1 # Understanding population structure in an evolutionary context:

2 ## population-specific F_{ST} and pairwise F_{ST}

3

4 Short running title: **Integrated F_{ST} population structure**

5

6

7 Shuichi Kitada[1], Reiichiro Nakamichi[2], and Hirohisa Kishino[3]
8

9

10 1 Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan

11 2 Japan Fisheries Research and Education Agency, Yokohama 236-8648, Japan

12 3 Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo

13 113-8657, Japan

14

15

16 Corresponding:

17 Shuichi Kitada,

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

19 Japan

20 +81-297-45-5267

21 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.

2

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

3

70   squares (GLS) of genome-wide population-specific F_{ST}, which takes into account residual

71   correlation due to population structure. We demonstrated the usefulness of our procedure by

72   conducting simulations that mimicked population colonization from a single ancestral

73   population, and we applied this procedure to published genotype data sets of humans, Atlantic

74   cod, and wild poplar.

75

76   We chose 377 microsatellite genotypes collected from human populations worldwide as the

77   first empirical data, because their evolutionary history, migration, and population structure has

78   been best studied (*e.g.*, Diamond, 1997; Rosenberg et al., 2002; Ramachandran et al., 2005;

79   Liu et al., 2006; Rutherford, 2016; Nielsen et al., 2017). Because the results are well known

80   by statistical/theoretical population geneticists and biologists, our new integrative F_{ST} analysis

81   on this data set could provide a good example of the usefulness of our new approach. A

82   single-nucleotide polymorphism (SNP) data set was obtained from a commercially important

83   fish, Atlantic cod (*Gadus morhua*), in the North Atlantic. The genotype data of 924 SNPs

84   were combined from two data sets, which included historical samples collected 50–80 years

85   ago and contemporary samples from the northern range margin of the species in Greenland,

86   Norway, and the Baltic Sea. The inclusion of both types of data might facilitate detection of

87   migration history of this highly migratory marine fish in a warming climate. The other SNP

88   data set was from a tree, wild poplar (*Populus trichocarpa*), in the American Pacific

89   Northwest. The samples were collected under different environmental conditions over an area

90   of 2,500 km near the Canadian–US border along with various environmental data and are thus

91   possibly useful for detecting environmental effects on population structure. The poplar data

92   contained 29,355 SNPs, and the corresponding CPU processing time will provide a practical

93   measure for scaling-up to genomic data. All F_{ST} estimators were computed using the R

4

94 package FinePop2_ver.0.2, which is available at CRAN. The R codes for our representation

95 method of population structure and simulations of population colonization used in this study

96 are available in the Supporting Information. This can be used for microsatellite and SNP

97 genotype data, and accepts Genepop format (Raymond & Rousset, 1995; Rousset, 2008),

98 which has been particularly widely used among biologists.

99

2 | MATERIALS AND METHODS

2.1 | Understanding population structure in evolutionary context

102 We integrated genome-wide population-specific and pairwise F_{ST} estimates (averaged over all

103 loci) on an unrooted neighbor-joining (NJ) tree (Saitou & Nei, 1987) and on a map of

104 sampling locations. We drew the NJ tree based on the distance matrix of genome-wide

105 pairwise F_{ST} values (averaged over all loci) using the `nj` function in the R package ape and

106 superimposed the magnitude of genome-wide population-specific F_{ST} values using a color

107 gradient on the NJ tree based on rgb $(1 - F_{ST,0}, 0, F_{ST,0})$, where $F_{ST,0} = (F_{ST} -$

108 $\min F_{ST})/(\max F_{ST} - \min F_{ST})$. This conversion represents the standardized magnitude of a

109 population-specific F_{ST} value at the sampling point, with colors between blue (for the largest

110 F_{ST}) and red (smallest F_{ST}). The F_{ST} maps were drawn using the `sf` package in R, where

111 sampling locations were plotted based on the longitudes and latitudes; they were visualized by

112 the population-specific F_{ST} color gradient, and the size of each sampling point was

113 proportional to the expected heterozygosity (H_e). Sampling points with pairwise F_{ST} values

114 smaller than a given threshold were connected by lines to visualize the image of gene flow

115 between populations. The R codes for the representation method of population structure for

116 the human data set are given in the Supporting Information.

117

5

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 Ω for the GLS function as follows. The population-specific F_{ST} estimator can be written as $ps\hat{F}_{ST}^i = \frac{\bar{y}^i}{\bar{x}}$. Using the Taylor series expansion for the first term, we inferred the asymptotic variance as

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

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

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

where the variance and covariance components were calculated by

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

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

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

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

140   determining outliers of locus-specific, population-specific F_{ST} values (Foll & Gaggiotti, 2008;

141   Coop et al., 2010).

142

2.3 | Applied F_{ST} estimators

144   Throughout this paper, notations consistent with those of Weir & Hill (2002) were used: i for

145   populations ($i = 1, \ldots, r$), u for alleles ($u = 1, \ldots, m$), and l for loci ($l = 1, \ldots, L$). We used

146   WG population-specific F_{ST} moment estimators, because we expected that population-specific

147   F_{ST} values reflect population history (Weir & Goudet, 2017). In our analyses, we extended

148   WG population-specific F_{ST} estimator to overall loci (genome-wide population-specific F_{ST}),

149   and the combined ratio estimator (Cochran, 1977) for overall loci (Buckleton et al. 2016) was:

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

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

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

153   (Supplemental Note). This is called the "ratio of averages" F_{ST} estimator, which

154   asymptotically converges to unbiased estimates of F_{ST} as the number of independent SNPs

155   increases (Weir & Cockerham, 1984; Weir & Hill, 2002; Bhatia et al., 2013).

156

157   We applied empirical (Beaumont & Balding, 2004) and full Bayesian (Foll & Gaggiotti,

158   2006) population-specific F_{ST} estimators. Beaumont & Balding (2004) maximized the

159   Dirichlet-multinomial marginal likelihood in their Equation 1 and estimated θ_{li}:

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

161   Here, θ_{li} is the scale parameter of the Dirichlet prior distribution for locus l and population i,

7

162    $\tilde{p}_{lu}$ is the observed frequency of allele u at locus l, n_{liu} is the observed allele count in

163   population i, and N_{li} is the total number of alleles. Importantly, $\bar{p}_{lu}$ is the mean allele

164   frequency over all subpopulations, whereas $\theta_{li}\bar{p}_{lu} = \alpha_{liu}$, where $\theta_{li} = \sum_{u=1}^{m_l} \alpha_{liu}$. The

165   parametrization reduces the number of parameters to be estimated. Based on a Dirichlet

166   (multi-allelic) and/or a beta (bi-allelic) scale parameter, population-specific F_{ST} values were

167   estimated for each locus using the following function of $\hat{\theta}_{li}$ (Beaumont & Balding, 2004):

168

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

169

170   We used Nei & Chesser's (1983) bias-corrected G_{ST} moment estimator (NC83) to estimate

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

172

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

173   where $\hat{H}_T$ and $\hat{H}_S$ are the unbiased estimators of total and within-population heterozygosity,

174   respectively (Supplemental Note). G_{ST} (Nei, 1973) is defined "by using the gene frequencies

175   at the present population, so that no assumption is required about the pedigrees of individuals,

176   selection, and migration in the past" (Nei, 1977). G_{ST} assumes no evolutionary history

177   (Holsinger & Weir, 2009), whereas NC83 does not consider any population replicates (Weir &

178   Cockerham, 1984; Excoffier, 2001). The pairwise F_{ST} values obtained from NC83 therefore

179   measured current population structures based on a fixed set of samples of subpopulations. Our

180   previous coalescent simulations demonstrated that NC83 performs the best among F_{ST}

181   estimators when estimating pairwise F_{ST} values, particularly for larger numbers of loci

182   (Kitada et al., 2017).

183

184   **2.4 | Computing F_{ST} values**

185   We converted the genotype data into Genepop format (Raymond & Rousset, 1995; Rousset,

186   2008) for implementation in the R package FinePop2_ver.0.2. We applied the bias-corrected

187   population-specific F_{ST} moment estimator (Weir & Goudet, 2017) (WG). Genome-wide WG

188   population-specific F_{ST} (Equation 3) values were computed using the `pop_specificFST`

189   function. In addition to the "ratio of averages" (Weir & Cockerham, 1984; Weir & Hill, 2002)

190   used for the F_{ST} functions in FinePop2_ver.0.2, we computed the "average of ratios" (Bhatia

191   et al., 2013) of the WG population-specific F_{ST} for human data for comparison by averaging

192   locus-specific, population-specific F_{ST} values over loci. We maximized Equation 4 and

193   estimated the empirical Bayesian population-specific F_{ST} (Beaumont & Balding, 2004) at

194   each locus according to Equation 5. We then averaged these values over all loci. For the full

195   Bayesian model, GESTE_ver. 2.0 (Foll & Gaggiotti, 2006) was used to compute genome-

196   wide population-specific F_{ST} values. F_{ST} is equal to G_{ST} for diploid random mating

197   populations (Excoffier, 2001; also see Supplemental Note); therefore, pairwise F_{ST} values

198   based on Nei & Chesser's bias corrected G_{ST} (1983) (NC83, Equation 6) were computed

199   using the `pop_pairwiseFST` function in FinePop2_ver.0.2. Expected heterozygosity was

200   calculated for each population with the `read.GENEPOP` function.

201

202   **2.5 | Simulations of population colonization**

203   To test the performance of our representation method, we conducted simulations that

204   mimicked colonization of populations from a single ancestral population (population 1). We

205   modeled three types of colonization, one- (Figure S1a), two-, and three-directional population

206   expansion (Figure 1a, d), with 24 demes (populations 2–25). We set the effective population

207   size of the ancestral population of $N_e = 10^5$ (twice the number of individuals in diploid

208   organisms in a random mating population). At the beginning of colonization, 1% of N_e

9

209   migrated into the adjacent vacant habitat once every 10 generations. The effective population

210   size of the newly derived population increased to $N_e = 10^4$ after one generation, and the

211   populations exchanged 1% of N_e genes with adjacent population(s) in every generation. Like

212   the ancestral population, 1% of N_e individuals migrated into the adjacent vacant habitat once

213   every 10 generations. We simulated the allele frequencies of SNPs in the ancestral and 24

214   derived populations.

215

216   The initial allele frequencies in the ancestral population, q, at 100,000 neutral SNP loci were

217   generated from the predictive equilibrium distribution, $f(q) \propto q^{-1}(1-q)^{-1}$ (Wright 1931).

218   Additionally, 10 newly derived SNPs were introduced to each existing population in each

219   generation. Therefore, in total, 35,100 SNPs were generated. When a new SNP emerged in a

220   population, we set the initial allele frequency of the newly derived SNP to 0.01 in the

221   population and 0 in the other populations. This mimicked new mutations that survived at the

222   initial phase after their birth. These 100,000 ancestral SNPs and 35,100 newly derived SNPs

223   were considered "unobserved." The allele frequencies of these SNPs were changed by random

224   drift under a binomial distribution in every generation. Many of the SNPs had reduced

225   frequencies of the derived allele over generations and lost their polymorphism. After 260

226   generations, SNPs that retained their polymorphism were randomly selected as "observed"

227   SNPs. In this simulation, we selected 10,000 ancestral SNPs and 500 newly derived SNPs.

228   Then, we generated 50 individuals for each population. Genotypes of these 10,500 SNPs were

229   randomly generated for each individual following the allele frequencies in the population to

230   which each individual belongs. Thus, we obtained "observed" genotypes of 1,250 individuals

231   (= 50 individuals × 25 populations) at 10,500 SNP loci (10,000 ancestral SNPs + 500 newly

232   derived SNPs by mutation). We converted the simulated genotypes into Genepop format

233   (Raymond & Rousset, 1995; Rousset, 2008). We computed genome-wide population-specific

234   F_{ST} and pairwise F_{ST} values between 25 populations and overlaid genome-wide population-

235   specific F_{ST} values on the unrooted NJ tree, as described in 2.1. The R codes for the

236   simulations are available in the Supporting Information.

237

238   **2.6 | Empirical data sets**

239   The human microsatellite data in Rosenberg et al. (2002) were retrieved from

240   https://web.stanford.edu/group/rosenberglab/index.html. We converted the data to Genepop

241   format (Raymond & Rousset, 1995; Rousset, 2008). We removed the Surui sample (Brazil)

242   from the data because that population was reduced to 34 individuals in 1961 as a result of

243   introduced diseases (Liu et al., 2006). We retained genotype data ($n = 1,035$) of 377

244   microsatellite loci from 51 populations categorized into six groups as in the original study: 6

245   populations from Africa, 12 from the Middle East and Europe, 9 from Central/South Asia, 18

246   from East Asia, 2 from Oceania, and 4 from America. Longitudes and latitudes of the

247   sampling sites were obtained from Cann et al. (2002).

248

249   The Atlantic cod SNP genotype data of 924 markers common to 29 populations reported in

250   Therkildsen et al. (2013a, b) and 12 populations in Hemmer‐Hansen et al. (2013a, b) were

251   combined. We compared genotypes associated with each marker in samples that were

252   identical between the two studies, namely, CAN08 and Western_Atlantic_2008, ISO02 and

253   Iceland_migratory_2002, and ISC02 and Iceland_stationary_2002, and standardized the gene

254   codes. We removed cgpGmo.S1035, whose genotypes were inconsistent between the two

255   studies. We also removed cgpGmo.S1408 and cgpGmo.S893, for which genotypes were

256   missing in several population samples in Therkildsen et al. (2013b). Temporal replicates in

257   Norway migratory, Norway stationary, North Sea, and Baltic Sea samples were removed for

258   simplicity. The final data set consisted of genotype data (n = 1,065) at 921 SNPs from 34

259   populations: 3 from Iceland, 25 from Greenland, 3 from Norway, and 1 each from Canada, the

260   North Sea, and the Baltic Sea. Two ecotypes (migratory and stationary) that were able to

261   interbreed but were genetically differentiated (Hemmer‐Hansen et al., 2013a; Berg et al.,

262   2016) were included in the Norway and Iceland samples. All individuals in the samples were

263   adults, and most were mature (Therkildsen et al., 2013a). The longitudes and latitudes of the

264   sampling sites in Hemmer‐Hansen et al. (2013a) were used. For the data from Therkildsen et

265   al. (2013a), approximate sampling points were estimated from the map of the original study,

266   and longitudes and latitudes were recorded.

267

268   Wild poplar SNP genotype data and environmental/geographical data were retrieved from the

269   original studies of McKown et al. (2014a, b). The genotype data contained 29,355 SNPs of

270   3,518 genes of wild poplar (n = 441) collected from 25 drainage areas (McKown et al.,

271   2014c). Details of array development and selection of SNPs are provided in Geraldes et al.

272   (2011, 2013). We converted the data to Genepop format (Raymond & Rousset, 1995; Rousset,

273   2008). The samples covered various regions over a range of 2,500 km near the Canadian–US

274   border at altitudes between 0 and 800 m (Supplemental Data). A breakdown of the 25

275   drainages (hereafter, subpopulations) is as follows: 9 in northern British Colombia (NBC), 2

276   in inland British Colombia (IBC), 12 in southern British Colombia (SBC), and 2 in Oregon

277   (ORE) (Geraldes et al., 2014). The original names of clusters and population numbers were

278   combined and used for our population labels (NBC1, NBC3,…, ORE30). Each sampling

279   location was associated with 11 environmental/geographical parameters: latitude (lat),

280   longitude (lon), altitude (alt), longest day length (DAY), frost-free days (FFD), mean annual

12

281 temperature (MAT), mean warmest month temperature (MWMT), mean annual precipitation

282 (MAP), mean summer precipitation (MSP), annual heat–moisture index (AHM), and summer

283 heat-moisture index (SHM) (Supplemental Data). The AHM was calculated in the original

284 study as (MAT+10)/(MAP/1000); a large AHM indicates extremely dry conditions.

285

3 | RESULTS

3.1 | Simulations of population colonization

288 In the one-directional simulation, our method correctly identified the ancestral population

289 with the highest genetic diversity, and populations were located in order from 1 to 25 on the

290 NJ tree (Figure S1b). In the two-directional simulation, our method correctly identified the

291 ancestral population and detected that populations were split at population 9 and expanded in

292 two directions, which was consistent with the simulation scenario (Figure 1b). In the three-

293 directional simulation, the ancestral population was closely located to the adjacent

294 populations 2, 9, and 17, but correctly detected three directions, as in the other simulations

295 (Figure 1e).

296

297 Our simulation results demonstrated that our method represents new insight into current

298 population structure that incorporates evolutionary history. Our results revealed that WG

299 population-specific F_{ST} values (standardized by a color gradient) identified the source

300 population and traced the evolutionary history of its derived populations based on genetic

301 diversity. In contrast, the NC83 pairwise F_{ST} estimator correctly estimated the current

302 population structure. Genome-wide WG population-specific F_{ST} values were negative in the

303 ancestral population and adjacent populations, and the phenomenon was particularly

304 significant in the one- and two-directional models (Figures S1c, 1c), whereas slightly negative

13

305 population-specific F_{ST} values were obtained in the ancestral and adjacent populations in the

306 three-directional model (Figures 1f). In contrast, H_e values were larger in the ancestral

307 population in the one- and two-directional models than in the three-directional model, though

308 the variation in H_e values was not substantial because of the relatively few generations (260)

309 in the simulation compared with real data (Figures S1b, 1b, e). Our simulation results

310 indicated that, when gene flow from other populations into the source population was limited,

311 relatively large H_e could be maintained, which resulted in substantial negative population-

312 specific F_{ST} values. Equation 3 calculates deviation of within-population heterozygosity

313 $(\widehat{H}_{Si} = 1 - \widetilde{M}_W^i)$ from between-population heterozygosity based on all different population

314 pairs $(\widehat{H}_B = 1 - \widetilde{M}_l^B)$. Thus, Equation 3 produces negative values of population-specific F_{ST}

315 in cases of $\widehat{H}_{Si} > \widehat{H}_B$ (see Discussion 4.3).

316

317 **3.2 | Humans**

318 The ordinal NJ tree of pairwise F_{ST} values divided the populations into five clusters: 1) Africa,

319 2) the Middle East, Europe, and Central/South Asia, 3) East Asia, 4) Oceania, and 5)

320 Americas (Figure S2). The WG population-specific F_{ST} estimator (Supplemental Data)

321 indicated that populations in Africa had the smallest F_{ST} values, followed by the Middle East,

322 Central/South Asia, Europe, and East Asia. The NJ tree integrated with population-specific

323 F_{ST} values inferred that human populations originated from Bantu Kenyans (having the

324 smallest F_{ST} value as shown in red) and expanded to Europe, Middle East, Central/South Asia,

325 and East Asia (Figure 2a,b). The Kalash were isolated from Europe/Middle East and

326 Central/South Asia populations. Middle/South American populations and

327 Papuans/Melanesians diverged from Central/South Asian and East Asian populations. As

328 indicated by sampling points with F_{ST} values below the 0.02 threshold, gene flow from Africa

14

329   was low. In contrast, gene flow was substantial within Eurasia but was much smaller than that

330   inferred from Eurasia to Oceania and America (Figure 3). As illustrated by sampling point

331   radii, H_e was high in Africa, the Middle East, Central/South Asia, Europe, and East Asia, but

332   relatively small in Oceania and America. The Kalash were less heterozygous than other

333   populations in Central/South Asia; there are approximately 4,000 individuals that live in

334   isolation in the highlands of northwestern Pakistan (Rutherford, 2016), and they speak an

335   Indo-European language (Rosenberg et al., 2002). The Karitiana in Brazil had the lowest

336   heterozygosity. H_e was highest in Africa and lowest in South America.

337

338   Bayesian population-specific F_{ST} values estimated using the methods of Beaumont & Balding

339   (2004) and Foll & Gaggiotti (2006) were nearly identical; however, in African populations,

340   they were higher than WG population-specific F_{ST} values (Figure S3). The distributions of

341   F_{ST} values obtained from the two Bayesian methods were very similar, with the smallest F_{ST}

342   values observed in the Middle East, Europe, and Central/South Asia (Supplemental Data). The

343   "ratio of averages" and "average of ratios" of the WG population-specific F_{ST} estimator were

344   almost identical in all populations for this data set (Figure S4).

345

346   **3.3 | Atlantic cod**

347   The populations were divided according to the ordinal NJ tree of the pairwise F_{ST} distance

348   matrix into four large clusters: 1) Canada, 2) Greenland west coast, 3) Greenland east coast,

349   Iceland, Norway, and 4) North and Baltic seas. Fjord populations (in purple) formed a sub-

350   cluster within the Greenland west coast, and migratory (orange) and stationary (magenta)

351   ecotypes also formed a sub-cluster (Figure S5). The lowest WG population-specific F_{ST} value

352   was in Canada (Supplemental Data). Greenland west-coast populations (in green in Figure

15

353   S5) generally had small F_{ST} values. Fjord populations had relatively higher F_{ST} values. F_{ST}

354   values were much higher for populations in Iceland, Norway, and the North Sea. The F_{ST}

355   value was the highest for BAS0607 from the Baltic Sea. Our integrated NJ tree with

356   population-specific F_{ST} values estimated that Atlantic cod originated from Canada (having the

357   smallest F_{ST} value as shown in red), migrated to the west coast of Greenland, and then

358   expanded their distribution to Iceland, Norway, the North Sea, and the Baltic Sea (Figures 4a,

359   S6). They might migrate to find new habitat (carrying capacity), and individuals with genomic

360   variation were able to adapt to changing environments and formed the current local

361   populations. The evolutionary history of Atlantic cod populations was clearly visualized on a

362   map (Figure 4b). H_e (indicated by circle radii) was very high in Canada and Greenland, low

363   in other areas, and lowest in the Baltic Sea. Based on pairwise F_{ST} values between sampling

364   points (< 0.02 threshold), substantial gene flow was detected between Greenland, Iceland, and

365   Norway. In contrast, gene flow was low from Canada and the North and Baltic seas.

366

367   **3.4 | Wild poplar**

368   The ordinal NJ tree based on pairwise F_{ST} distant matrix divided populations into three large

369   clusters: 1) IBC, 2) SBC, 3) NBC, and 4) ORE (Figure S7). The population represented by

370   sample ORE30 was isolated from ORE29. Population-specific F_{ST} values were lowest in

371   SBC27, IBC15, and IBC16 (Supplemental Data). Samples collected from areas close to the

372   SBC coast had higher population-specific F_{ST} values than other SBC samples. NBC samples

373   had population-specific F_{ST} values similar to those of SBC. Among NBC samples, NBC8 had

374   the smallest population-specific F_{ST}, and NBC5 had the highest value, followed by NBC6 and

375   NBC7. Wild poplar could have expanded their distribution by their fluffy seeds being blown

376   away by wind, and individuals that had genomic variation were able to adapt to local

16

377   environments, which formed current local populations. Our integrated NJ tree with

378   population-specific F_{ST} values showed that wild poplar originated from Inner BC and

379   expanded in three directions with environmental adaptation, namely, to the BC southern coast,

380   northern BC and south-western Alaska, and Oregon (Figures 5a, S8). H_e was highest in

381   SBC27, IBC15, and IBC16, and lowest in NBC5. WG population-specific F_{ST}-based

382   visualization of wild-poplar evolutionary history (Figure 5b) revealed that SBC27, IBC15,

383   and IBC16 had the smallest F_{ST} values (in red) and large H_e, whereas NBC5, NBC6, and

384   SBC22 had the largest F_{ST} values (in blue) and lowest H_e. As inferred by pairwise F_{ST} values

385   between sampling points connected by yellow lines (< 0.02 threshold), substantial gene flow

386   was observed among populations.

387

388   To avoid multicollinearity, we excluded seven out of 11 environmental variables that were

389   significantly correlated with each other, namely, lat, lon, alt, FFD, MWMT, MSP, and AHM.

390   Our GLS of genome-wide population-specific F_{ST} values on the four environmental variables

391   (DAY, MAT, MAP, and SHM) indicated that DAY, MAP, and SHM were significant (Table 1).

392   All estimates were positive, which indicated that higher population-specific F_{ST} values were

393   expected for longer DAY (longer daylight time), higher MAP (abundant rain), and higher

394   SHM (dry summers), and these values reflected directions of population expansion. The

395   scatter plot of DAY and SHM (each population colored by the population-specific F_{ST} value)

396   suggested three directions of population range expansion; the wild popular that originated

397   from IBC15 expanded its distribution to NBC, where it adapted to longer DAY; that which

398   originated from SBC27 expanded to SBC seashores, where it adapted to lower SHM, and to

399   ORE29 and ORE30, where it adapted to higher SHM (Figure 6a). This was consistent in the

400   scatter plot of DAY and MAP, which demonstrated that the expansion in SBC could have

17

401 been facilitated by adaptation to higher MAP (Figure 6b).

402

3.5 | CPU times

404 Using a laptop computer with an Intel Core i7-8650U CPU, only 89.8 s of CPU time was

405 required to compute WG population-specific F_{ST} estimates and SEs of wild poplar (29,355

406 SNPs; 25 populations, $n = 441$). Alternatively, 120.7 s was required to obtain pairwise F_{ST}

407 (NC83) between all population pairs. Based on the results, we may need 50 min to compute

408 WG population-specific F_{ST} and 70 min to compute pairwise NC83 F_{ST} estimates for 1

409 million SNPs using this laptop. This computation could be much faster if we used a

410 workstation.

411

4 | DISCUSSION

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

415 In our analysis, genome-wide WG population-specific F_{ST} values successfully illustrated

416 human evolutionary history, and indicated that humans originated in Kenya, expanded from

417 the Middle East into Europe and from Central/South Asia into East Asia, and then possibly

418 migrated to Oceania and America (Figures 2, 3). Kenya is located just below Ethiopia, where

419 the earliest anatomically modern humans were found from fossils (Nielsen et al., 2017). Our

420 results are also in good agreement with the highest levels of genetic diversity being detected

421 in Africa (Rosenberg et al., 2002), the relationship uncovered between genetic and geographic

422 distance (Ramachandran et al., 2005), the shortest colonization route from East Africa (Liu et

423 al., 2006), and major migrations inferred from genomic data (Nielsen et al., 2017). The

424 genome-wide WG population-specific F_{ST} values are consistent with results obtained from 24

18

425   forensic STR markers (Buckleton et al., 2016). Our analysis identified a source population

426   and traced the evolutionary history. Our two-directional simulation corresponded to the

427   human data and supported the results of the analysis.

428

429   The Atlantic cod data also corresponded to our two-directional simulation. The evolutionary

430   history of Atlantic cod was again clearly visualized (Figure 4). Our analysis indicated that

431   Atlantic cod originated in Canada (CAN08). The population-specific F_{ST} value of CAN08

432   was very small, -0.21 ± 0.02 (SE), which was caused by the highest H_e value (population-

433   specific heterozygosity) of CAN08, which was much greater than $1 - \tilde{M}_l^B$ (between

434   population heterozygosity) in Equation 3 (Figure S9). This suggested that the population

435   expansion of Atlantic cod began by minimal gene flow from Canada. They might have first

436   expanded to the west coast of Greenland before spreading to Iceland, the North Sea, Norway,

437   and the Baltic Sea. This result was consistent with genomic evidence that Atlantic cod inhabit

438   both sides of the Atlantic Ocean and evolved from a common evolutionary origin (Berg et al.,

439   2017). The migratory ecotypes characterized by deeper and more offshore habitats and long-

440   distance migrations (Hemmer‐Hansen et al., 2013a) may have played an important role in

441   this expansion. In the original Atlantic cod study (Therkildsen et al., 2013a), strong

442   differentiation of CAN08 was found at neutral markers, which prompted the authors to

443   suggest that Greenland populations were the result of colonization from Iceland rather than

444   from refugial populations in southern North America. In our study, CAN08 had the highest

445   H_e, which was lower in Iceland than in Greenland (Figures 4b, S9); this result implies that

446   Icelandic populations were the descendants of colonists from Greenland, which in turn

447   originated in Canada. The BAS0607 sample from the Baltic Sea had the highest population-

448   specific F_{ST} and the lowest heterozygosity values, which suggests that Baltic cod is the

19

449 newest population. This result agrees with the findings of a previous study, which identified

450 Baltic cod as an example of a species subject to ongoing selection for reproductive success in

451 a low salinity environment (Berg et al., 2015).

452

453 The wild poplar data corresponded to our three-directional simulation. Although the samples

454 used in this study might not cover the whole distribution range of wild poplar, which extends

455 from southern California to northern Alaska, genome-wide population-specific F_{ST} values

456 suggested that wild poplar trees in southern British Colombia (SBC27) and inland British

457 Colombia (IBC15, 16) are the closest to the ancestral population. The largest population-

458 specific F_{ST} value was found in the population with the smallest heterozygosity, SBC22,

459 which may have resulted from a bottleneck (Geraldes et al., 2014). The wild popular

460 expanded in three directions as they adapted to local environments: coastal British Colombia

461 (SBC; abundant rain), southern Oregon (ORE30; mostly dry summers), and northern British

462 Colombia (NBC; long periods of daylight) (Figure 6). Changes in environmental factors could

463 be inferred at the end points of population expansion. To relate SNPs to environmental

464 changes, functional roles of mutation that underpinned environmental adaptation should be

465 examined (many may be associated with functional loss).

466

467 Our results from the simulations and three case studies demonstrated that WG population-

468 specific F_{ST} values identified the source population and traced the evolutionary history of its

469 derived population's history based on genetic diversity.

470

471 **4.2 | Genome-wide population-specific F_{ST} detects key environments that**

472 **promote adaptation**

20

473 Our GLS of genome-wide population-specific F_{ST} values revealed that long daylight hours,

474 abundant rainfall, and dry summer conditions are the key environmental factors that

475 influenced the evolution of wild poplar (Table 1). This analysis was conducted because

476 divergent selection in an environmental gradient can impact genome-wide population

477 structure (Nosil et al., 2009; Orsini et al., 2013), and prior studies examined geographic

478 distance and habitat differences between populations as variables that impact population

479 structure (Bradbury & Bentzen, 2007; Jorde et al., 2015; Kitada et al., 2017). The results

480 suggested that wild popular originated from IBC and expanded its distribution to NBC by

481 adapting to longer day lengths, to SBC seashores adapting to the rainy environment, and to

482 ORE adapting to dry summer conditions (Figure 6). A previous study on wild poplar revealed

483 that genes involved in drought response were identified as F_{ST} outliers along with other genes

484 related to transcriptional regulation and nutrient uptake (Geraldes et al., 2014), which is a

485 finding consistent with our GLS results. Our results were also consistent with the F_{ST} outlier

486 test of the original study (Geraldes et al., 2014), in which Bayescan (Foll & Gaggiotti, 2008)

487 revealed that genes involved in circadian rhythm and response to red/far-red light had high

488 locus-specific global F_{ST} values. Moreover, the first principal component of SNP allele

489 frequencies was significantly correlated with day length, and a previous enrichment analysis

490 for population structuring uncovered genes related to circadian rhythm and photoperiod

491 (McKown et al., 2014a). Our results were in agreement with the previous findings, which

492 show the usefulness of using GLS of genome-wide population-specific F_{ST} to infer

493 environmental adaptation and population expansion of species.

494

495 **4.3 | Properties of F_{ST} moment estimators**

496 Previous studies have suggested or indicated that the "ratio of averages" works better than the

21

497 "average of ratios" as the number of independent SNPs increases (Cochran, 1977; Weir &

498 Cockerham, 1984; Weir & Hill, 2002; Bhatia et al., 2013). In regard to the WG population-

499 specific F_{ST} estimator, similar results were obtained for the 377 human microsatellite loci

500 using either the "ratio of averages" or the "average of ratios" (Figure S4). This similar

501 outcome may have been due to the relatively small variation in the locus-specific global F_{ST}

502 values (Figure not shown) and the relatively large number of alleles (12 ± 4) of human

503 microsatellites.

504

505 To explicitly show the underlying mechanism, we used the observed heterozygosity of

506 population i ($\widehat{H}_{Si}$) as derived in Nei & Chesser (1983) (Supplemental Note). When the number

507 of loci (L) increases, the average observed heterozygosity over all loci converges to its

508 expected value according to the law of large numbers as

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

509

510 The observed heterozygosity thus converges to the expected value:

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

511

512 Similarly, $\widehat{H}_S$ and $\widehat{H}_T$ converge to their expected values. This example indicates that the

513 numerators and denominators of bias-corrected F_{ST} moment estimators, whether global,

514 pairwise, or population-specific, converge to their true means and provide unbiased estimates

515 of F_{ST} in population genomics analyses with large numbers of SNPs. Our analyses show that

516 genomic data highlight the usefulness of the bias-corrected moment estimators of traditional

517 F_{ST} developed in the early 1980s (Nei & Chesser, 1983; Weir & Cockerham, 1984) and

518 population-specific F_{ST} (Weir & Goudet, 2017).

519

22

520   To estimate pairwise F_{ST}, our previous coalescent simulations based on ms (Hudson, 2002)

521   showed that NC83 performed best among the present F_{ST} estimators for cases with 10,000

522   SNPs (Kitada et al., 2017). Other F_{ST} moment estimators within an ANOVA framework

523   produce values approximately double those of true values when used to estimate pairwise F_{ST}.

524   NC83 considers a fixed set of population samples; in contrast, the other F_{ST} moment

525   estimators consider replicates of a set of populations (Weir & Cockerham, 1984; Holsinger &

526   Weir, 2009). The models for replicates of population samples were considered to

527   appropriately estimate global F_{ST} and/or mean ancestral coefficient, but cause over-estimation

528   when used to estimate pairwise F_{ST} (Kitada et al., 2017).

529

530   The WG population-specific F_{ST} moment estimator measures population genetic diversity

531   under the framework of relatedness of individuals and identifies the population with the

532   largest genetic diversity as the ancestral population. This estimator thus works to infer

533   evolutionary history through genetic diversity. The WG population-specific F_{ST} estimator is

534   based on allele matching probabilities, where within-population observed heterozygosity can

535   be written as $1 - \widetilde{M}_W^i$. When Hardy–Weinberg equilibrium is assumed ($\widehat{H}_{0i} = \widehat{H}_{Si}$), the

536   preceding formula is equivalent to the NC83 unbiased estimator of the gene diversity of

537   population i $\left(\widehat{H}_{Si}\right)$ (Supplemental Note):

538
$$1 - \widetilde{M}_W^i = \frac{2n_i}{2n_i - 1}\left(1 - \sum_{u=1}^{m} \tilde{p}_{iu}^2\right) = \widehat{H}_{Si} \, .$$

539   Another variable, $\widetilde{M}^B$, is "average over pairs of populations of between-population-pair

540   matching" (Weir & Goudet, 2017). $\widetilde{M}^B$ is the homozygosity over pairs of populations, and

541   we can write observed heterozygosity over pairs of populations as $1 - \widetilde{M}^B = \widehat{H}_B$. $\widehat{H}_B$ is an

542   estimator for the denominator of Hudson et al. (1992). When using only allele frequencies, the

543   population-specific F_{ST} estimator can be written in terms of gene diversity as

544

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

545   This formulation is reasonable, because WG population-specific F_{ST} uses "allele matching,

546   equivalent to homozygosity and complementary to heterozygosity as used by Nei, rather than

547   components of variance" (Weir & Goudet, 2017). Weir & Goudet (2017) also gave the

548   relation between $E[G_{ST}]$ and their notation θ^{B} and θ^{W} in their Equation 2. In our three case

549   studies, a linear relationship between H_{e} of each population ($= H_{Si}$) and $\mathrm{ps}\widehat{F}_{ST}^{i}$ was evident

550   (Figure S9), which was exemplified in Equation 7. The coefficient of determination, R^{2}, was

551   0.91 for 51 human populations ($n = 1{,}035$), 0.993 for 34 Atlantic cod populations ($n = 1{,}065$),

552   and 0.82 for 25 wild poplar populations ($n = 441$). The goodness of fit to the linear function

553   should depend on population sample size (number of individuals).

554

555   In the Atlantic cod case study, CAN08 had the highest H_{e} (Figure 4b) and a very large

556   negative population-specific F_{ST} value of -0.21 ± 0.019 compared with the maximum value

557   of 0.22 ± 0.014 in BAS0607 (Figure S9, Supplemental Data). The Atlantic cod data

558   corresponded to the two-directional model of our simulations, where the WG population-

559   specific F_{ST} value was significantly negative in the ancestral population, whereas H_{e} was the

560   largest (Figure 1). Our consistent results between the simulations and Atlantic cod case study

561   indicate that, when gene flow from other populations into the source population is limited,

562   relatively large H_{e} ($\widehat{H}_{Si}$) is maintained in the source population. In such cases with $\widehat{H}_{Si} >$

563     $\widehat{H}_{B}$, Equation 7 produces negative values of population-specific F_{ST}.

564

565   **4.4 | Shrinkage in Bayesian F_{ST} estimators**

566   We drew the integrated NJ tree with the empirical Bayesian population-specific F_{ST} values

567   (Beaumont & Balding, 2004), which showed that the Hazara, Pakistan population was

568   genetically closest to human ancestors (Figure 7a). Our F_{ST} map indicated that the Middle

569   East, Europe, and Central/South Asia were centers of human origin (Figure 7b), which was

570   consistent with that from the full Bayesian population-specific F_{ST} estimator (Foll &

571   Gaggiotti, 2006) and population-specific F_{ST} estimators (figure not shown). The results

572   obtained with Bayesian estimators were a consequence of Equation 4, which uses the mean

573   allele frequency over subpopulations ($\bar{p}_{lu}$) to reduce the number of parameters to be

574   estimated. The locations of the 51 human populations were as follows: 21 from the Middle

575   East, Europe, and Central/South Asia, 18 from East Asia, 6 from Africa, 2 from Oceania, and

576   4 from America. The mean allele frequency ($\bar{p}_{lu}$) reflected the weight of samples from the

577   Middle East, Europe, and Central/South Asia, thereby resulting in these areas being identified

578   as centers of origin. Instead of $\bar{p}_{lu}$, the full Bayesian method uses allele frequencies in the

579   ancestral population, p_{lu}, which are generated from a noninformative Dirichlet prior,

580   $p_{lu} \sim Dir$ (1, ... ,1). Our results indicate that not enough information is available to estimate

581   allele frequencies in the ancestral population assumed in the models. The shrinkage effect on

582   allele frequencies in Bayesian inference (Stein, 1956) may shift population-specific F_{ST}

583   values toward the average of the whole population. Indeed, Bayesian population-specific F_{ST}

584   values were higher for African populations than WG population-specific F_{ST} values and close

585   to those for East Asia (Figures S2). In contrast, because of shrinkage toward mean allele

586   frequencies, maximum likelihood and Bayesian estimators of locus-specific global F_{ST}

587   improve the power to detect genes under environmental selection (Beaumont & Balding,

588   2004). Our empirical Bayes pairwise F_{ST} estimator (EBF_{ST}; Kitada et al., 2007), which is

589   based on Equation 4, is also useful in cases involving a relatively small number of

25

590 polymorphic marker loci, such as microsatellites; it performs best by averaging large

591 sampling variation of allele frequencies in populations with small sample sizes, particularly in

592 high gene flow scenarios (Kitada et al., 2017). However, this approach suffers from a

593 shrinkage effect similar to that of Bayesian population-specific F_{ST} estimators. We note that

594 the shrinkage effect on allele frequencies can enhance the bias of EBF_{ST} and other Bayesian

595 F_{ST} estimators, particularly in genome analyses where large numbers of SNPs are used.

596

5 | CONCLUSIONS

598 WG population-specific F_{ST} moment estimator identifies the source population and traces the

599 evolutionary history of its derived population's history based on genetic diversity. In contrast,

600 NC83 pairwise F_{ST} moment estimator represents the current population structure. By

601 integrating estimates from both estimators on NJ trees and maps of sampling locations, we

602 obtained a picture of current population structure by incorporating evolutionary history. Our

603 GLS analysis of genome-wide population-specific F_{ST}, which takes the correlation between

604 population-specific F_{ST} values into account, provides insights into how a species has adapted

605 to key environments and expanded its distribution. Given a large number of loci, bias-

606 corrected F_{ST} moment estimators, whether global, pairwise, or population-specific, provide

607 unbiased estimates of F_{ST} supported by the law of large numbers. Genomic data highlight the

608 usefulness of the bias-corrected moment estimators of F_{ST}. All F_{ST} moment estimators

609 described in this paper have reasonable CPU times as implemented in FinePop2 and can also

610 be used in population genomics studies. Our new practical procedure is expected to have a

611 wide range of applications, because there are R scripts that can implement our representation

612 method and simulations of population colonization.

613

ACKNOWLEDGEMENTS

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.

27

641 https://doi.org/10.1111/j.1365-294X.2004.02125.x

642 Beaumont, M. A. (2005). Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology*
643 *and Evolution*, *20*, 435–440. https://doi.org/10.1016/j.tree.2005.05.017

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

648 Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M., Knutsen, H., ... & Jentoft, S.
649 (2016). Three chromosomal rearrangements promote genomic divergence between
650 migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*, 23246.
651 https://doi.org/10.1038/srep23246

652 Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., ... &
653 Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is
654 associated with large inversions. *Heredity*, *119*, 418–428.
655 https://doi.org/10.1038/hdy.2017.54

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

659 Bradbury, I. R., & Bentzen, P. (2007). Non-linear genetic isolation by distance: implications
660 for dispersal estimation in anadromous and marine fish populations. *Marine Ecology*
661 *Progress Series*, *340*, 245–257. doi:3354/meps340245

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

665 Cann, H. M., De Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., ... & Chen, Z.
666 (2002). A human genome diversity cell line panel. *Science*, *296*, 261–262. DOI:
667 10.1126/science.296.5566.261b

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

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

672 Diamond, J. (1997). *Guns, Germs and Steel: The Fates of Human Societies*. London, UK:
673 Random House.

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

677 Foll, M., & Gaggiotti, O. E. (2006). Identifying the environmental factors that determine the
678 genetic structure of populations. *Genetics*, *174*, 875–891.
679 https://doi.org/10.1534/genetics.106.059451

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

683 Gaggiotti, O. E., & Foll, M. (2010). Quantifying population structure using the F-model.
684 *Molecular Ecology Resources*, *10*, 821–830. https://doi.org/10.1111/j.1755-
685 0998.2010.02873.x

686 Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., ... & Jones, S. J. (2011).
687 SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome
688 resequencing. *Molecular Ecology Resources*, *11*, 81–92. https://doi.org/10.1111/j.1755-

689   0998.2010.02960.x

690 Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., ... &
691   Porth, I. (2013). A 34K SNP genotyping array for *Populus trichocarpa*: design,
692   application to the study of natural populations and transferability to other *Populus*
693   species. *Molecular Ecology Resources*, *13*, 306–323. https://doi.org/10.1111/1755-
694   0998.12056

695 Geraldes, A., Farzaneh, N., Grassa, C. J., McKown, A. D., Guy, R. D., Mansfield, S. D., ... &
696   Cronk, Q. C. (2014). Landscape genomics of *Populus trichocarpa*: the role of
697   hybridization, limited gene flow, and natural selection in shaping patterns of population
698   structure. *Evolution*, *68*, 3260–3280. https://doi.org/10.1111/evo.12497

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

703 Hemmer‑Hansen, J., Nielsen, E. E., Therkildsen, N. O., Taylor, M. I., Ogden, R., Geffen, A.
704   J., ... & FishPopTrace Consortium. (2013). A genomic island linked to ecotype
705   divergence in A tlantic cod, Dryad, Dataset, https://doi.org/10.5061/dryad.9gf10

706 Holsinger, K.E., & Weir, B. S. (2009). Genetics in geographically structured populations:
707   defining, estimating and interpreting F_{ST}. *Nature Review Genetics*, *9*, 639–650.
708   https://doi.org/10.1038/nrg2611

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

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

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

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

720 Kitada, S., Nakamichi, R., & Kishino, H. (2017). The empirical Bayes estimators of fine-scale
721   population structure in high gene flow species. *Molecular Ecology Resources*, *17*,
722   1210–1222. https://doi.org/10.1111/1755-0998.12663

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

726 McKown, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., Cronk, Q. C., ... &
727   Douglas, C. J. (2014a). Geographical and environmental gradients shape phenotypic
728   trait variation and genetic structure in *Populus trichocarpa*. *New Phytologist*, *201*,
729   1263–1276. https://doi.org/10.1111/nph.12601

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

734 McKown, A. D., Guy, R. D., Quamme, L., Klápště, J., La Mantia, J., Constabel, C. P., ... &
735   Azam, M. S. (2014c). Association genetics, geography and ecophysiology link stomatal
736   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

30

785   Therkildsen, N. O., Hemmer‐Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C.,
786   Meldrup, D., ... & Nielsen, E. E. (2013). Spatiotemporal SNP analysis reveals
787   pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*.,
788   v2, Dryad, Dataset, https://doi.org/10.5061/dryad.rd250
789   Weir, B. S., & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population
790   structure. *Evolution*, *38*, 1358–1370.
791   Weir, B. S., & Hill, W. G. (2002). Estimating *F*-statistics. *Annual Review of Genetics*, *36*,
792   721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940
793   Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and
794   relatedness. *Genetics*, *206*, 2085–2103. https://doi.org/10.1534/genetics.116.198424
795   Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.
796   Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354.
797   https://doi.org/10.1111/j.1469-1809.1949.tb02451.x
798

799   **SUPPORTING INFORMATION**

800   Additional Supporting Information may be found in the Supporting Information section at the
801   end of the article.

802

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*

803 DAY; longest day length (hours),
804 MAT; mean annual temperature (°C),
805 MAP; mean annual precipitation (mm),
806 SHM; summer heat-moisture index,
807 *p < 0.05 and **p < 0.01
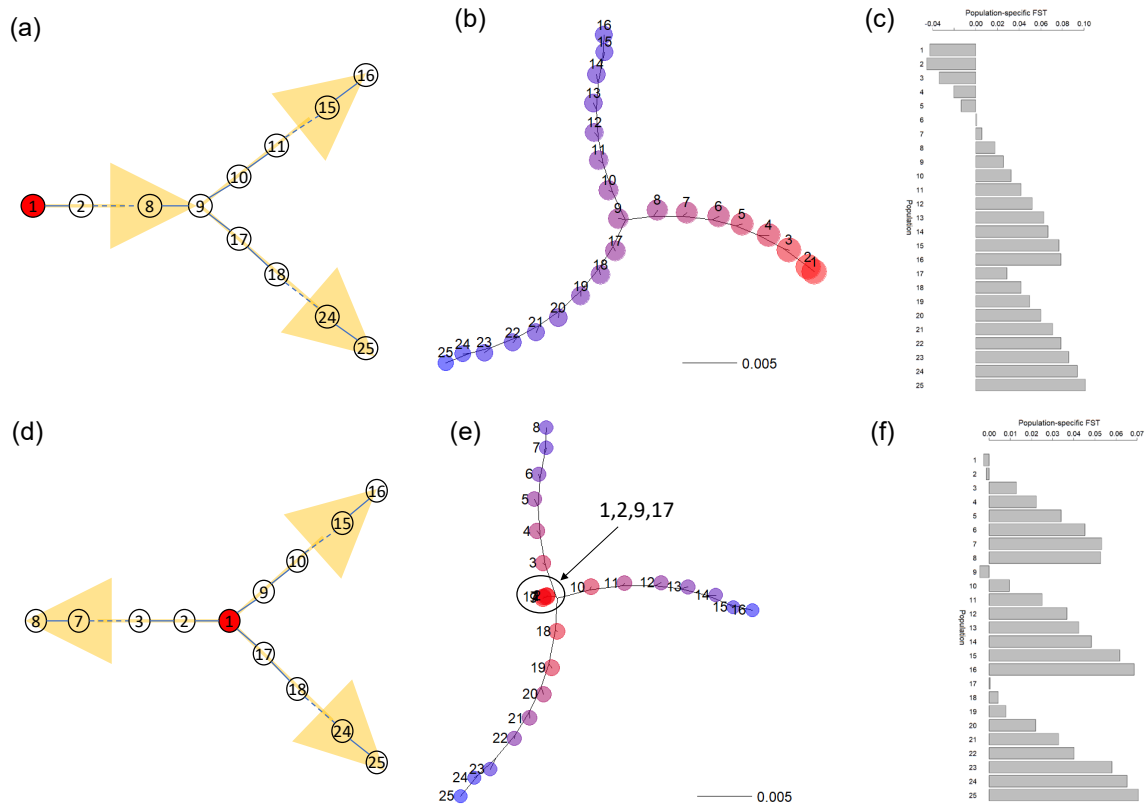
FIGURE 1 Results from simulations of population colonization. (a) Two- and (d) three-directional colonization models. Population 1 in red is ancestral, and arrows show the direction of colonization. Neighbor-joining unrooted trees based on pairwise F_{ST} distance matrix overlaid with population-specific 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 $(H_e \times 10)^3$. Population-specific F_{ST} values for 25 simulated populations are presented for the (c) two- and (f) three-directional models.

(a)

(b)
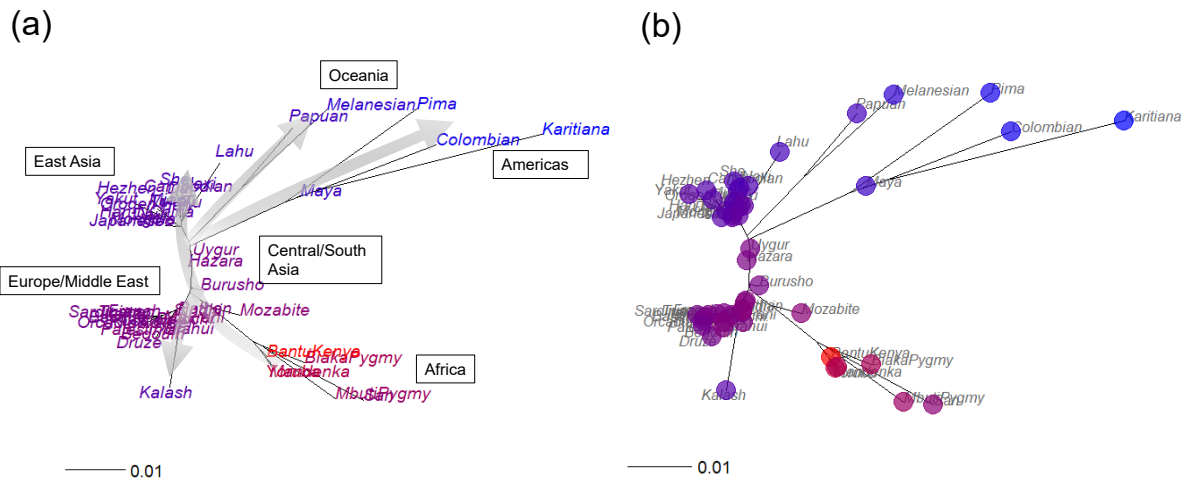
FIGURE 2 Population structure of 51 human populations (n = 1,035; 377 microsatellites). The unrooted NJ tree based on pairwise F_{ST} overlaid with population-specific F_{ST} values on (a) population labels and (b) population nodes. The arrows show inferred routes of population range expansion. The color of each population shows the magnitude of population-specific F_{ST} values. Data from Rosenberg *et al.* (2002).
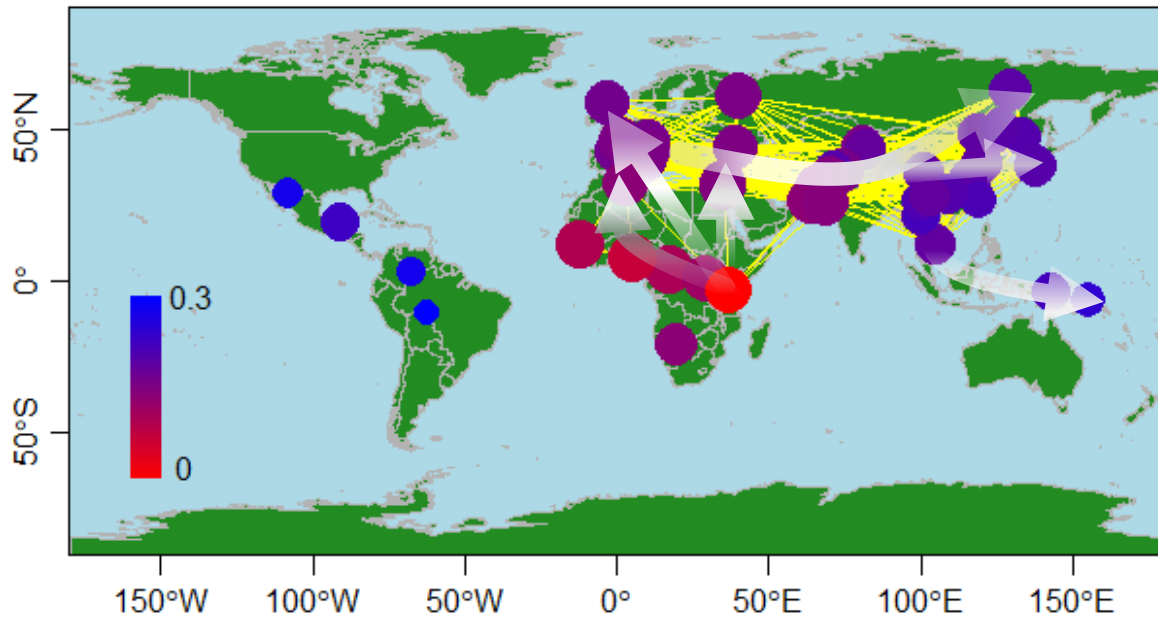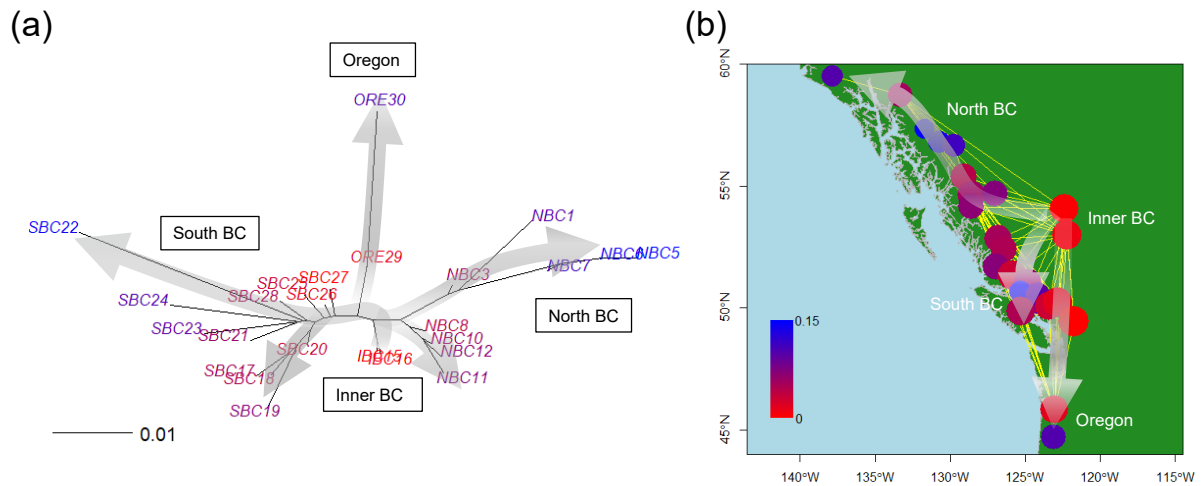
FIGURE 3 Map of the population structure of 51 human populations. The color of each population shows the magnitude of population-specific F_{ST} values. Populations connected by yellow lines are those with pairwise $F_{ST} < 0.01$. The radius of each sampling point is proportional to the level of *He* as visualized by H_e^{12}. The arrows show inferred routes of population expansion. Data from Rosenberg et al. (2002).
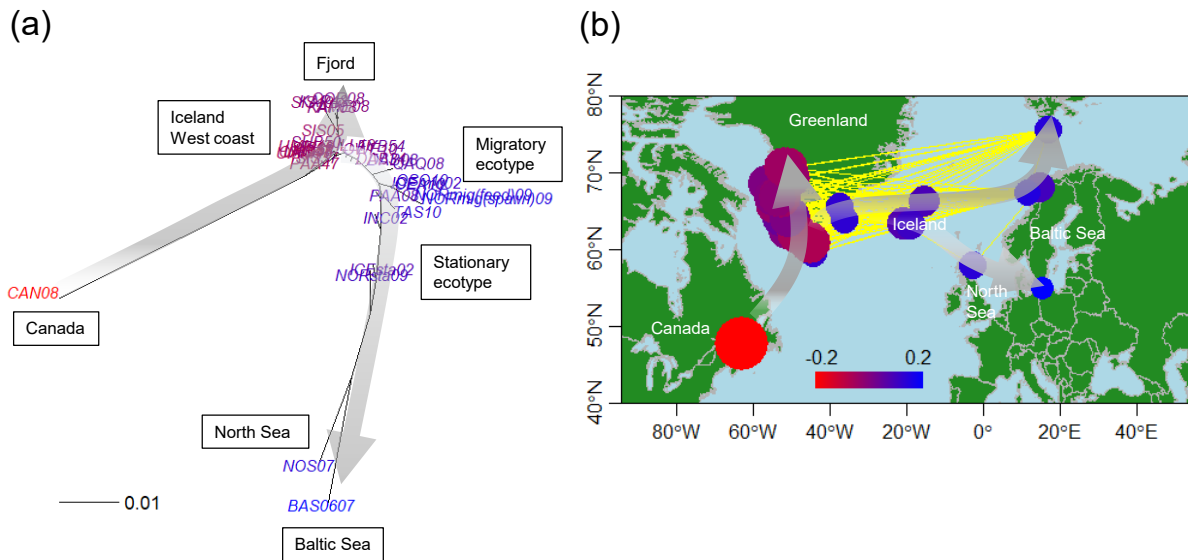
832
833
834 **FIGURE 4** Population structure of 34 geographical samples of wild Atlantic cod (n =
835 1,065; 921 SNPs). (a) Unrooted NJ tree based on pairwise F_{ST} and population-
836 specific F_{ST} values. (b) Map of the population structure of the Atlantic cod
837 populations. The color of each population shows the magnitude of population-specific
838 F_{ST} values. Populations connected by yellow lines are those with pairwise F_{ST} < 0.04.
839 The radius of each sampling point is proportional to the level of heterozygosity (He)
840 as visualized by H_e^{100}. The arrows show inferred routes of population expansion.
841 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 (He) as visualized by H_e^{100}. The arrows show inferred routes of population expansion. Data from McKown et al. (2014b).
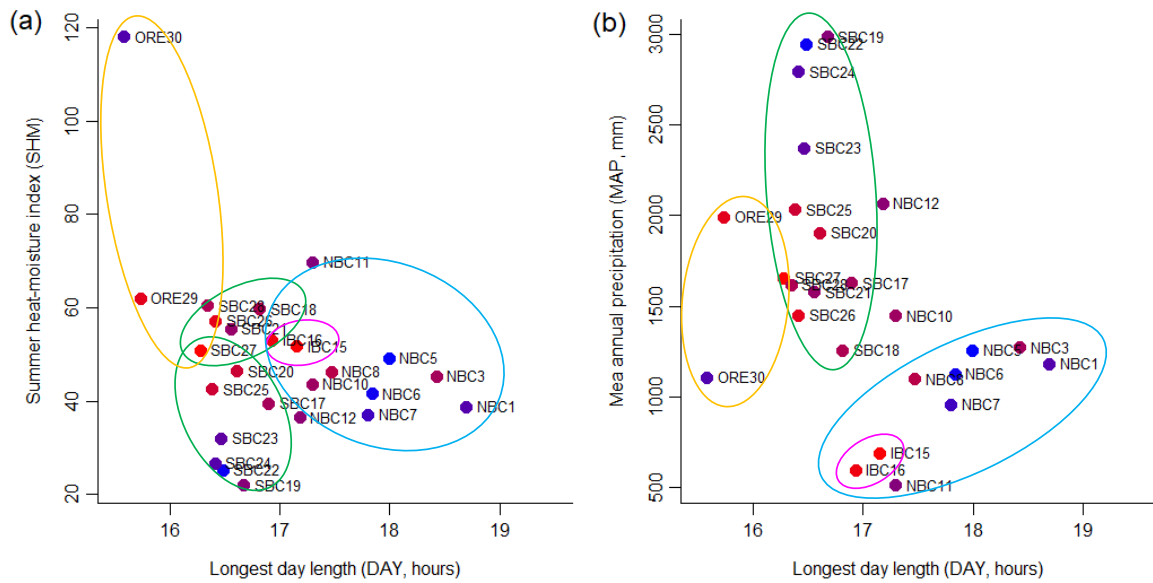
853
854 **FIGURE 6** Population range expansion and environmental adaptation. Longest day
855 length vs. (a) summer heat-moisture index and (b) mean annual precipitation for 25
856 geographical samples of wild poplar. The color of each population shows the
857 magnitude of population-specific F_{ST} values. The circles show inferred population
858 expansion from IBC15, IBC16, and SBC27. The color of the circles refers to the
859 population clusters (see, Figure S6).
860
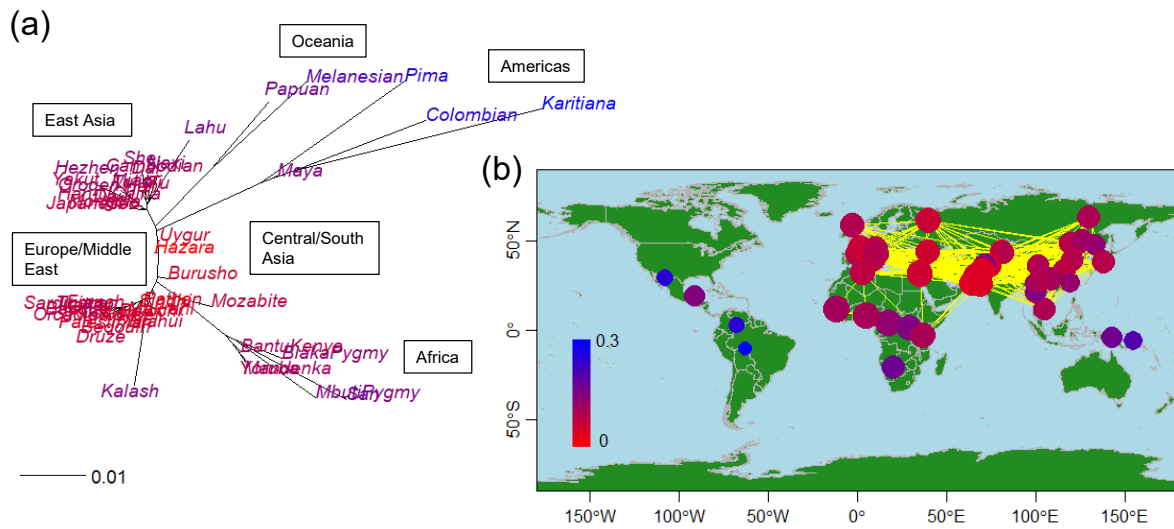
FIGURE 7 Population structure of 51 human populations inferred based on Bayesian population-specific F_{ST} values. (a) Unrooted NJ tree based on pairwise F_{ST} overlaid with population-specific F_{ST} values. (b) Map of the population structure. The color of each population shows the magnitude of Bayesian 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_e^{12}. Data from Rosenberg et al. (2002).