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
§The Research Institute of Evolutionary Biology, Tokyo 138-0098, Japan

Short running title:
Population-specific and pairwise F_{ST}

Keywords:
demographic history, genetic diversity, migration, population structure, range expansion

Corresponding author:
Shuichi Kitada
Tokyo University of Marine Science and Technology, Minato-ku, Konan 4-5-7, Tokyo 108-8477, Japan
+81-297-45-5267
kitada@kaiyodai.ac.jp

Abstract

Populations are shaped by their history. 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 the current population structure and a population's history of range expansion, we propose a representation method that overlays population-specific F_{ST} estimates on a sampling location map, and on an unrooted neighbor-joining tree and a multi-dimensional scaling plot inferred from a pairwise F_{ST} distance matrix. We examined the usefulness of our procedure using simulations that mimicked population colonization from an ancestral population and by analyzing published human, Atlantic cod, and wild poplar data. Our results demonstrated that population-specific F_{ST} values identify the source population and trace the evolutionary history of its derived populations. Conversely, pairwise F_{ST} values represent the current population structure. By integrating the results of both estimators, we obtained a new picture of the 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 daylight hours are long in summer, to seashores with abundant rainfall, and to the south with dry summers. Genomic data highlight the power of the bias-corrected moment estimators of F_{ST}, whether global, pairwise, or population-specific, that provide unbiased estimates of F_{ST}. All F_{ST} moment estimators described in this paper have reasonable process times and are useful in population genomics studies. The R codes for our method and simulations are available in the Supplemental Material.

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). For such objectives, Wright's F_{ST} (Wright 1951) is commonly used to quantify the genetic divergence of populations, and there have been many informative reviews of F_{ST} estimators (*e.g.*, Rousset 2004, 2007; Balloux and Lugon-Moulin 2002; Weir and Hill 2002; Beaumont 2005; Excoffier 2007; Holsinger and Weir 2009; Gaggiotti and Foll 2010; Bhatia *et al*. 2013). The traditional F_{ST} estimators have been defined as the ratio of the between-population variance to the total variance of allele frequencies (Wright 1965; Cockerham 1969, 1973; Weir and Cockerham 1984; Balloux and Lugon-Moulin 2002; Excoffier 2007; Holsinger and Weir 2009). An alternative approach for estimating population differentiation is to use population-specific F_{ST} estimators (Balding and Nichols 1995; Nicholson *et al.* 2002; Weir and Hill 2002; Weir *et al.* 2005; Gaggiotti and Foll 2010; Weir and Goudet 2017). 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). In addition to model-based methods, moment estimators of population-specific F_{ST} have been derived (Weir and Hill 2002; Weir and Goudet 2017). A large number of approaches exist for estimating F_{ST} that have different underlying assumptions (global, pairwise, or population-specific F_{ST}) and the framework used, such as frequentist and/or Bayesian. There have been many comprehensive reviews of traditional and population-specific F_{ST} estimators, as indicated above, most of which were written from the viewpoint of theoretical issues. Conversely, there has been no formal comparative study to describe their differences in terms of the evolutionary scenarios that best describe the data. Although these issues are well understood among statistical geneticists and theoretical population geneticists, empirical researchers, particularly those working in non-model organisms, could benefit from studies that address the problem.

In practice, the F_{ST} value estimated from a set of population samples is called the global F_{ST} that measures population differentiation over all populations (*e.g.*, Pérez-Lezaun *et al*. 1997). Additionally, F_{ST} values between pairs of population samples (pairwise F_{ST}, Reynolds *et al*. 1983) 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 and ecology (*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 influence population structure (Nosil *et al*. 2009; Orsini *et al*. 2013), and researchers have 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). To identify the adaptive divergence of a gene among populations, locus-population-specific F_{ST} was developed using empirical Bayes (Beaumont and Balding 2004) and full Bayesian methods (BayeScan) (Foll and Gaggiotti 2008). The Bayesian methods have been applied in many cases of various species to identify outlier single-nucleotide polymorphisms (SNPs) (*e.g.*, Therkildsen *et al*. 2013a; Geraldes *et al*. 2014; Limborg *et al*. 2012). However, genome-wide population-specific F_{ST} is new among biologists. Despite the expected usefulness of genome-wide population-specific F_{ST} in

evolutionary biology (Weir and Goudet 2017), applications have been sparse to date (*e.g.*, Nicholson *et al.* 2002; Weir *et al.* 2005; Foll and Gaggiotti 2006; Buckleton *et al.* 2016; Rougemont *et al*. 2020).

Traditional F_{ST} estimators were originally developed to estimate F_{ST} over a metapopulation (global F_{ST}) based on a set of population samples (Cockerham 1969; Nei and Chesser 1983; Weir and Cockerham 1984; Excoffier 2007; Rousset 2007). In this study, we use Nei and Chesser's (1983) bias-corrected G_{ST} moment estimator (hereafter, NC83) as a pairwise F_{ST} estimator (Supplemental Note). The pairwise F_{ST} between populations (i, j) is defined as

$$\mathrm{pw}F_{\mathrm{ST}}^{ij} = \frac{H_T^{ij} - H_S^{ij}}{H_T^{ij}} = 1 - \frac{H_S^{ij}}{H_T^{ij}} \quad (1)$$

where H_T^{ij} is total heterozygosity over all populations and H_S^{ij} is within-population heterozygosity.

We apply Weir and Goudet's (2017) bias-corrected moment estimator of population-specific F_{ST} (hereafter, WG) (Supplemental Note). When only allele frequencies are used, the WG population-specific F_{ST} at a locus is defined as

$$\mathrm{ps}F_{\mathrm{ST}}^{i} = \frac{M_w^i - M^B}{1 - M^B}$$

where M_W^i is the within-population matching of two distinct alleles of population i and M^B is the between-population-pair matching average over pairs of populations i, i'. M^B is homozygosity over pairs of populations. We represent heterozygosity over all pairs of populations as $1 - M^B = H_B$, and $1 - M_W^i = H_{Si}$. Therefore,

$$\mathrm{ps}F_{\mathrm{ST}}^{i} = \frac{H_B - H_{Si}}{H_B} = 1 - \frac{H_{Si}}{H_B}. \quad (2)$$

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). H_B is heterozygosity for all pairs of populations, whereas H_T^{ij} in Equation 1 is heterozygosity for the pair of populations. Equation 2 shows that WG population-specific F_{ST} measures population-specific genetic diversity (H_{Si}) under the framework of the relatedness of individuals and identifies the population with the greatest genetic diversity as the ancestral or oldest population. Because populations close to the ancestral population have had more opportunities for mutations than recently founded populations (Liu *et al*., 2006), they are likely to have the highest heterozygosity and low values of population-specific F_{ST}. Thus, WG population-specific F_{ST} works to infer evolutionary history through genetic diversity in terms of heterozygosity under the assumption that the ancestral population had the highest genetic diversity. By combining population-specific and pairwise F_{ST} estimates, we can infer the present population structure, which reflects evolutionary history.

In this study, our objective is to demonstrate to empirical population geneticists and biologists how the two types of genome-wide F_{ST} estimators can be combined to help elucidate the population structure (pairwise F_{ST}) in the evolutionary context (population-specific F_{ST}). In our approach, the current population structure is represented by an unrooted neighbor-joining (NJ) tree (Saitou and Nei 1987) and a multi-dimensional scaling (MDS) plot based on pairwise F_{ST} values. The colors of the populations (names and/or sampling points) based on the WG genome-wide population-specific F_{ST} values enable the inference of the historical

order of population colonization. We also present a representation on a geographical map, which is useful for visually understanding population history in a distribution range.

First, we examine the usefulness of our procedure using stepping-stone simulations that mimic population colonization from a single ancestral population for five scenarios of population range expansion. We then apply our approach to three datasets of human, Atlantic cod (*Gadus morhua*), and wild poplar (*Populus trichocarpa*). 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; Hellenthal *et al*. 2014; Rutherford 2016; Nielsen *et al*. 2017). These patterns are well known by statistical/theoretical population geneticists and biologists; therefore, testing our integrative F_{ST} analysis on this dataset could provide a good example of the usefulness of the practical approach. In humans, genome-wide SNPs have been used to infer population structure and admixture (Pickrell and Pritchard 2012; Lipson *et al*. 2013; Hellenthal *et al*. 2014; Bradburd *et al*. 2016). Conversely, microsatellite markers, which have a larger number of alleles than SNPs, despite the smaller number of loci, are still usual tools in ecology and conservation biology (*e.g*., Sylvester *et al*. 2018; D'Aloia *et al*. 2020; Kattenberg *et al*. 2020; Karamanlidis *et al*. 2021). We use microsatellite data for illustrative purposes. The Atlantic cod SNP were genotyped from mature fish samples collected from the North Atlantic from the northern range margin of the species in Greenland, Norway, and the Baltic Sea. Two ecotypes (migratory and stationary) that were able to interbreed were genetically differentiated (Hemmer-Hansen *et al*., 2013a; Berg *et al*., 2016). The inclusion of both types of data may help the understanding of the demographic history of highly migratory marine fish. The wild poplar SNP data were collected from the American Pacific Northwest. The wind-pollinated tree can spread pollen over a wide area, and the samples covered various regions over a range of 2,500 km near the Canadian–US border at altitudes between 0 and 800 m. Each sampling location was associated with 11 environmental and geographical parameters. The analysis of environmental variables and population-specific F_{ST} values may provide a good example for understanding the history of the range expansion of a wind-pollinated tree. By analyzing different types of data with species-specific ecology and migration history, the usefulness of our approach may be identified to enable us to understand the current population structure in an evolutionary context.

We computed pairwise and population-specific F_{ST} estimators 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 Supplemental Material. They can be used for microsatellite and SNP genotype data in the Genepop format (Raymond and Rousset 1995; Rousset 2008), which has been particularly widely used among biologists.

MATERIALS AND METHODS

Population colonization simulations

To test the performance of our visual representation, we conducted simulations that mimicked the colonization of populations from a single ancestral population (population 1). We modeled five types of stepping-stone colonization: one, two, and three-directional population expansion; three-directional grid colonization from an edge; and eight-directional grid colonization from the center, with 24 demes (populations 2–25) (Figure 1A–E). 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. For convenience, we considered one simulation cycle to be one generation. The effective population size of the newly derived population increased to the same size as 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. We also examined the cases in which the effective population size of the ancestral population was 10 times greater ($N_e = 10^5$) than that of the newly derived population ($N_e = 10^4$).

We generated the initial allele frequencies in the ancestral population, q, at 100,000 neutral SNP loci from the predictive equilibrium distribution, $f(q) \propto q^{-1}(1-q)^{-1}$ (Wright 1931). Additionally, we introduced 10 newly derived SNPs to each existing population in each generation. 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 in the initial phase after their birth. We considered these 100,000 ancestral SNPs and newly derived SNPs to be "unobserved." We changed the allele frequencies of these SNPs using random drift under a binomial distribution in every generation. The frequencies of the derived alleles reduced for many of the SNPs over simulation generations and lost their polymorphism. After 260 simulation generations, we randomly selected SNPs that retained their polymorphism as "observed" SNPs. For grid colonization models (Figure 1D–E), we randomly selected polymorphic SNPs after 100 simulation generations. In this simulation, we selected 10,000 ancestral SNPs and 500 newly derived SNPs. Then, we generated 50 individuals for each population. We randomly generated the genotypes of these 10,500 SNPs 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. To examine the effect of generations on genetic diversity in newly derived SNPs, we also selected 9,000 and 7,000 ancestral SNPs, and 1,000 and 3,000 newly derived SNPs, respectively. We converted the simulated genotypes into Genepop format (Raymond & Rousset, 1995; Rousset, 2008). We then computed genome-wide population-specific and pairwise F_{ST} values between the 25 populations.

Visual representation of population structure and demographic history

We integrated genome-wide population-specific and pairwise F_{ST} estimates on a map of sampling locations on an NJ tree and MDS plot. We visualized the magnitude of the genome-wide population-specific F_{ST} values using a color gradient 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 standardized magnitude of a population-specific F_{ST} value at the sampling point, with colors between red (for the smallest F_{ST}) and blue (for the largest F_{ST}). We drew the F_{ST} map using the `sf` package in R, where we plotted sampling locations based on the longitudes and latitudes. We connected sampling points with genome-wide pairwise F_{ST} values smaller than a given threshold using 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), $F_{ST} \approx \frac{1}{4N_e m+1}$, where N_e is the effective population size and m is the average rate of migration between populations (Slatkin 1987). For the threshold of $F_{ST} =$

0.02, there are $4N_e m \approx 49$ migrants per generation (see Whitlock and Mccauley 1999; Waples and Gaggiotti 2006). We plotted the genome-wide population-specific F_{ST} values on a dot chart with standard errors estimated using Equation S5 (Supplemental Note). We drew the NJ tree based on the distance matrix of the genome-wide pairwise F_{ST} values using the `nj` function in the R package ape. We performed MDS analysis on the pairwise F_{ST} distance matrix using the `cmdscale` function in R. We used the cumulative contribution ratio up to the kth axis $(j = 1, \ldots, k, \ldots, K)$ as the explained variation measure, which we computed using the R 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 $\lambda_j < 0$. We colored the sampling locations on the F_{ST} maps, dot charts, NJ trees, and MDS plots using a color gradient of the magnitude of genome-wide population-specific F_{ST} values. We also examined a diverging color palette instead of blue to red to test the resolution using RColorBrewer on the F_{ST} maps.

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. We computed genome-wide pairwise F_{ST} values (NC83, Equation S3) using the `pop_pairwiseFST` function in FinePop2. We calculated expected heterozygosity for each population using the `read.GENEPOP` function. We computed the genome-wide WG population-specific F_{ST} (Equation S4) values using the `pop_specificFST` function. We applied Bayesian population-specific F_{ST} estimators on human data. We maximized Equation S7 and estimated the empirical Bayesian population-specific F_{ST} (Beaumont and Balding 2004) at each locus according to Equation S8. Then we averaged these values over all loci. For the full Bayesian model, we used GESTE_ver. 2.0 (Foll and Gaggiotti 2006) to compute the genome-wide population-specific F_{ST} values. We examined the shrinkage effect of the Bayesian population-specific F_{ST} estimator on inferring the ancestral population using a set of subsamples (37 populations) chosen from 51 populations.

Inferring environmental effects on the observed population structure

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 $(j = 1, \ldots, s)$:

$$\mathrm{ps}\hat{F}_{\mathrm{ST}}^{i} = \beta_0 + \beta_1 x_1 + \cdots + \beta_s x_s + \varepsilon_i, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Omega), \quad (i = 1, \ldots, r), \quad (3)$$

where Ω is the variance matrix of population-specific F_{ST}. We correlated residuals because of the population structure; therefore, the effective sample size was 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 for generalized least squares (GLS). We performed this analysis on the wild poplar dataset, for which 11 environmental/geographical parameters were available for each sampling location. We used Equations S5 and S6 for the components of the variance matrix Ω in Equation 3. We performed regression using the `GLS` function in FinePop2_ver.0.2.

Three empirical datasets

We retrieved the human microsatellite data used in Rosenberg *et al*. (2002) from https://web.stanford.edu/group/rosenberglab/index.html. 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. We obtained the longitudes and latitudes of the sampling sites from Cann *et al*. (2002) (Table S1).

We combined the Atlantic cod SNP genotype data of 924 SNPs common to 29 populations reported in Therkildsen *et al*. (2013a, b) and 12 populations reported in Hemmer-Hansen *et al*. (2013a, b). We compared the genotypes associated with each marker in samples that were identical in the two studies, that is, 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 the genotypes were missing in several population samples in Therkildsen *et al*. (2013b). For simplicity, we removed temporal replicates from the Norway migratory, Norway stationary, North Sea, and Baltic Sea samples. The final dataset consisted of genotype data (n = 1,065) for 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. All individuals in the samples were adults, and most were mature (Therkildsen *et al*., 2013a). We used the longitudes and latitudes of the sampling sites in Hemmer-Hansen *et al*. (2013a). For the data from Therkildsen *et al*. (2013a), we estimated approximate sampling points from the map of the original study and recorded the longitudes and latitudes (Table S2).

We retrieved wild poplar SNP genotype data and environmental/geographical data 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 the array development and selection of SNPs are provided in Geraldes *et al*. (2011, 2013). A breakdown of the 25 drainages (hereafter, populations) is as follows: 9 in northern British Columbia (NBC), 2 in inland British Columbia (IBC), 12 in southern British Columbia (SBC), and 2 in Oregon (ORE) (Geraldes *et al*., 2014). We combined the original names of the clusters and population numbers, and used them as our population labels (NBC1, NBC3,…, ORE30). We associated each sampling location 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.

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 material.

RESULTS

Simulations of population colonization

First, we examined the effect of the number of simulation generations on genetic diversity in newly derived SNPs using the eight-directional grid simulation (Figure 1E). WG population-specific F_{ST} correctly identified the ancestral population and traced the population

history, and population structure reflected the population history regardless of the numbers of ancestral SNPs (9,000 and 7,000) and newly derived SNPs (1,000 and 3,000) selected after 100 simulation generations (Figures S1). The result was consistent with the case that used 10,500 SNP loci (10,000 ancestral SNPs + 500 newly derived SNPs) (Figures 1J, O, S2E). In the following analysis, we used the results based on 10,500 SNP loci to generate clearer results, even for limited numbers of simulation generations (260 and/or 100).

In the one-directional simulation (Figure 1A), population-specific F_{ST} correctly identified the ancestral population with the highest genetic diversity (Figure S2A), and populations were located in order from 1 to 25 on the NJ tree (Figure 1F). The first axis of the MDS plot explained 93% of the variance of the pairwise F_{ST} distance matrix and indicated population expansion from population 1 to population 25 (Figure 1K). In the two-directional simulation (Figure 1B), our analysis correctly identified the ancestral population (Figure S2B) and detected that populations were split at population 9 and expanded in two directions (Figure 1G), which was consistent with the simulation scenario. The first axis of the MDS plot identified population expansion from population 1 to population 25 and explained 56% of the variance of the pairwise F_{ST} distance matrix, whereas the second axis identified another manner of population expansion from population 1 to population 16 and explained 40% of the variation (Figure 1L). In the three-directional simulation (Figure 1C), the ancestral population was also correctly identified (Figure S2C). It was closely located to the adjacent populations 2, 9, and 17, but correctly detected three directions (Figure 1H). The first axis of the MDS plot identified population expansion from population 1 to populations 16 and 25, and explained 57% of the variance of pairwise F_{ST}, whereas the second axis identified population expansion from population 1 to populations 8 and 16 and explained the 38% variation (Figure 1M).

In the three-directional grid colonization model from an edge (Figure 1D), population-specific F_{ST} correctly identified the ancestral population (Figure S2D) and pairwise F_{ST} detected that populations expanded in three directions (Figure 1I), which agreed with the simulation scenario. The first axis of the MDS plot identified population expansion from population 1 to other edge populations (populations 11, 15, and 25), and explained 60% of the variance of the pairwise F_{ST} distance matrix, whereas the second axis indicated genetic differentiation between populations 24, 25 and 15, 19, 22, and explained 17% of the variation (Figure 1N). In the eight-directional grid colonization model (Figure 1E), population-specific F_{ST} identified the ancestral population (Figure S2E) and pairwise F_{ST} estimated that populations expanded in five directions from the center (Figure 1J). The first axis of the MDS plot identified vertical population expansion from population 1 to populations 24 and 25 and explained 40% of the variance of the pairwise F_{ST} distance matrix, and the second axis indicated horizontal population expansion from population 1 to populations 24 and 23, which explains 30% of the variation (Figure 1O). We obtained similar results in the cases in which the effective population size of the ancestral population was 10 times greater ($N_e = 10^5$) than that of the newly derived population ($N_e = 10^4$) (Figure S3). We obtained very similar results from different data for more than 20 simulations (figures not shown).

Humans

The F_{ST} map (Figure 2A) shows integrated information from genome-wide population-specific and pairwise F_{ST}, which visualizes population structure with the migration and demographic history of human populations in terms of genetic diversity. Interestingly, Bantu Kenyans had the smallest F_{ST} value (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 the sampling points connected by yellow lines with pairwise F_{ST} values below the 0.02 threshold (Figure 2A), genetic connectivity from Africa was low. Conversely, migration was substantial within Eurasia but much smaller than that inferred from Eurasia to Oceania and America. 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 indicated 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 S4). The first axis of the MDS plot highlighted difference between African and American populations and explained 44% of the variance of the pairwise F_{ST} distance matrix, whereas the second axis indicated genetic differentiation between Melanesian and Karitiana populations, and explained the 19% variation (Figure 2D).

The 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 and the smallest F_{ST} values observed in the Middle East, Europe, and Central/South Asia (Figure S5A, Table S4). However, in African populations, they were higher than the WG population-specific F_{ST} values (Figure S5B). Our F_{ST} map based on the empirical Bayesian population-specific F_{ST} values indicated that the Middle East, Europe, and Central/South Asia were centers of human origin (Figure 3A), which was consistent with that from the full Bayesian population-specific F_{ST} estimator (figure not shown). Our integrated NJ tree showed that the Hazara, Pakistan population was genetically closest to the human ancestors (Figure S6A). 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. 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 S5), the area with the highest population-specific F_{ST} values shifted toward Central/South Asia and East Asia (Figure 3B), whereas Bantu Kenyans had the smallest WG population-specific F_{ST} value (Figure 3C); this was consistent with the results from the full dataset (Figure 2A). The integrated NJ trees provided similar results (Figure S6B, C).

Atlantic cod

The F_{ST} map (Figure 4A) visualizes the population structure, migration, and genetic diversity of the Atlantic cod populations. The Canadian population had the smallest population-specific F_{ST} value (shown in red) and the greatest H_e. H_e was also high in Greenland, low in other areas, and lowest in the Baltic Sea. Figure 4B shows the order of population-specific F_{ST} values from Canada to the Baltic sea (Table S6). Greenland west coast populations (green in Figure S7) generally had small population-specific F_{ST} values, whereas fjord populations (violet) had relatively higher population-specific F_{ST} values. The population-specific 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 4A), substantial migration was suggested between Greenland, Iceland, and Norway. Conversely, migration could be low from Canada to Greenland and from Iceland and Norway to the North and Baltic Seas. Our integrated NJ tree with population-specific F_{ST} values (Figure 4C) inferred 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. According to the ordinal NJ tree of the pairwise F_{ST} distance matrix (Figure S7), the populations were divided 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 (in purple) formed a sub-cluster within the Greenland west coast, and migratory (orange) and stationary (magenta) ecotypes also formed a sub-cluster. The first axis of the MDS plot characterized the differentiation between Canadian and North Sea/Baltic Sea populations and explained 72% of the variance of the pairwise F_{ST} distance matrix, whereas the second axis highlighted the differentiation between Norwegian migratory type populations from Canadian and North Sea/Baltic Sea populations, which explains the 22% variation (Figure 4D).

Wild poplar

The F_{ST} map (Figure 5A) indicated that population-specific F_{ST} values were lowest in southern British Columbia (SBC27) and inner British Columbia (IBC15 and IBC16) (shown in red, Figure S8). The sampling points connected by yellow lines (< 0.02 threshold pairwise F_{ST} values) indicated migration between all populations. H_e was highest in SBC27, IBC15, and IBC16, and lowest in northern British Columbia (NBC5). Figure 5B shows that samples collected from areas close to the SBC coast had higher population-specific F_{ST} values than other SBC samples (Table S7). The NBC samples had population-specific F_{ST} values similar to those of SBC. Among the 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 5C) suggested that wild poplar originated from around SBC27 and Inner British Columbia, and expanded in three directions: 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 four large clusters: 1) IBC, 2) SBC, 3) NBC, and 4) ORE (Figure S8). The population represented by sample ORE30 was isolated from ORE29. The first axis of the MDS plot characterized the differentiation between southern and northern populations and explained the 51% variance of the pairwise F_{ST} distance matrix, whereas the second axis characterized the southernmost ORE30 population, and explained the 18% variation (Figure 5D).

To avoid multicollinearity, we excluded seven out of 11 environmental variables that were significantly correlated with each other: lat, lon, alt, FFD, MWMT, MSP, and AHM (Table S3). 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 might reflect the directions of population expansion. The scatter plot of DAY and SHM (each population colored by the population-specific F_{ST} value) (Figure 6A) suggested three directions of population range expansion; the wild poplar that might originate from around SBC27 and IBC15 expanded its distribution to NBC, where daylight hours are long in summer; expanded to SBC seashores with lower SHM (humid summer and abundant rainfall), and to the south (ORE29 and ORE30) with higher SHM (dry summer). This was consistent in the scatter plot of DAY and

MAP (Figure 6B).

Diverging color palette

RColorBrewer had 35 color palettes. Each palette had a minimum of eight colors. Two palettes had 12 colors (maximum) and nine had 11 colors. We chose a color palette with 10 colors (RdYlBu). The color gradient better identified the middle range of population-specific F_{ST} values, but failed to detect the ancestral population (Figure S9).

CPU times

Using a laptop computer with an Intel Core i7-8650U CPU, only 89.8 s of CPU time was required to compute the 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 the pairwise F_{ST} (NC83) between all population pairs. Based on these results, 50 min may be required to compute WG population-specific F_{ST} and 70 min to compute pairwise NC83 F_{ST} estimates for 1 million SNPs using that laptop. This computation may be much faster if we execute it on a workstation.

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 identified the source population and traced the evolutionary history of its derived populations based on genetic diversity (heterozygosity estimated from each population). The NC83 pairwise F_{ST} estimator correctly estimated the current population structure. As explained in the Introduction, the population-specific F_{ST} estimator is a rescaling of expected heterozygosity, and we expect a linear relationship between expected heterozygosity and population F_{ST}. This shows that the population-specific F_{ST} estimator implicitly assumes that populations closest to the ancestral population have the highest heterozygosity. In our three case studies, a linear relationship between H_e of each population ($= H_{Si}$) and $\mathrm{ps}\hat{F}^{i}_{\mathrm{ST}}$ was evident (Figure S10). 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 the sample size (number of individuals). Our simulations evaluated the performance of the population-specific F_{ST} estimator for such cases. However, in populations that experienced extensive admixture events, heterozygosity was enhanced, whereas a bottleneck in the ancestral population reduced heterozygosity. In such cases, the population-specific F_{ST} estimator misidentifies the ancestral population.

In our analysis, the 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).

Our analysis indicated that Atlantic cod might originate in Canada (CAN08). Figure 4 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 Figures 4C and S7. In our study, CAN08 had the highest H_e, which was lower in Iceland than in Greenland; 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, suggesting 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). In the Atlantic cod case study, CAN08 had the highest H_e 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 S10, Table S6). The WG population-specific F_{ST} value can be negative (Weir and Goudet 2017). In the one and two-directional models of our simulations, the WG population-specific F_{ST} value was significantly negative in the ancestral population, whereas H_e was the largest (Figure S2A,B). 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, a relatively large H_e ($\widehat{H}_{Si}$) is maintained in the source population. In such cases with $\widehat{H}_{Si} > \widehat{H}_B$, Equation 2 produces negative values for population-specific F_{ST}.

Although the wild poplar samples used in this study might not cover the entire distribution range of the species, which extends from southern California to northern Alaska, Montana, and Idaho (Geraldes *et al*. 2013), the genome-wide population-specific F_{ST} values suggested three directions of population expansion of wild poplar: from southern British Columbia (SBC27) and inland British Columbia (IBC15, IBC16) to coastal British Colombia, southern Oregon, and northern British Columbia (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).

Our continuous color gradient from red to blue successfully detected the ancestral population, whereas the R diverging color palettes, which had a limited number of colors (maximum of 12), better identified the middle range of population-specific F_{ST} values (Figure S9). Both color gradients may be useful.

The two types of genomewide F_{ST}s enabled us to infer the current population structure and the history of range expansion, which can also be calculated for the windows in the genomes. By comparing the pairwise F_{ST} among windows, it is possible to identify the genomic regions that are largely different between the populations (*e.g.*, Yi *et al*. 2010). Likewise, by comparing population-specific F_{ST} among windows, it may be possible for us to identify unique populations and their corresponding unique genomic regions.

Genome-wide population-specific F_{ST} detects key environments that relate to population expansion

578 Our GLS of genome-wide population-specific F_{ST} values revealed that long daylight hours,
579 abundant rainfall, and dry summer conditions are the key environmental factors that relate to
580 the demographic history of wild poplar (Table 1). Wild poplar could have expanded its
581 distribution by its fluffy seeds being blown away by the wind. In the NJ unrooted tree, the
582 root of the populations cannot be fixed without out-group populations. There is no direct
583 evidence, but we can infer from the population-specific F_{ST} values that wild poplar seems to
584 have spread from southern British Columbia (SBC27) and inland British Columbia (IBC15,
585 16), and expanded its distribution to NBC, where daylight hours are long in summer, to SBC
586 seashores with the rainy environment, and to southern Oregon (ORE30) with dry summer
587 conditions (Figures 5,6). A previous study on wild poplar revealed that genes involved in
588 drought response were identified as F_{ST} outliers using Bayescan (Foll and Gaggiotti 2008)
589 (Geraldes *et al*. 2014). The F_{ST} outlier test of Geraldes and colleagues also revealed that genes
590 involved in the circadian rhythm and response to red/far-red light had high locus-specific
591 global F_{ST} values. The first principal component of SNP allele frequencies of the polar tree
592 was significantly correlated with day length, and a previous enrichment analysis for
593 population structuring uncovered genes related to the circadian rhythm and photoperiod
594 (McKown *et al*. 2014a). Our results were in agreement with previous findings, showing the
595 usefulness of using the GLS of genome-wide population-specific F_{ST} to infer environmental
596 effects on the population expansion of species.
597

598 **Genome-wide F_{ST} moment estimators converge to their true means**

599 Previous studies have suggested or indicated that the "ratio of averages" works better than the
600 "average of ratios" as the number of independent SNPs increases (Reynolds *et al*. 1983; Weir
601 and Cockerham 1984; Bhatia *et al*. 2013). Because "the combined ratio estimate (ratio of
602 averages) is much less subject to the risk of bias than the separate estimate (average of ratios)"
603 (Cochran 1977), scholars recommend using the "ratio of averages" estimators (Bhatia *et al*.
604 2013). To explicitly show the underlying mechanism, we used the observed heterozygosity of
605 population i ($\widehat{H}_{Si}$) as derived by Nei and Chesser (1983) (Supplemental Note). When the
606 number of loci (L) increases, the average observed heterozygosity over all loci converges to
607 its expected value according to the law of large numbers as

608 $$\frac{1}{L}\sum_{l=1}^{L}\left(1-\sum\nolimits_{u=1}^{m}\tilde{p}_{iu}^{2}\right)\rightarrow\frac{1}{L}\sum_{l=1}^{L}\left(1-E\left[\sum\nolimits_{u=1}^{m}\tilde{p}_{iu}^{2}\right]\right)$$

609 The observed gene diversity thus converges to the expected value:

610 $$\widehat{H}_{Si}\left(1-\frac{1}{n_i}\right)+\frac{\widehat{H}_{0i}}{2n_i}\ \rightarrow H_{Si}\left(1-\frac{1}{n_i}\right)+\frac{H_{0i}}{2n_i}$$

611 Similarly, $\widehat{H}_S$ and $\widehat{H}_T$ converge to their expected values. This example indicates that the
612 numerators and denominators of bias-corrected F_{ST} moment estimators, whether global,
613 pairwise, or population-specific, converge to their true means and provide unbiased estimates
614 of F_{ST} in population genomics analyses with large numbers of SNPs.
615

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

618 In the Bayesian framework, the population specific F_{ST} is the coefficient of the genetic drift
619 that represent the among-population variation of the allele frequencies at neutral loci from the
620 allele frequencies of the ancestral population. The allele frequency of the ancestral population
621 is assumed to be the among-population mean allele frequencies. The analysis of human
622 population (Figure 3A) provokes the need to take account of geographical heterogeneity of

sampling fraction of populations. The unbiased estimate of the allele frequency of the ancestral population will be obtained as the weighted among-population average. The weights are inversely proportional to the sampling fractions.

The shrinkage effect on allele frequencies in Bayesian inference (Stein, 1956) may shift population-specific F_{ST} values toward the average of the entire population. Because of the shrinkage toward mean allele frequencies, the maximum likelihood and Bayesian estimators of locus-specific global F_{ST} improve the power to detect genes under environmental selection (Beaumont and Balding 2004; Foll and Gaggiotti 2008). An empirical Bayes genome-wide pairwise F_{ST} estimator (Kitada *et al*. 2007) is useful in cases involving a small number of polymorphic marker loci, particularly in high gene flow scenarios, but it suffers from the shrinkage effect when larger numbers of loci are used. The shrinkage of allele frequencies should affect inference in genome-wide population-specific F_{ST}, particularly in cases when samples (populations) were not a representative of the populations.

Conclusions

The WG genome-wide population-specific F_{ST} moment estimator can identify the source population and trace the evolutionary history of the derived populations based on genetic diversity under the assumption that populations closest to the ancestral population have the highest heterozygosity. Conversely, the NC83 genome-wide pairwise F_{ST} moment estimator represents the current population structure. By integrating population-specific and pairwise estimates on F_{ST} maps, NJ trees, and MDS plots, we obtain a picture of population structure by incorporating evolutionary history. Our GLS analysis of genome-wide population-specific F_{ST}, which takes the correlation between populations into account, provides insights into how a species expanded its distribution in different environments. 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}.

ACKNOWLEDGMENTS

We express our appreciation to editors Andrew Kern, Jeffrey Ross-Ibarra, Mark Beaumont, and Nicholas Barton for reviewing this manuscript and providing constructive comments. We also thank the reviewers for providing essential comments that significantly improved the earlier versions of 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

Bradburd, G. S., P. L. Ralph, and G. M. Coop, 2016. A spatial framework for understanding population structure and admixture. *PLoS Genetics* **12**: e1005703. https://doi.org/10.1371/journal.pgen.1005703

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

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

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

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

D'Aloia, C. C., Andrés, J. A., Bogdanowicz, S. M., McCune, A. R., Harrison, R. G. *et al.* 2020 Unraveling hierarchical genetic structure in a marine metapopulation: A comparison of three high‐throughput genotyping approaches. *Molec. Ecol.* **29**: 2189–2203. https://doi.org/10.1111/mec.15405

Excoffier, L., 2007 Analysis of population subdivision, pp. 980–1020 in *Handbook of Statistical Genetics, 3rd ed.* edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

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

Hellenthal, G., G. B. Busby, G. Band, J. F. Wilson, C. Capelli *et al.* 2014. A genetic atlas of human admixture history. *Science* **343**: 747–751. https://doi.org/10.1126/science.1243518

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

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

Karamanlidis, A.A., T. Skrbinšek, G. Amato, P. Dendrinos, S. Gaughran *et al*. 2021 Genetic and demographic history define a conservation strategy for earth's most endangered pinniped, the Mediterranean monk seal *Monachus monachus*. *Sci. Rep.* **11**: 373. https://doi.org/10.1038/s41598-020-79712-1

Kattenberg, J. H., Z. Razook, R. Keo, C. Koepfli, C. Jennison *et al.* 2020 Monitoring *Plasmodium falciparum* and *Plasmodium vivax* using microsatellite markers indicates limited changes in population structure after substantial transmission decline in Papua New Guinea. *Molec. Ecol.* **29**: 4525–4541. https://doi.org/10.1111/mec.15654

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

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

Lipson, M., P. R. Loh, A. Levin, D. Reich, N. Patterson *et al.* 2013 Efficient moment-based inference of admixture parameters and sources of gene flow. *Molec. Biol. Evol.* **30**: 1788–1802. https://doi.org/10.1093/molbev/mst099

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

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.* **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

Pickrell, J., and J. K. Pritchard, 2012 Inference of population splits and mixtures from

genome-wide allele frequency data. *PLOS Genetics* **8**: e1002967. https://doi.org/10.1371/journal.pgen.1002967

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.

Reynolds, J., B. S., Weir, and C. C. Cockerham, 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics, **105**: 767–779. https://doi.org/10.1093/genetics/105.3.767

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

Rousset, F., 2007 Inferences from spatial population genetics, pp. 945–979 in *Handbook of Statistical Genetics, 3rd ed.* edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

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.

Sylvester, E. V., R. G. Beiko, P. Bentzen, I. Paterson, J. B. Horne *et al.* 2018 Environmental extremes drive population structure at the northern range limit of Atlantic salmon in North America. *Mol. Ecol.* **27**: 4026–4040. https://doi.org/10.1111/mec.14849

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
Wright, S., 1965 The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution **19**: 395–420. https://doi.org/10.2307/2406450
Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo *et al.* 2010 Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**: 75–78. https://doi.org/10.1126/science.1190371

890

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

Variable	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*

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

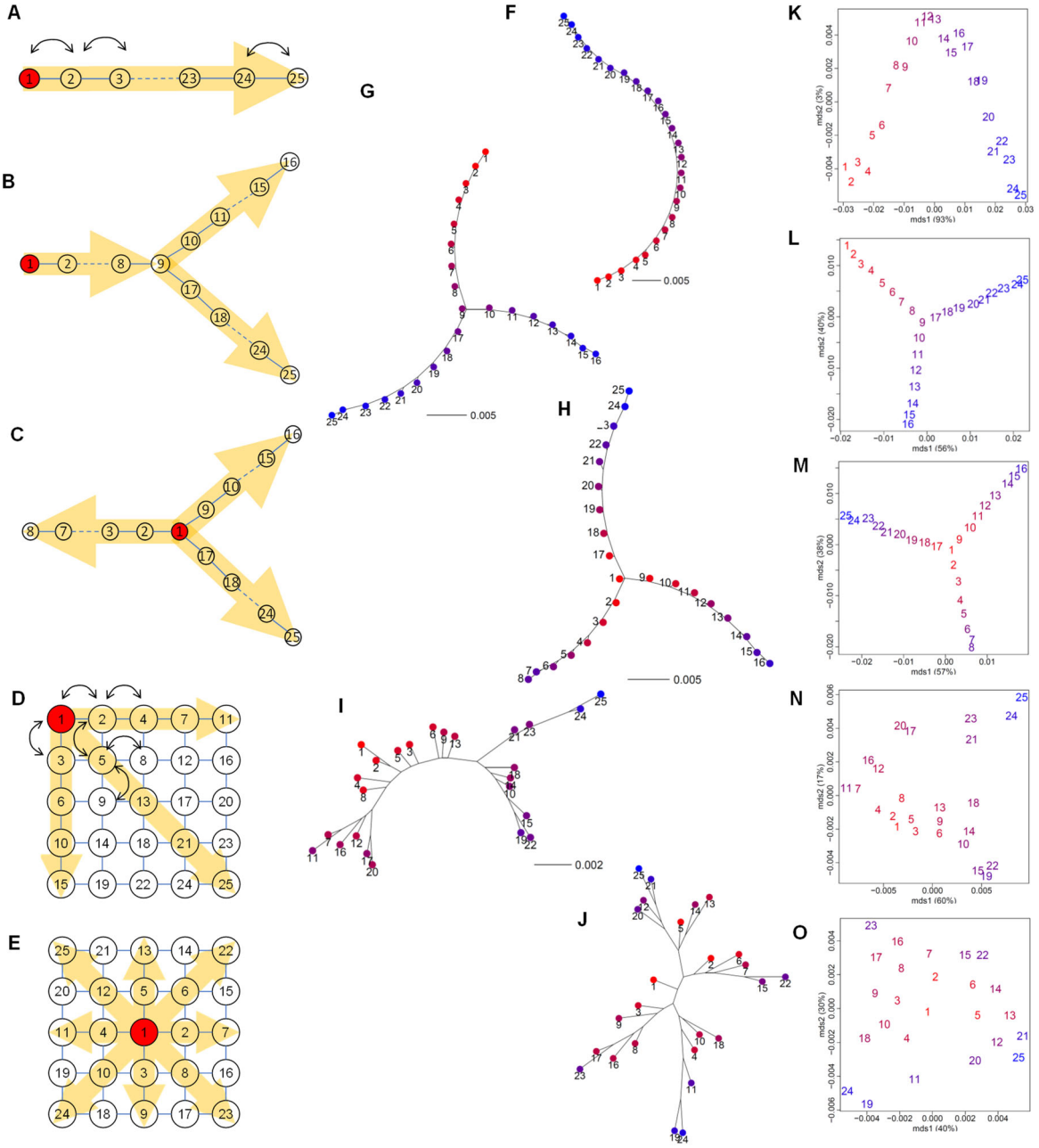

Figure 1 Results from population colonization simulations. Schematic diagrams of the models: (A) one, (B) two, (C) three-directional colonization, (D) three-directional grid colonization, and (E) eight-directional grid colonization. Population 1 in red is ancestral, and the yellow arrows indicate the direction of colonization. Lines show opportunities for migration. The effective population size of the newly derived population increased to the same size as the ancestral population ($N_e = 10^4$) after one simulation generation, and each population exchanged 1% of N_e genes with adjacent population(s) in every generation, as indicated by the arrows (see the text). Neighbor-joining (NJ) unrooted trees (F–J) and multi-dimensional scaling (MDS) plots (K–O) based on the pairwise F_{ST} distance matrix overlaid with population-specific F_{ST} values for each model. The color of each population shows the magnitude of population-specific F_{ST} values between red (for the smallest F_{ST}) and blue (for the largest F_{ST}).

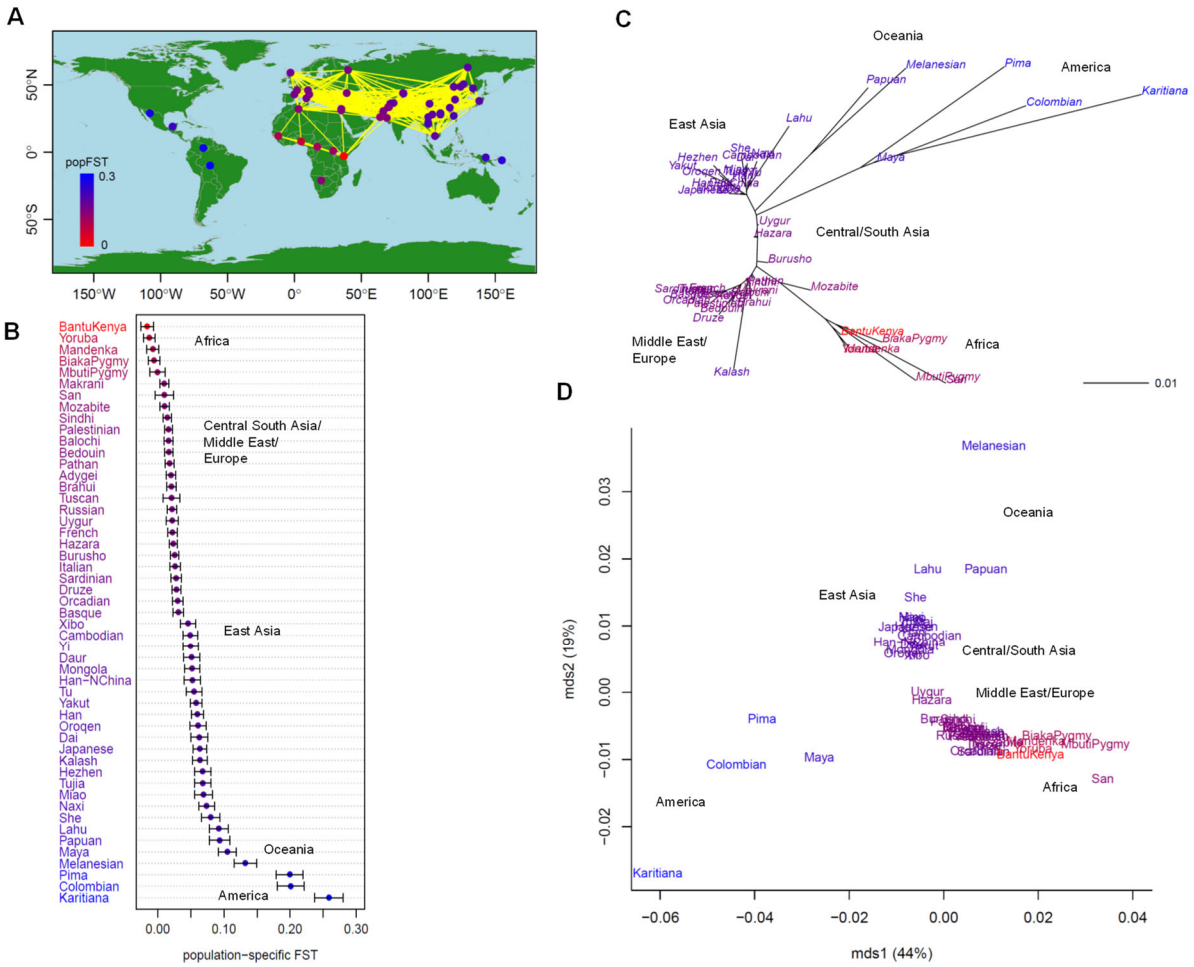

Figure 2 Population structure of 51 human populations (n = 1,035; 377 microsatellites). (A) Map showing population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. (B) Distribution of population-specific F_{ST} values ± 2×SE. (C) Neighbor-joining (NJ) unrooted tree and (D) multi-dimensional scaling (MDS) based on pairwise F_{ST} overlaid with population-specific F_{ST} values on population labels. The color of each population indicates the magnitude of population-specific F_{ST} values between red (for the smallest F_{ST}) and blue (for the largest F_{ST}).

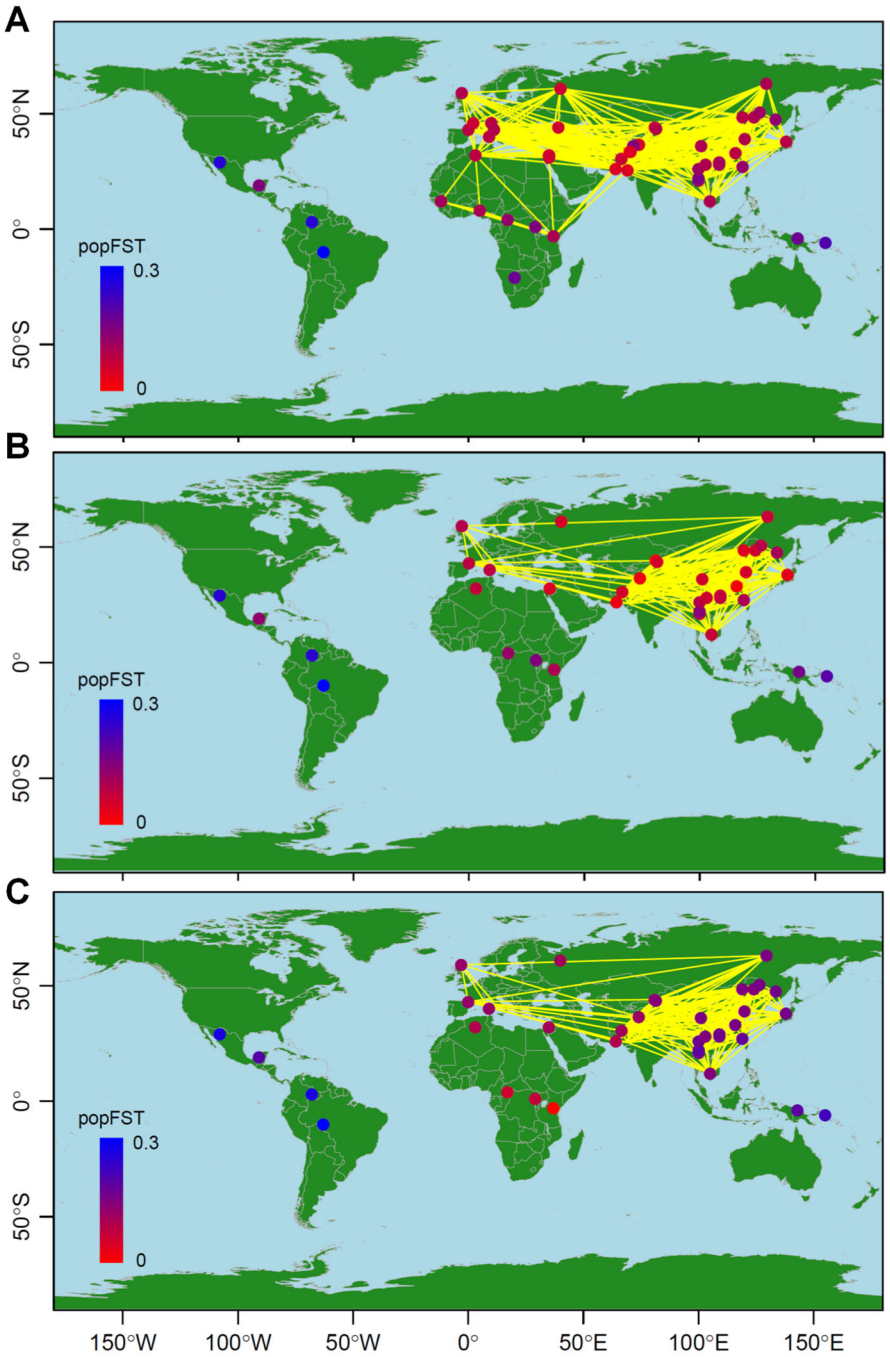

Figure 3 Population structure of humans based on Bayesian and moment estimators of population-specific F_{ST}. Results from the Bayesian population-specific F_{ST} estimator using (A) 51 samples and (B) 37 subsamples, and from (C) the WG population-specific F_{ST} moment estimator using 37 subsamples. 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. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. The color of each population indicates the magnitude of population-specific F_{ST} values between red (for the smallest F_{ST}) and blue (for the largest F_{ST}).

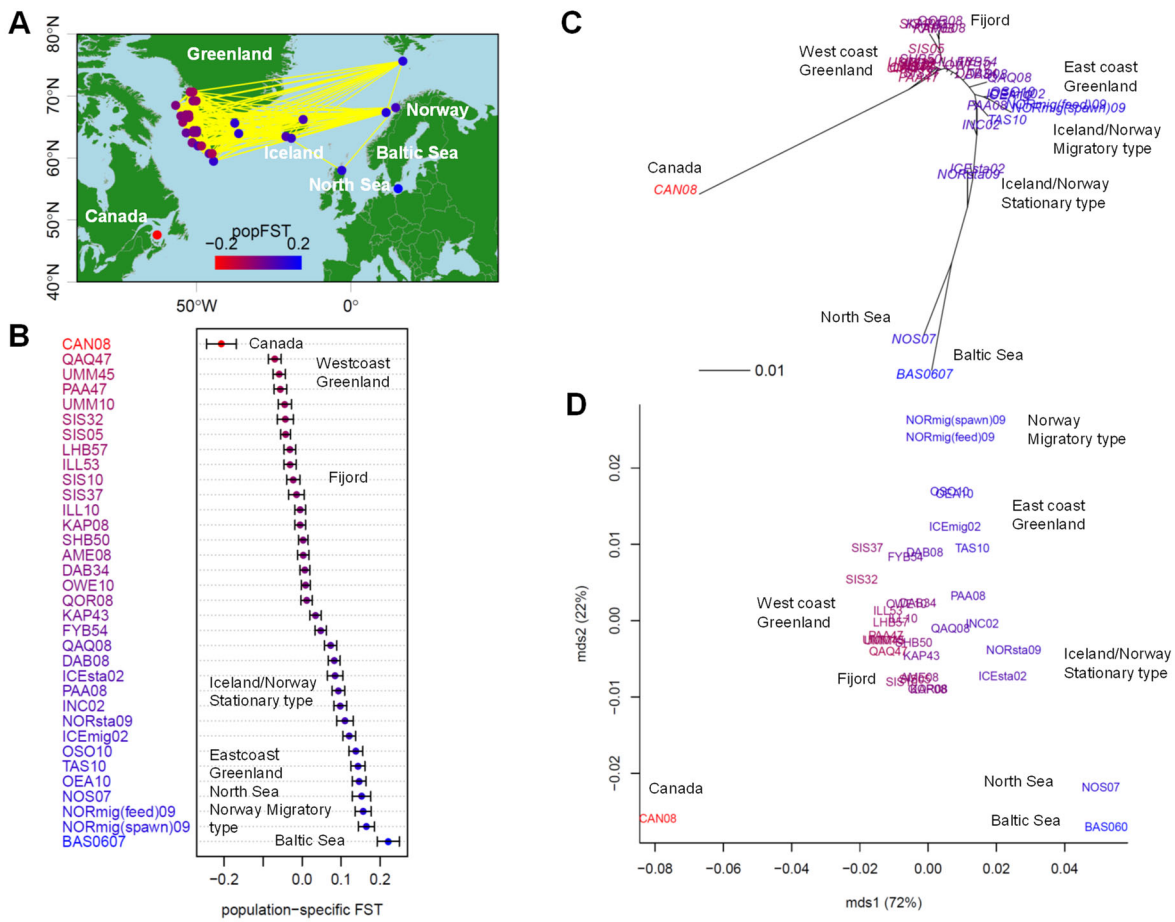

Figure 4 Population structure of 34 geographical samples of wild Atlantic cod (*n* = 1,065; 921 SNPs). (A) Map showing population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. (B) Distribution of population-specific F_{ST} values ± 2×SE. (C) Neighbor-joining (NJ) unrooted tree and (D) multi-dimensional scaling (MDS) 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 between red (for the smallest F_{ST}) and blue (for the largest F_{ST}).

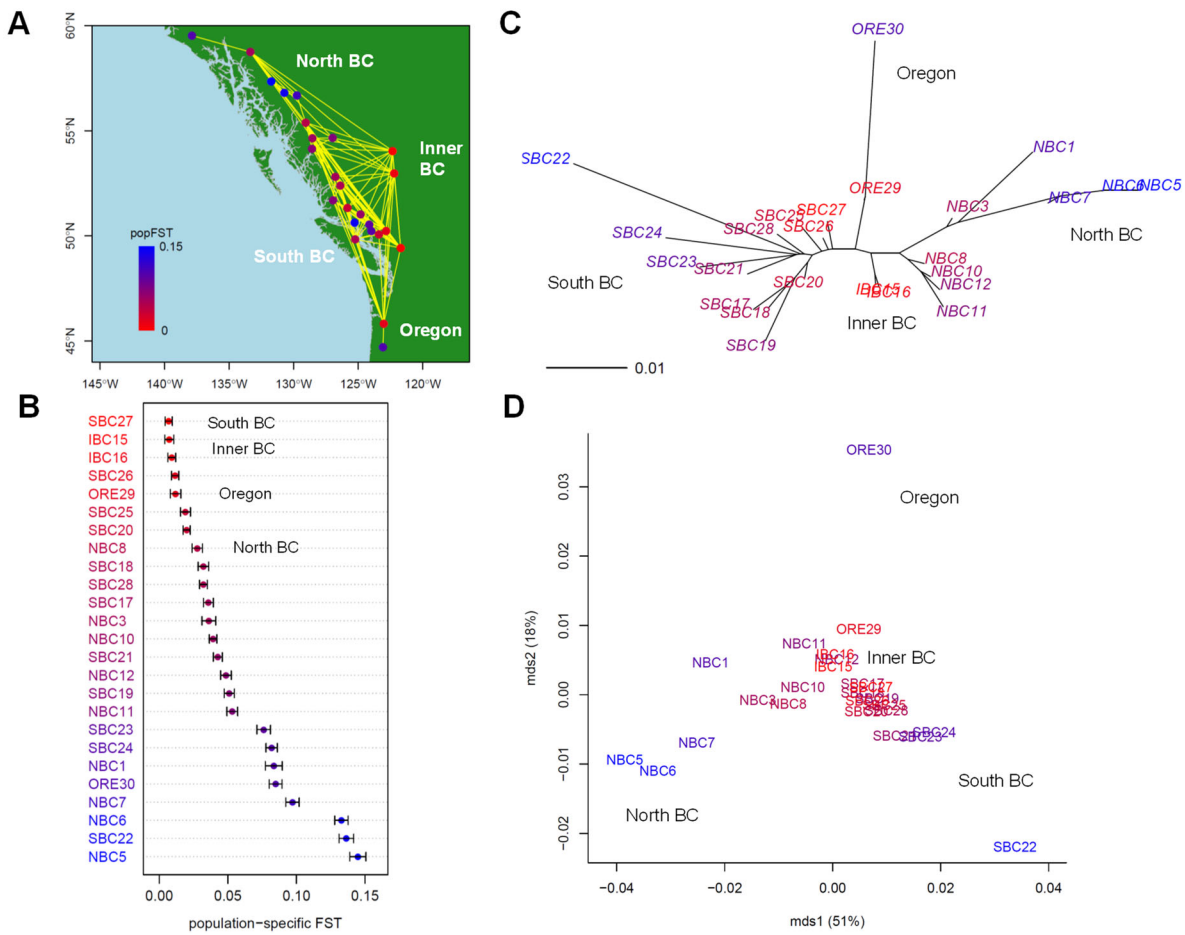

Figure 5 Population structure for 25 geographical samples of wild poplar (n = 441; 29,355 SNPs). (A) Map showing population connectivity with the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.02$. (B) Distribution of population-specific F_{ST} values ± 2×SE. (C) Neighbor-joining (NJ) unrooted tree and (D) multi-dimensional scaling (MDS) based on pairwise F_{ST} overlaid with population-specific F_{ST} values on population labels. The color of each population indicates the magnitude of population-specific F_{ST} values between red (for the smallest F_{ST}) and blue (for the largest F_{ST}).

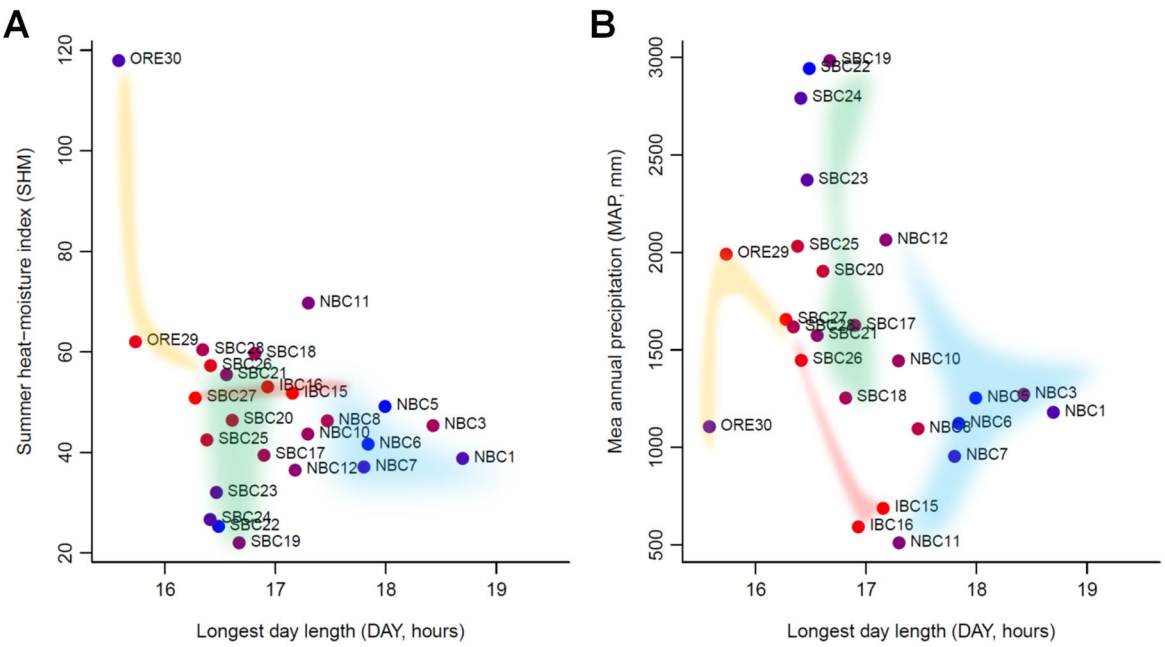

Figure 6 Population range expansion and key environmental variables. Longest day length vs. (A) summer heat-moisture index and (B) mean annual precipitation for 25 geographical samples of wild poplar. The colored areas by the population clusters (see Figure S8) show the inferred population expansion from IBC15, IBC16, and SBC27. The color of each population shows the magnitude of population-specific F_{ST} values between red (for the smallest F_{ST}) and blue (for the largest F_{ST}).