Understanding population structure in an evolutionary context:

population-specific $F_{ST}$ and pairwise $F_{ST}$

Short running title: Integrated $F_{ST}$ population structure

Shuichi Kitada¹, Reiichiro Nakamichi², and Hirohisa Kishino³

1 Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan
2 Japan Fisheries Research and Education Agency, Yokohama 236-8648, Japan
3 Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan

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. Wright’s $F_{ST}$ measures current 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 novel 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.

Keywords:

Adaptation, evolution, genetic diversity, migration, population structure
1 | INTRODUCTION

Quantifying genetic relationships among populations is of substantial interest in population biology, ecology, and human genetics (Weir & 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 & Hill, 2002; Weir & 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 & Lugon-Moulin, 2002; Weir & Hill, 2002; Beaumont, 2005; Holsinger & Weir, 2009; Gaggiotti & 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. Although population-specific $F_{ST}$ is expected to have a wide range of applications (Weir and Goudet, 2017), there have been no formal comparative studies that describe their differences between traditional and population-specific $F_{ST}$ estimators, and how they scale-up to genomic data is not known. Software has not been provided for the population-specific $F_{ST}$ moment estimator, which makes it difficult biologists to this approach.

In this study, we propose a novel representation that overlays genome-wide population-specific and pairwise $F_{ST}$ estimates (average over loci) to understand population structure in an evolutionary context. We visualized current population structure on a clustering tree and on a map of sampling locations based on the two different $F_{ST}$ estimates. The environment experienced by the population under range expansion was estimated by the generalized least
squares (GLS) of genome-wide population-specific $F_{ST}$, which takes into account residual
correlation due to population structure. We demonstrated the usefulness of our procedure by
conducting simulations that mimicked population colonization from a single ancestral
population, and we applied this procedure to published genotype data sets of humans, Atlantic
cod, and wild poplar.

We chose 377 microsatellite genotypes collected from human populations worldwide as the
first empirical data, because their evolutionary history, migration, and population structure has
been best studied (e.g., Diamond, 1997; Rosenberg et al., 2002; Ramachandran et al., 2005;
Liu et al., 2006; Rutherford, 2016; Nielsen et al., 2017). Because the results are well known
by statistical/theoretical population geneticists and biologists, our new integrative $F_{ST}$ analysis
on this data set could provide a good example of the usefulness of our new approach. A
single-nucleotide polymorphism (SNP) data set was obtained from a commercially important
fish, Atlantic cod ($Gadus morhua$), in the North Atlantic. The genotype data of 924 SNPs
were combined from two data sets, which 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 in a warming climate. The other SNP
data set was from a tree, wild poplar ($Populus trichocarpa$), in the American Pacific
Northwest. The samples were collected under different environmental conditions over an area
of 2,500 km near the Canadian–US border along with various environmental data and are thus
possibly useful for detecting environmental effects on population structure. The poplar data
contained 29,355 SNPs, and the corresponding CPU processing time will provide a practical
measure for scaling-up to genomic data. 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 of population structure and simulations of population colonization used in this study are available in the Supporting Information. This can be used for microsatellite and SNP genotype data, and accepts Genepop format (Raymond & Rousset, 1995; Rousset, 2008), which has been particularly widely used among biologists.

2 | MATERIALS AND METHODS

2.1 | Understanding population structure in evolutionary context

We integrated genome-wide population-specific and pairwise $F_{ST}$ estimates (averaged over all loci) on an unrooted neighbor-joining (NJ) tree (Saitou & Nei, 1987) and on a map of sampling locations. We drew the NJ tree based on the distance matrix of genome-wide pairwise $F_{ST}$ values (averaged over all loci) 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 $\text{rgb}(1 - F_{ST,0}, 0, F_{ST,0})$, where $F_{ST,0} = (F_{ST} - \text{min}F_{ST})/(\text{max}F_{ST} - \text{min}F_{ST})$. 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 heterozygosities ($H_e$). Sampling points with pairwise $F_{ST}$ values smaller than a given threshold were connected by lines to visualize the image of gene flow between populations. The R codes for the representation method of population structure for the human data set are given in the Supporting Information.
2.2 | Inferring environmental selection from 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. 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 overestimates the precision. To take the correlation into account, we used GLS with the GLS function in FinePop2_ver.0.2. We derived the components of the variance–covariance matrix $\Omega$ for the GLS function as follows. The population-specific $F_{ST}$ estimator can be written as $psF_{ST} = \tilde{y}^i$. Using the Taylor series expansion for the first term, we inferred the asymptotic variance as

$$V[psF_{ST}^i] \approx \frac{\tilde{y}_i^2}{\bar{x}^2} \left( \frac{V[\bar{x}]}{\bar{x}^2} + \frac{V[\tilde{y}_i]}{\tilde{y}_i^2} - \frac{2Cov[\bar{x}, \tilde{y}_i]}{\bar{x}\tilde{y}_i} \right)$$  \hspace{1cm} (1)

Similarly, the asymptotic covariances between population-specific $F_{ST}$ values of $i, j$ populations were obtained by

$$Cov[psF_{ST}^i, psF_{ST}^j] \approx \frac{\tilde{y}_i\tilde{y}_j}{\bar{x}^4} V[\bar{x}] - \frac{\tilde{y}_i}{\bar{x}^3} Cov[\bar{x}, \tilde{y}_i] - \frac{\tilde{y}_j}{\bar{x}^3} Cov[\bar{x}, \tilde{y}_j]$$  \hspace{1cm} (2)

where the variance and covariance components were calculated by

$$V[\bar{x}] = \frac{1}{L(L-1)} \sum_{l=1}^{L} (x_l - \bar{x})^2, \quad V[\tilde{y}_i] = \frac{1}{L(L-1)} \sum_{l=1}^{L} (y^i_l - \bar{y}_i)^2,$$

$$Cov[\bar{x}, \tilde{y}_i] = \frac{1}{L(L-1)} \sum_{l=1}^{L} (x_l - \bar{x})(y^i_l - \bar{y}_i)$$ and

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

This analysis was performed on the wild poplar data set, for which 11 environmental/geographical parameters were available for each sampling location. 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 & Gaggiotti, 2008; Coop et al., 2010).

2.3 | Applied $F_{ST}$ estimators

Throughout this paper, notations consistent with those of Weir & Hill (2002) were used: $i$ for populations ($i = 1, \ldots, r$), $u$ for alleles ($u = 1, \ldots, m$), and $l$ for loci ($l = 1, \ldots, L$). We used WG population-specific $F_{ST}$ moment estimators, because we expected that population-specific $F_{ST}$ values reflect population history (Weir & 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:

$$psF^i_{ST} = \frac{\sum_{l=1}^{L}(M^i_{W,l} - M^P_l)}{\sum_{l=1}^{L}(1 - M^P_l)},$$  \hspace{1cm} (3)

where $M^i_{W,l}$ is the unbiased within-population matching of two distinct alleles of population $i$, and $M^P_l$ 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 & Cockerham, 1984; Weir & Hill, 2002; Bhatia et al., 2013).

We applied empirical (Beaumont & Balding, 2004) and full Bayesian (Foll & Gaggiotti, 2006) population-specific $F_{ST}$ estimators. Beaumont & Balding (2004) maximized the Dirichlet-multinomial marginal likelihood in their Equation 1 and estimated $\theta_{li}$:

$$L_{li}(\theta_{li}|n_{i1}, \ldots, n_{im}) = \frac{\Gamma(\theta_{li})}{\Gamma(N_{li} + \theta_{li})} \prod_{u=1}^{m} \frac{\Gamma(n_{iu} + \theta_{li} \hat{p}_{lu})}{\Gamma(\theta_{li} \hat{p}_{lu})}. \hspace{1cm} (4)$$

Here, $\theta_{li}$ 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, \( \tilde{p}_{lu} \) is the mean allele frequency over all subpopulations, whereas \( \theta_{li} \tilde{p}_{lu} = \alpha_{liu} \), 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 \( \tilde{\theta}_{li} \) (Beaumont & Balding, 2004):

\[
psF_{ST,l} = \frac{1}{\tilde{\theta}_{li} + 1}. \tag{5}
\]

We used Nei & Chesser’s (1983) bias-corrected \( G_{ST} \) moment estimator (NC83) to estimate pairwise \( F_{ST} \) over loci in our analysis:

\[
pwF_{ST} = \frac{\sum_{i=1}^{L}(\bar{H}_{T,l} - \bar{H}_{S,l})}{\sum_{i=1}^{L}\bar{H}_{T,l}}, \tag{6}
\]

where \( \bar{H}_T \) and \( \bar{H}_S \) 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 & Weir, 2009), whereas NC83 does not consider any population replicates (Weir & Cockerham, 1984; Excoffier, 2001). The pairwise \( F_{ST} \) values obtained from NC83 therefore measured current population structures based on a fixed set of samples of subpopulations. 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).

### 2.4 | Computing \( F_{ST} \) values
We converted the genotype data into Genepop format (Raymond & Rousset, 1995; Rousset, 2008) for implementation in the R package FinePop2_ver.0.2. We applied the bias-corrected population-specific $F_{ST}$ moment estimator (Weir & Goudet, 2017) (WG). 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 & Cockerham, 1984; Weir & Hill, 2002) used for the $F_{ST}$ functions in FinePop2_ver.0.2, 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 & 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 & Gaggiotti, 2006) was used to compute genome-wide population-specific $F_{ST}$ values. $F_{ST}$ is equal to $G_{ST}$ for diploid random mating populations (Excoffier, 2001; also see Supplemental Note); therefore, pairwise $F_{ST}$ values based on Nei & Chesser’s bias corrected $G_{ST}$ (1983) (NC83, Equation 6) were computed using the `pop_pairwiseFST` function in FinePop2_ver.0.2. Expected heterozygosity was calculated for each population with the `read.GENEPOP` function.

### 2.5 Simulations of population colonization

To test the performance of our representation method, we conducted simulations that mimicked colonization of populations from a single ancestral population (population 1). We modeled three types of colonization, one- (Figure S1a), two-, and three-directional population expansion (Figure 1a, d), with 24 demes (populations 2–25). We set the effective population size of the ancestral population of $N_e = 10^5$ (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. The effective population size of the newly derived population increased to $N_e = 10^4$ after one generation, and the populations 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 generations and lost their polymorphism. After 260 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 belongs. 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 $F_{ST}$ and pairwise $F_{ST}$ values between 25 populations and overlaid genome-wide population-specific $F_{ST}$ values on the unrooted NJ tree, as described in 2.1. The R codes for the simulations are available in the Supporting Information.

2.6 | Empirical data sets

The human microsatellite data in Rosenberg et al. (2002) were retrieved from [https://web.stanford.edu/group/rosenberglab/index.html](https://web.stanford.edu/group/rosenberglab/index.html). We converted the data to Genepop format (Raymond & 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).

The Atlantic cod SNP genotype data of 924 markers common to 29 populations reported in Therkildsen et al. (2013a, b) and 12 populations in Hemmer-Hansen et al. (2013a, b) were combined. 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. 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.

Wild poplar 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 (Raymond & Rousset, 1995; Rousset, 2008). The samples covered various regions over a range of 2,500 km near the Canadian–US border at altitudes between 0 and 800 m (Supplemental Data). A breakdown of the 25 drainages (hereafter, subpopulations) 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/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) (Supplemental Data). The AHM was calculated in the original study as (MAT+10)/(MAP/1000); a large AHM indicates extremely dry conditions.

3 | RESULTS

3.1 | 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 S1b). 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 1b). 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 1e).

Our simulation results demonstrated that our method represents new 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. In contrast, 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 (Figures S1c, 1c), whereas slightly negative
population-specific \( F_{ST} \) values were obtained in the ancestral and adjacent populations in the three-directional model (Figures 1f). 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 compared with real data (Figures S1b, 1b, e). 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. Equation 3 calculates deviation of within-population heterozygosity \( (\hat{H}_{Si} = 1 - \hat{M}_{Wi}^i) \) from between-population heterozygosity based on all different population pairs \( (\hat{H}_B = 1 - \hat{M}_{Bl}^B) \). Thus, Equation 3 produces negative values of population-specific \( F_{ST} \) in cases of \( \hat{H}_{Si} > \hat{H}_B \) (see Discussion 4.3).

3.2 | Humans

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) Americas (Figure S2). The WG population-specific \( F_{ST} \) estimator (Supplemental Data) indicated that populations in Africa had the smallest \( F_{ST} \) values, followed by the Middle East, Central/South Asia, Europe, and East Asia. The NJ tree integrated with population-specific \( F_{ST} \) values inferred that human populations originated from Bantu Kenyans (having the smallest \( F_{ST} \) value as shown in red) and expanded to Europe, Middle East, Central/South Asia, and East Asia (Figure 2a,b). The Kalash were isolated from Europe/Middle East and Central/South Asia populations. Middle/South American populations and Papuans/Melanesians diverged from Central/South Asian and East Asian populations. As indicated by sampling points with \( F_{ST} \) values below the 0.02 threshold, gene flow from Africa...
was low. In contrast, gene flow was substantial within Eurasia but was much smaller than that inferred from Eurasia to Oceania and America (Figure 3). As illustrated by sampling point radii, $H_e$ was high in Africa, the Middle East, Central/South Asia, Europe, and East Asia, but relatively small in Oceania and America. The Kalash were less heterozygous than other populations in Central/South Asia; there are approximately 4,000 individuals that live in isolation in the highlands of northwestern Pakistan (Rutherford, 2016), and they speak an Indo-European language (Rosenberg et al., 2002). The Karitiana in Brazil had the lowest heterozygosity. $H_e$ was highest in Africa and lowest in South America.

Bayesian population-specific $F_{ST}$ values estimated using the methods of Beaumont & Balding (2004) and Foll & Gaggiotti (2006) were nearly identical; however, in African populations, they were higher than WG population-specific $F_{ST}$ values (Figure S3). 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 (Supplemental Data). 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 S4).

### 3.3 Atlantic cod

The populations were divided according to the ordinal NJ tree of the pairwise $F_{ST}$ distance matrix 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 (Figure S5). The lowest WG population-specific $F_{ST}$ value was in Canada (Supplemental Data). Greenland west-coast populations (in green in Figure
S5) generally had small $F_{ST}$ values. Fjord populations had relatively higher $F_{ST}$ values. $F_{ST}$ values were much higher for populations in Iceland, Norway, and the North Sea. The $F_{ST}$ value was the highest for BAS0607 from the Baltic Sea. Our integrated NJ tree with population-specific $F_{ST}$ values estimated that Atlantic cod originated from Canada (having the smallest $F_{ST}$ value as shown in red), migrated to the west coast of Greenland, and then expanded their distribution to Iceland, Norway, the North Sea, and the Baltic Sea (Figures 4a, S6). They might migrate to find new habitat (carrying capacity), and individuals with genomic variation were able to adapt to changing environments and formed the current local populations. The evolutionary history of Atlantic cod populations was clearly visualized on a map (Figure 4b). $H_e$ (indicated by circle radii) was very high in Canada and Greenland, low in other areas, and lowest in the Baltic Sea. Based on pairwise $F_{ST}$ values between sampling points (< 0.02 threshold), substantial gene flow was detected between Greenland, Iceland, and Norway. In contrast, gene flow was low from Canada and the North and Baltic seas.

### 3.4 | Wild poplar

The ordinal NJ tree based on 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. Population-specific $F_{ST}$ values were lowest in SBC27, IBC15, and IBC16 (Supplemental Data). Samples collected from areas close to the SBC coast had higher population-specific $F_{ST}$ values than other SBC samples. 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. 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, which formed current local populations. Our integrated NJ tree with population-specific $F_{ST}$ values showed that wild poplar originated from Inner BC and expanded in three directions with environmental adaptation, namely, to the BC southern coast, northern BC and south-western Alaska, and Oregon (Figures 5a, S8). $H_e$ was highest in SBC27, IBC15, and IBC16, and lowest in NBC5. WG population-specific $F_{ST}$-based visualization of wild-poplar evolutionary history (Figure 5b) revealed that SBC27, IBC15, and IBC16 had the smallest $F_{ST}$ values (in red) and large $H_e$, whereas NBC5, NBC6, and SBC22 had the largest $F_{ST}$ values (in blue) and lowest $H_e$. As inferred by pairwise $F_{ST}$ values between sampling points connected by yellow lines ($< 0.02$ threshold), substantial gene flow was observed among populations.

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) 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 (Figure 6a). This was consistent in the scatter plot of DAY and MAP, which demonstrated that the expansion in SBC could have
been facilitated by adaptation to higher MAP (Figure 6b).

3.5 | 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.

4 | DISCUSSION

4.1 | Population-specific $F_{ST}$ traced population history as reflected by genetic diversity

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 (Figures 2, 3). 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 identified a source population and traced the evolutionary history. Our two-directional simulation corresponded to the human data and supported the results of the analysis.

The Atlantic cod data also corresponded to our two-directional simulation. The evolutionary history of Atlantic cod was again clearly visualized (Figure 4). Our analysis indicated that Atlantic cod originated in Canada (CAN08). The population-specific $F_{ST}$ value of CAN08 was very small, $-0.21 \pm 0.02$ (SE), which was caused by the highest $H_e$ value (population-specific heterozygosity) of CAN08, which was much greater than $1 - \tilde{M}_B$ (between population heterozygosity) in Equation 3 (Figure S9). This 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. 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 (Figures 4b, S9); 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 heterozygosity values, 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).

The wild poplar data corresponded to our three-directional simulation. Although the samples used in this study might not cover the whole distribution range of wild poplar, which extends from southern California to northern Alaska, genome-wide population-specific $F_{ST}$ values suggested that wild poplar trees in southern British Colombia (SBC27) and inland British Colombia (IBC15, 16) are the closest to the ancestral population. 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). The wild popular expanded in three directions as they adapted to local environments: coastal British Colombia (SBC; abundant rain), southern Oregon (ORE30; mostly dry summers), and northern British Colombia (NBC; long periods of daylight) (Figure 6). Changes in environmental factors could be inferred at the end points of population expansion. To relate SNPs to environmental changes, functional roles of mutation that underpinned environmental adaptation should be examined (many may be associated with functional loss).

Our results from the simulations and three case studies demonstrated that WG population-specific $F_{ST}$ values identified the source population and traced the evolutionary history of its derived population’s history based on genetic diversity.

4.2 | 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). This analysis was conducted because divergent selection in an environmental gradient can impact genome-wide population structure (Nosil et al., 2009; Orsini et al., 2013), and prior studies examined geographic distance and habitat differences between populations as variables that impact population structure (Bradbury & Bentzen, 2007; Jorde et al., 2015; Kitada et al., 2017). The results suggested that wild popular originated from IBC and expanded its distribution to NBC by adapting to longer day lengths, to SBC seashores adapting to the rainy environment, and to ORE adapting to dry summer conditions (Figure 6). 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 & 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.

4.3 | 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 (Cochran, 1977; Weir & Cockerham, 1984; Weir & 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 S4). 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 $i$ ($\hat{H}_{Si}$) as derived in Nei & 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

$$\frac{1}{L} \sum_{l=1}^{L} \left( 1 - \sum_{u=1}^{m} \hat{p}_{lu}^2 \right) \rightarrow \frac{1}{L} \sum_{l=1}^{i} \left( 1 - E \left[ \sum_{u=1}^{m} \hat{p}_{lu}^2 \right] \right).$$

The observed heterozygosity thus converges to the expected value:

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

Similarly, $\hat{H}_S$ and $\hat{H}_T$ 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 & Chesser, 1983; Weir & Cockerham, 1984) and population-specific $F_{ST}$ (Weir & 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 & Cockerham, 1984; Holsinger & 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 $F_{ST}$ (Kitada et al., 2017).

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 $1 - \tilde{M}_W^i$. When Hardy–Weinberg equilibrium is assumed ($\tilde{H}_{0i} = \tilde{H}_{Si}$), the preceding formula is equivalent to the NC83 unbiased estimator of the gene diversity of population $i$ ($\tilde{H}_{Si}$) (Supplemental Note):

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

Another variable, $\tilde{M}_B^i$, is “average over pairs of populations of between-population-pair matching” (Weir & Goudet, 2017). $\tilde{M}_B^i$ is the homozygosity over pairs of populations, and we can write observed heterozygosity over pairs of populations as $1 - \tilde{M}_B^i = \tilde{H}_B$. $\tilde{H}_B$ 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

$$ps\hat{F}_{ST}^i = \frac{\bar{M}_w^i - \bar{M}^B}{1 - \bar{M}^B} = \frac{\bar{H}_B - \bar{H}_{Si}}{\bar{H}_B} = 1 - \frac{\bar{H}_{Si}}{\bar{H}_B}.$$ (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, rather than components of variance” (Weir & Goudet, 2017). Weir & Goudet (2017) also gave the relation between $E[\text{G}_{ST}]$ and their notation $\theta^B$ and $\theta^W$ in their Equation 2. In our three case studies, a linear relationship between $H_e$ of each population ($= H_{Si}$) and $ps\hat{F}_{ST}^i$ was evident (Figure S9), which was exemplified in Equation 7. The coefficient of determination, $R^2$, was 0.91 for 51 human populations ($n = 1,035$), 0.993 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 $-0.21 \pm 0.019$ compared with the maximum value of $0.22 \pm 0.014$ in BAS0607 (Figure S9, Supplemental Data). The Atlantic cod data corresponded to the two-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 1). 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 $H_e$ ($\bar{H}_{Si}$) is maintained in the source population. In such cases with $\bar{H}_{Si} > \bar{H}_B$, Equation 7 produces negative values of population-specific $F_{ST}$.

4.4 | Shrinkage in Bayesian $F_{ST}$ estimators
We drew the integrated NJ tree with the empirical Bayesian population-specific $F_{ST}$ values (Beaumont & Balding, 2004), which showed that the Hazara, Pakistan population was genetically closest to human ancestors (Figure 7a). Our $F_{ST}$ map indicated that the Middle East, Europe, and Central/South Asia were centers of human origin (Figure 7b), which was consistent with that from the full Bayesian population-specific $F_{ST}$ estimator (Foll & Gaggiotti, 2006) and population-specific $F_{ST}$ estimators (figure not shown). The results obtained with Bayesian estimators were a consequence of Equation 4, which uses the mean allele frequency over subpopulations ($\bar{p}_{lu}$) to reduce the number of parameters to be estimated. The locations of the 51 human populations were as follows: 21 from the Middle East, Europe, and Central/South Asia, 18 from East Asia, 6 from Africa, 2 from Oceania, and 4 from America. The mean allele frequency ($\bar{p}_{lu}$) reflected the weight of samples from the Middle East, Europe, and Central/South Asia, thereby resulting in these areas being identified as centers of origin. Instead of $\bar{p}_{lu}$, the full Bayesian method uses allele frequencies in the ancestral population, $p_{lu}$, which are generated from a noninformative Dirichlet prior, $p_{lu} \sim Dir (1, ..., 1)$. 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, Bayesian population-specific $F_{ST}$ values were higher for African populations than WG population-specific $F_{ST}$ values and close to those for East Asia (Figures S2). 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 & Balding, 2004). Our empirical Bayes pairwise $F_{ST}$ estimator (EB$F_{ST}$; 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, this approach 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 $EBF_{ST}$ and other Bayesian $F_{ST}$ estimators, particularly in genome analyses where large numbers of SNPs are used.

5 | CONCLUSIONS

WG population-specific $F_{ST}$ moment estimator identifies the source population and traces the evolutionary history of its derived population’s history based on genetic diversity. In contrast, NC83 pairwise $F_{ST}$ moment estimator represents the current population structure. By integrating estimates from both estimators on NJ trees and maps of sampling locations, we obtained a picture of current 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 $F_{ST}$. All $F_{ST}$ moment estimators described in this paper have reasonable CPU times as implemented in FinePop2 and can also be used in population genomics studies. Our new practical procedure is expected to have a wide range of applications, because there are R scripts that can implement our representation method and simulations of population colonization.
ACKNOWLEDGEMENTS

We appreciate the essential comments on the early version by the reviewers, which 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.

AUTHOR CONTRIBUTIONS

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

DATA ACCESSIBILITY STATEMENT

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. The R codes to perform our representation method and simulations of population colonization are available in the Supporting Information.

ORCID

Shuichi Kitada: http://orcid.org/0000-0001-5838-0374
Nakamichi Reiichiro: http://orcid.org/0000-0001-5789-7689
Hirohisa Kishino: https://orcid.org/0000-0002-3244-359X

REFERENCES

Balloux, F., & Lugon-Moulin, N. (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, 11, 155–165. https://doi.org/10.1046/j.0962-1083.2001.01436.x
Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13, 969–980.
Beaumont, M. A. (2005). Adaptation and speciation: what can FST tell us? Trends in Ecology and Evolution, 20, 435–440. https://doi.org/10.1016/j.tree.2005.05.017

Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., ... & Andre, C. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic cod (Gadus morhua L.). Genome Biology and Evolution, 7, 1644–1663. https://doi.org/10.1093/gbe/evv093

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M., Knutsen, H., ... & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. Scientific Reports, 6, 23246. https://doi.org/10.1038/srep23246

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., ... & Jakobsen, K. S. (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., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. Genome Research, 23, 1514–1521. http://www.genome.org/cgi/doi/10.1101/gr.154831.113

Bradbury, I. R., & Bentzen, P. (2007). Non-linear genetic isolation by distance: implications for dispersal estimation in anadromous and marine fish populations. Marine Ecology Progress Series, 340, 245–257. doi:10.3354/meps340245

Buckleton, J., Curran, J., Goudet, J., Taylor, D., Thiery, A., & Weir, B. S. (2016). Population-specific Fst values for forensic STR markers: A worldwide survey. Forensic Science International: Genetics, 23, 91–100. https://doi.org/10.1016/j.fsigen.2016.03.004

Cann, H. M., De Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., ... & Chen, Z. (2002). A human genome diversity cell line panel. Science, 296, 261–262. DOI: 10.1126/science.296.5566.261b

Cockran, W. G. (1977). Sampling Techniques. New York, USA: Wiley.

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (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. London, UK: Random House.

Excoffier, L (2001). Analysis of population subdivision, In D. J. Balding, M. Bishop and C. Cannings (Eds.) Handbook of Statistical Genetics (pp. 271–307). Chichester, UK: Wiley.

Foll, M., & Gaggiotti, O. E. (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., & Gaggiotti, O. E. (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., & Foll, M. (2010). Quantifying population structure using the F-model. Molecular Ecology Resources, 10, 821–830. https://doi.org/10.1111/j.1755-0998.2010.02873.x

Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., ... & Jones, S. J. (2011). SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. Molecular Ecology Resources, 11, 81–92. https://doi.org/10.1111/j.1755-
A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Molecular Ecology Resources*, 13, 306–323. https://doi.org/10.1111/1755-0998.12056

Hemmer - Hansen, J., Nielsen, E. E., Therkildsen, N. O., Taylor, M. I., Ogden, R., Geffen, A. J., ... & FishPopTrace Consortium. (2013). A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology*, 22, 2653–2667. https://doi.org/10.1111/mec.12284

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., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132, 583–589.

Jorde, P. E., Søvik, G., Westgaard, J. I., Albretsen, J., André, C., Hvingel, C., ... & Jørstad, K. E. (2015). Genetically distinct populations of northern shrimp, *Pandalus borealis*, in the North Atlantic: adaptation to different temperatures as an isolation factor. *Molecular Ecology*, 24, 1742–1757. https://doi.org/10.1111/mec.13158

Liu, H., Prugnolle, F., Manica, A., & Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. *The American Journal of Human Genetics*, 79, 230–237. https://doi.org/10.1086/505436

McKown, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., Cronk, Q. C., ... & Douglas, C. J. (2014a). Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New Phytologist*, 201, 1263–1276. https://doi.org/10.1111/nph.12601

McKown, A. D., Klápště, J., Guy, R. D., Geraldes, A., Porth, I., Hannemann, J., ... & Cronk, Q. C. (2014b). Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist*, 203, 535–553. https://doi.org/10.1111/nph.12815

McKown, A. D., Guy, R. D., Quamme, L., Klápště, J., La Mantia, J., Constabel, C. P., ... & Azam, M. S. (2014c). Association genetics, geography and ecophysiology link stomatal patterning in *Populus trichocarpa* with carbon gain and disease resistance trade-offs.
Molecular Ecology, 23, 5771–5790. https://doi.org/10.1111/mec.12969

Nei, M. (1973). Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences, 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. Annals of Human Genetics, 41, 225–233. https://doi.org/10.1111/j.1469-1809.1977.tb01918.x

Nei, M., & Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. Annals of Human Genetics, 47, 253–259. https://doi.org/10.1111/j.1469-1809.1983.tb00993.x

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

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

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J., & De Meester, L. (2013). Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. Molecular Ecology, 22, 5983–5999. https://doi.org/10.1111/mec.12561

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

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

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, 2002 Genetic structure of human populations. Science 298: 2381–2385. DOI: 10.1126/science.1078311

Rousset, F. (2001). Inferences from spatial population genetics, In D. J. Balding, M. Bishop and C. Cannings (Eds.) Handbook of Statistical Genetics (pp. 239–269). Chichester, UK: Wiley.

Rousset, F., (2004). Genetic Structure and Selection in Subdivided Populations. Princeton, USA: Princeton University Press.

Rousset, F. (2008). Genepop’007: a complete reimplementation of the Genepop software for Windows and Linux. Molecular Ecology Resources, 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. New York, USA: The Experiment.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

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

Therkildsen, N. O., Hemmer - Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C., Meldrup, D., ... & Nielsen, E. E. (2013). Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod Gadus morhua. Evolutionary Applications, 6, 690–705. https://doi.org/10.1111/eva.12055
Therkildsen, N. O., Hemmer-Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C., Meldrup, D., ... & Nielsen, E. E. (2013). 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

Weir, B. S., & Cockerham, C. C. (1984). Estimating $F$-statistics for the analysis of population structure. Evolution, 38, 1358–1370.

Weir, B. S., & Hill, W. G. (2002). Estimating $F$-statistics. Annual Review of Genetics, 36, 721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940

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

Wright, S. (1931). Evolution in Mendelian populations. Genetics, 16, 97–159.

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

**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the Supporting Information section at the end of the article.
Table 1 Regression of genome-wide population-specific $F_{ST}$ of 25 wild poplar populations on the 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 simulations of population colonization. (a) Two- and (d) three-directional colonization models. Population 1 in red is ancestral, and arrows show the direction of colonization. Neighbor-joining unrooted trees based on pairwise $F_{ST}$ distance matrix overlaid with population-specific $F_{ST}$ values for the (b) two- and (e) three-directional models. 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 $(He \times 10)^3$. Population-specific $F_{ST}$ values for 25 simulated populations are presented for the (c) two- and (f) three-directional models.
FIGURE 2 Population structure of 51 human populations \((n = 1,035; 377\) microsatellites). The unrooted NJ tree based on pairwise \(F_{ST}\) overlaid with population-specific \(F_{ST}\) values on (a) population labels and (b) population nodes. The arrows show inferred routes of population range expansion. The color of each population shows the magnitude of population-specific \(F_{ST}\) values. Data from Rosenberg et al. (2002).
FIGURE 3 Map of the population structure of 51 human populations. The color of each population shows the magnitude of population-specific $F_{ST}$ values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.01$. The radius of each sampling point is proportional to the level of $He$ as visualized by $H_e^{12}$. The arrows show inferred routes of population expansion. Data from Rosenberg et al. (2002).
FIGURE 4 Population structure of 34 geographical samples of wild Atlantic cod ($n = 1,065; 921$ SNPs). (a) Unrooted NJ tree based on pairwise $F_{ST}$ and population-specific $F_{ST}$ values. (b) Map of the population structure of the Atlantic cod populations. The color of each population shows the magnitude of population-specific $F_{ST}$ values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.04$. The radius of each sampling point is proportional to the level of heterozygosity ($H_e$) as visualized by $H_e^{100}$. The arrows show inferred routes of population expansion.

Data are combined from Therkildsen et al. (2013) and Hemmer-Hansen et al. (2013).
FIGURE 5 Population structure for 25 geographical samples of wild poplar (n = 441; 29,355 SNPs). (a) Unrooted NJ tree based on pairwise $F_{ST}$ overlaid with population-specific $F_{ST}$ values. (b) Map of the population structure of the wild poplar populations. The color of each population shows 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 heterozygosity ($H_e$) as visualized by $H_e^{100}$. The arrows show inferred routes of population expansion. Data from McKown et al. (2014b).
FIGURE 6 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 S6).
FIGURE 7 Population structure of 51 human populations inferred based on Bayesian population-specific $F_{ST}$ values. (a) Unrooted NJ tree based on pairwise $F_{ST}$ overlaid with population-specific $F_{ST}$ values. (b) Map of the population structure. The color of each population shows the magnitude of Bayesian population-specific $F_{ST}$ values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.01$. The radius of each sampling point is proportional to the level of $He$ as visualized by $H_1^2$. Data from Rosenberg et al. (2002).