecol. 18: 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Orsini, L., J. Vanoverbeke, I. Swillen, J. Mergeay, and L. De Meester, 2013 Drivers of population genetic differentiation in the wild: isolation by dispersal limitation,

isolation by adaptation and isolation by colonization. Molec, Ecol. 22, 5983–5999. https://doi.org/10.1111/mec.12561

Palsbøll, P. J., M. Berube, and F. W. Allendorf, 2007 Identification of management units using population genetic data. Trends Ecol. Evol. 22: 11–16. https://doi.org/10.1016/j.tree.2006.09.003

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

Petrou, E. L., J. E. Seeb, L. Hauser, M. J. Witteveen, W. D. Templin *et al.* 2014 Fine-scale sampling reveals distinct isolation by distance patterns in chum salmon (*Oncorhynchus keta*) populations occupying a glacially dynamic environment. Cons. Genet. 15: 229–243. https://doi.org/10.1007/s10592-013-0534-3

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.* 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

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 *et al.* 2002 Genetic structure of human populations. Science 298: 2381–2385. https://doi.org/10.1126/science.1078311

Rougemont, Q., J. S. Moore, T. Leroy, E. Normandeau, E. B. Rondeau *et al.* 2020 Demographic history shaped geographical patterns of deleterious mutation load in a broadly distributed Pacific Salmon. PLoS Genetics 16: e1008348.

https://doi.org/10.1371/journal.pgen.1008348

Rousset, F., 1997 Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145: 12193–1228.

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., 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.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molec. Biol. Evol. 4: 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Schwartz, M. K., G. Luikart, and R. S. Waples, 2007 Genetic monitoring as a promising tool for conservation and management. Trends Ecol. Evol. 22: 25–33. https://doi.org/10.1016/j.tree.2006.08.009

Slatkin, M. 1987 Gene flow and the geographic structure of natural populations. Science, 236: 787–792. https://doi.org/10.1126/science.3576198

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.

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

Waples, R. S., and O. Gaggiotti, 2006 What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15: 1419–1439. https://doi.org/10.1111/j.1365-294X.2006.02890.x

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. https://doi.org/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

Whitlock, M. C., and D. E. Mccauley, 1999 Indirect measures of gene flow and migration: $F_{\mathrm{ST}} \neq 1(4Nm + 1)$. Heredity 82: 117–125.

https://doi.org/10.1046/j.1365-2540.1999.00496.x

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

Table 1 Regression of genome-wide population-specific F_{ST} of 25 wild poplar populations on environmental variables

Variables	Estimate	SE	*Z*	*p*
DAY	0.0489	0.0164	2.99	0.003**
MAT	-0.0088	0.0086	-1.03	0.305
MAP	0.0001	0.0000	2.79	0.005**
SHM	0.0022	0.0009	2.38	0.018*

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

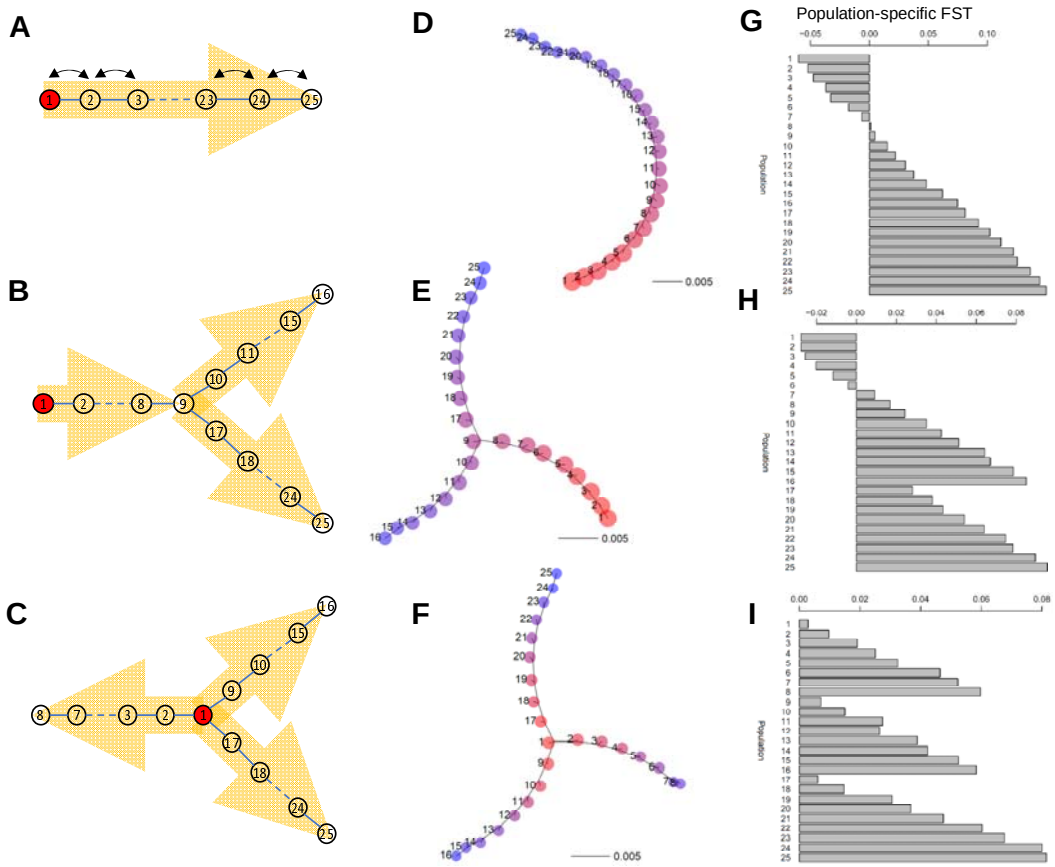

Figure 1 Results from population colonization simulations. Schematic diagrams of the models: (A) one-, (B) two-, and (C) three-directional colonization. Population 1 in red is ancestral, and arrows show the direction of colonization. Each population exchanged 1% of N_e genes with adjacent population(s) in every generation, as shown by arrows (see text). Neighbor-joining unrooted trees based on pairwise F_{ST} distance matrix overlaid with population-specific F_{ST} values for each model (D–F). The color of each population shows the magnitude of population-specific F_{ST} values. The radius of each population is proportional to the level of *He*, as visualized by $(H_e \times 10)^{2.5}$. Population-specific F_{ST} values for 25 simulated populations are presented for each model (G–I).

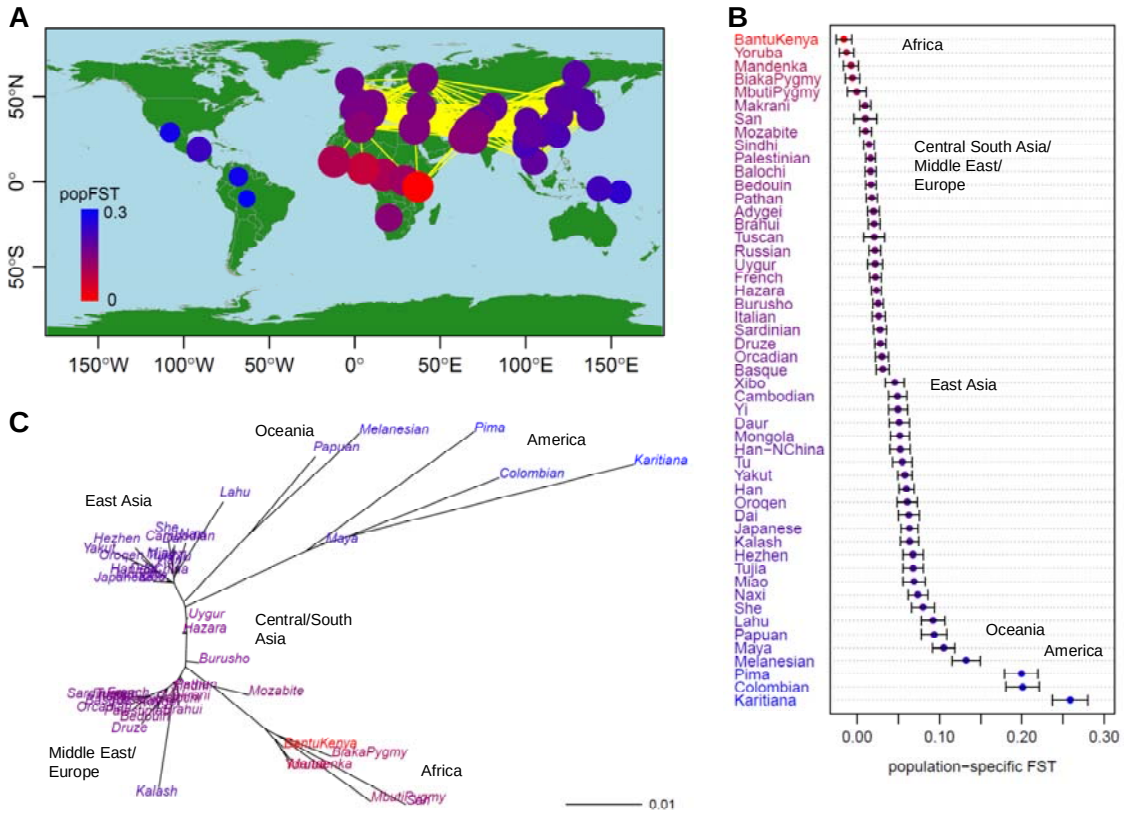

Figure 2 Population structure of 51 human populations (*n* = 1,035; 377 microsatellites). (A) Map showing genetic diversity and population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The radius of each sampling point is proportional to the level of *He*, as visualized by H_e^{12}. (B) Distribution of population-specific F_{ST} values ± 2×SE. (C) Unrooted NJ tree based on pairwise F_{ST} overlaid with population-specific F_{ST} values on population labels. The color of each population shows the magnitude of population-specific F_{ST} values.

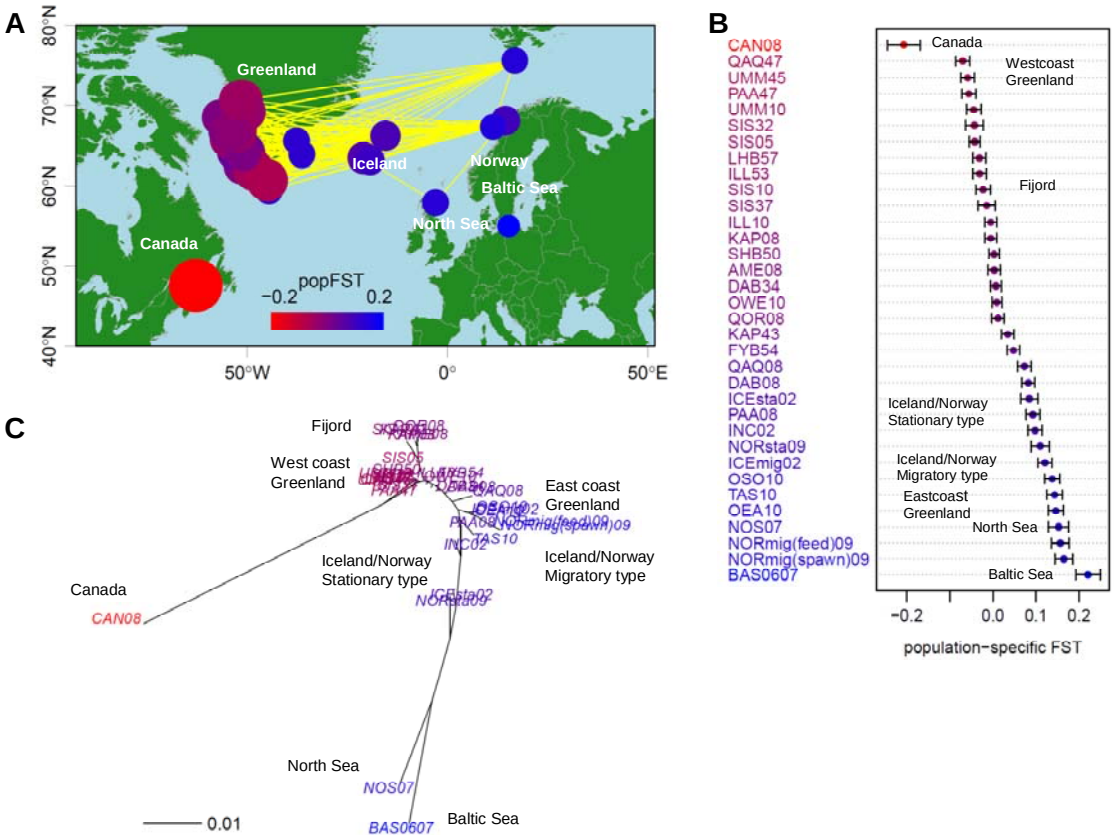

Figure 3 Population structure of 34 geographical samples of wild Atlantic cod (*n* = 1,065; 921 SNPs). (A) Map showing genetic diversity and population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The radius of each sampling point is proportional to the level of *He*, as visualized by H_e^{50}. (B) Distribution of population-specific F_{ST} values ± 2×SE. (C) Unrooted NJ tree based on pairwise F_{ST} overlaid with population-specific F_{ST} values on population labels. The color of each population shows the magnitude of population-specific F_{ST} values.

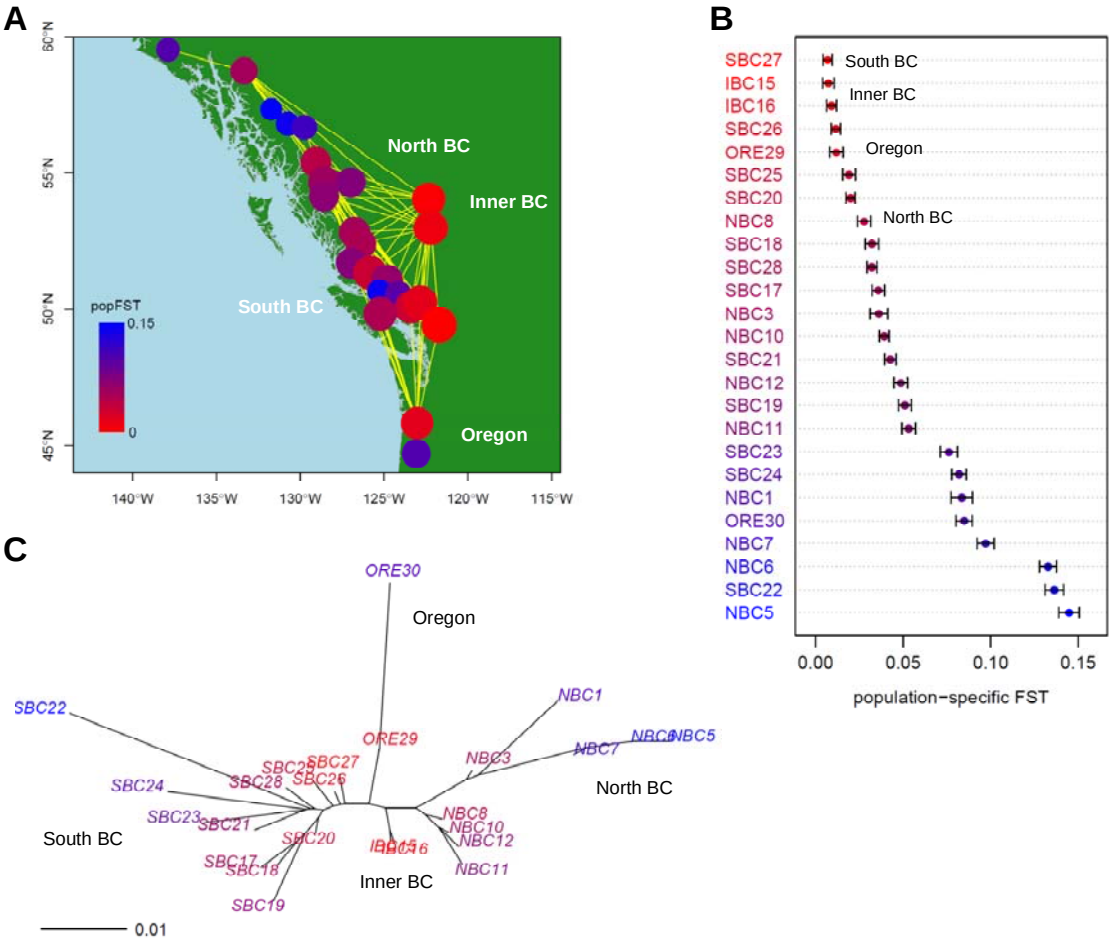

Figure 4 Population structure for 25 geographical samples of wild poplar (*n* = 441; 29,355 SNPs). (A) Map showing genetic diversity and population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The radius of each sampling point is proportional to the level of *He*, as visualized by H_e^{100}. (B) Distribution of population-specific F_{ST} values ± 2×SE. (C) Unrooted NJ tree based on pairwise F_{ST} overlaid with population-specific F_{ST} values on population labels. The color of each population shows the magnitude of population-specific F_{ST} values.

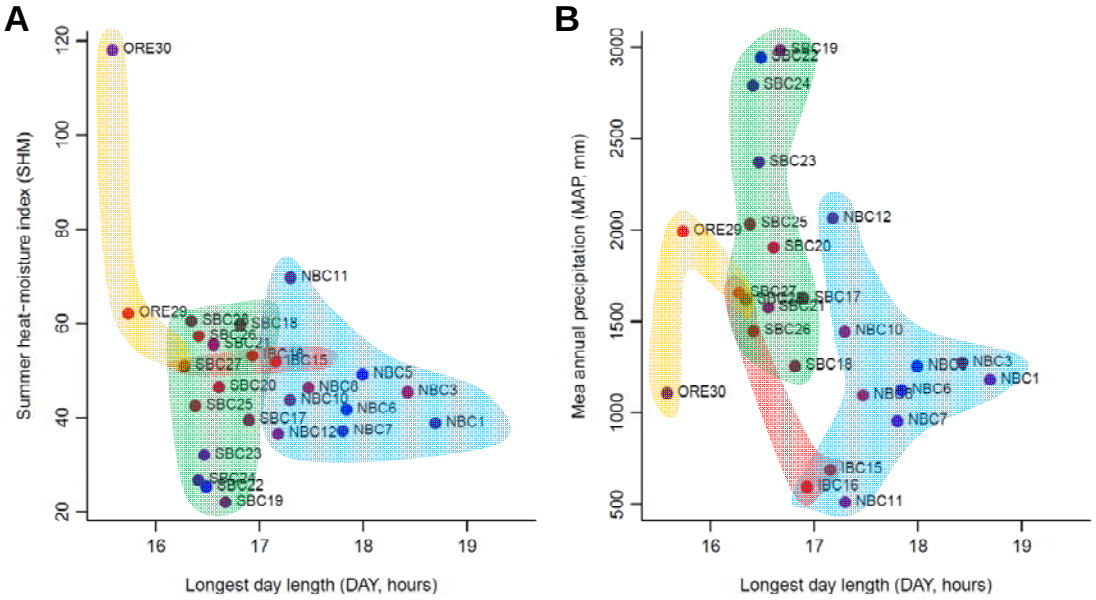

Figure 5 Population range expansion and environmental adaptation. Longest day length vs. (A) summer heat-moisture index and (B) mean annual precipitation for 25 geographical samples of wild poplar. The color of each population shows the magnitude of population-specific F_{ST} values. The circles show inferred population expansion from IBC15, IBC16, and SBC27. The color of the circles refers to the population clusters (see, Figure S5).

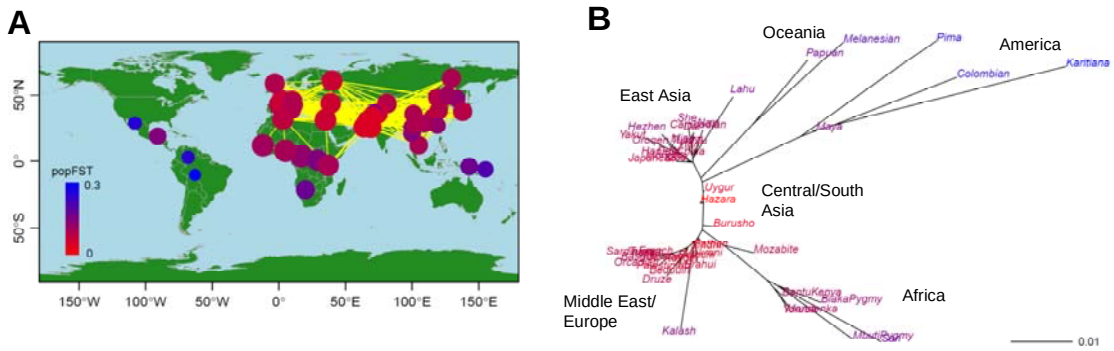

Figure 6 Population structure of 51 human populations inferred based on Bayesian population-specific F_{ST} values. (A) Map showing genetic diversity and population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The radius of each sampling point is proportional to the level of *He*, as visualized by H_e^{12}. (b) Unrooted NJ tree based on pairwise F_{ST} overlaid with population-specific F_{ST} values on population labels. The color of each population shows the magnitude of population-specific F_{ST} values.

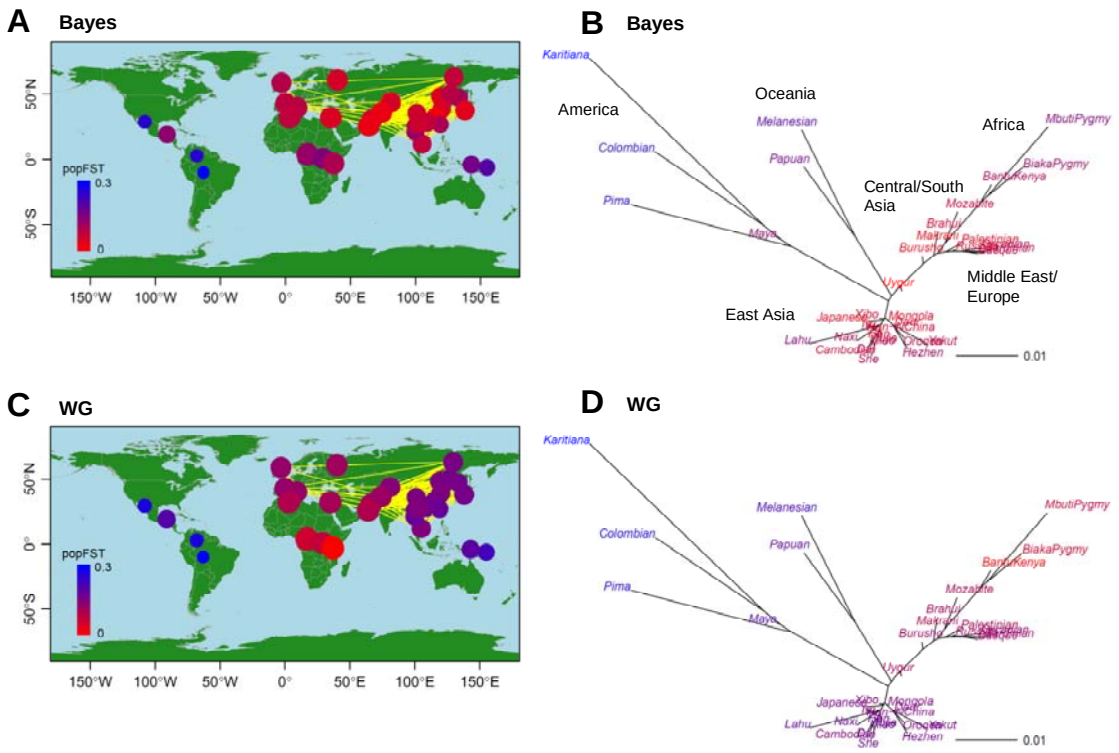

Figure 7 Population structure of 37 subsamples of human populations inferred based on Bayesian and WG population-specific F_{ST} values. The numbers of sampling locations of the subsamples were as follows: 3 from Africa, 6 from the Middle East/Europe, 9 from Central/South Asia, 18 from East Asia, 2 from Oceania, and 4 from America. Maps show genetic diversity and population connectivity with the magnitude of (A) Bayesian and (C) WG population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The radius of each sampling point is proportional to the level of *He*, as visualized by H_e^{12}. Unrooted NJ tree based on pairwise F_{ST} overlaid with (B) Bayesian and (D) WG population-specific F_{ST} values on population labels. The color of each population shows the magnitude of population-specific F_{ST} values.