Estimating recent migration and population size surfaces

Hussein Al-Asadi*[1,2], Desislava Petkova[3], Matthew Stephens*[2,4], and John Novembre*[1,4]

[1]Committee on Evolutionary Biology, University of Chicago

[2]Department of Statistics, University of Chicago

[3]Department of Statistics, Oxford University

[4]Department of Human Genetics, University of Chicago

to whom correspondence shall be addressed*

July 9, 2018

Abstract

In many species a fundamental feature of genetic diversity is that genetic similarity decays with geographic distance; however, this relationship is often complex, and may vary across space and time. Methods to uncover and visualize such relationships have widespread use for analyses in molecular ecology, conservation genetics, evolutionary genetics, and human genetics. While several frameworks exist, a promising approach is to infer maps of how migration rates vary across geographic space. Such maps could, in principle, be estimated across time to reveal the full complexity of population histories. Here, we take a step in this direction: we present a method to infer separate maps of population sizes and migration rates for different time periods from a matrix of genetic similarity between every pair of individuals. Specifically, genetic similarity is measured by counting the number of long segments of haplotype sharing (also known as identity-by-descent tracts). By varying the length of these segments we obtain parameter estimates for qualitatively different time periods. Using simulations, we show that the method can reveal time-varying migration rates and population sizes, including changes that are not detectable when ignoring haplotypic structure. We apply the method to a dataset of contemporary European individuals (POPRES), and provide an integrated analysis of recent population structure and growth over the last ~3,000 years in Europe. Software implementing the methods is available at https://github.com/halasadi/MAPS.

1

1   Introduction

Populations exist on a physical landscape and often have limited dispersal. As a result, most genetic data exhibit a pattern of isolation by distance (Wright, 1943), which is simply to say that populations closer to each other geographically are more similar genetically. Furthermore, the degree of isolation by distance can vary across space and time (Manel et al., 2003). For instance, in a mountainous area of a terrestrial species' range, a pair of individuals may be more divergent from each other than a pair of individuals separated by the same distance in a flat and open area of the habitat. Additionally, the degree of isolation by distance can change over time – for example, if dispersal patterns are changing over time. Such spatial and temporal heterogeneity is an important aspect of population biology, and understanding it is crucial to solving problems in ecology (Turner et al., 2001), conservation genetics (Segelbacher et al., 2010), evolution (Rousset, 2004), and human genetics (Rosenberg et al., 2005).

Several methods have been developed to reveal spatial heterogeneity in patterns of isolation by distance (Womble, 1951; Barbujani et al., 1989; Guillot et al., 2005, 2009; Caye et al., 2016; Petkova et al., 2016; Bradburd et al., 2016, 2017). Some methods are based on explicitly modeling the spatial structure in the data (Guillot et al., 2005, 2009; Petkova et al., 2016; Bradburd et al., 2016, 2017); others take non-parametric approaches (e.g. Womble, 1951; Barbujani et al., 1989); while other methods ignore the spatial configuration of the samples and rely on researchers to make a *post hoc* geographic interpretation of the results (e.g. Pritchard et al., 2000; Patterson et al., 2006). However, none of these methods can be flexibly applied to address temporal heterogeneity in isolation by distance patterns, and new methods are needed.

One source of information for inferring changes in demography across time is the density of mutations observed in pairwise sequence comparisons (Li and Durbin, 2011; Schraiber and Akey, 2015). For example, when individuals are similar along a long segment of their chromosomes, it suggests that these segments share a recent common ancestor (Palamara et al., 2012). These segments are often called "identity-by-descent" tracts, although here we prefer the term "long pairwise shared coalescence" (lPSC) segments (as identity by descent traditionally required a definition of a founder generation, which is not clear in most data applications). A key feature of these segments is that filtering them by length provides a means to interrogate different periods of population history. The longest segments reflect the most recent population history, whereas shorter segments reflect longer periods of time. Recent analyses using lPSC segments suggest that they can reveal fine-scale spatial and temporal patterns of population structure that are not evident with genotype-based methods

2

65 such as principal components analysis (Ralph and Coop, 2013; Lawson et al., 2012; Leslie

66 et al., 2015).

67   Here we develop a new method to infer spatial and temporal heterogeneity in population

68 sizes and migration rates. The method takes as input geographic coordinates for a set of

69 individuals sampled across a spatial landscape, and a matrix of their genetic similarities as

70 measured by sharing of lPSC segments. It then infers two maps, one representing dispersal

71 rates across the landscape, and another representing population density. Crucially, building

72 these maps using different lengths of lPSC segments can help reveal changes in dispersal

73 rates and population sizes over time.

74   Our method is based on a stepping-stone model where randomly-mating subpopulations

75 are connected to neighboring subpopulations in a grid. Such models are parameterized

76 by a vector of population sizes $(\vec{N})$ and a sparse migration rate matrix ($\mathbf{M}$). Stepping-

77 stone models with a large number of demes can approximate spatially continuous population

78 models (Barton et al., 2002; Baharian et al., 2016), and this can be exploited to produce

79 maps of approximate dispersal rates and population density across continuous space.

80   Our method builds upon a method developed for estimating effective migration surfaces

81 (EEMS) (Petkova et al., 2016). While EEMS infers local rates of effective migration relative

82 to a global average, here we can explicitly infer absolute parameter values by leveraging lPSC

83 segments and modeling the recombination process [$\vec{N}$ and $\mathbf{M}$ values in the stepping-stone

84 model, and effective spatial density function $D_e(\vec{x})$ and dispersal rate function $\sigma(\vec{x})$ in the

85 continuous limit]. We call this method MAPS, for inferring Migration And Population-size

86 Surfaces.

87   We test MAPS on coalescent simulations and apply it to a European subset of 2,224

88 individuals from the POPRES data (Nelson et al., 2008). In simulations, we show that MAPS

89 can infer both time-resolved migration barriers and population sizes across the habitat. In

90 empirical data, we infer dispersal rates $\sigma(\vec{x})$ and population densities $D_e(\vec{x})$ across different

91 time periods in Europe.

2   Results

2.1   Outline of the MAPS method

94 MAPS estimates demography using the number of Pairwise Shared Coalescence (PSC) seg-

95 ments of different lengths shared between individuals. We define a PSC segment between

96 (haploid) individuals to be a genomic segment with a single coalescent time across its length

97 (Figure 1A). Long PSC (lPSC) segments tend to have a recent coalescent time, and so man-
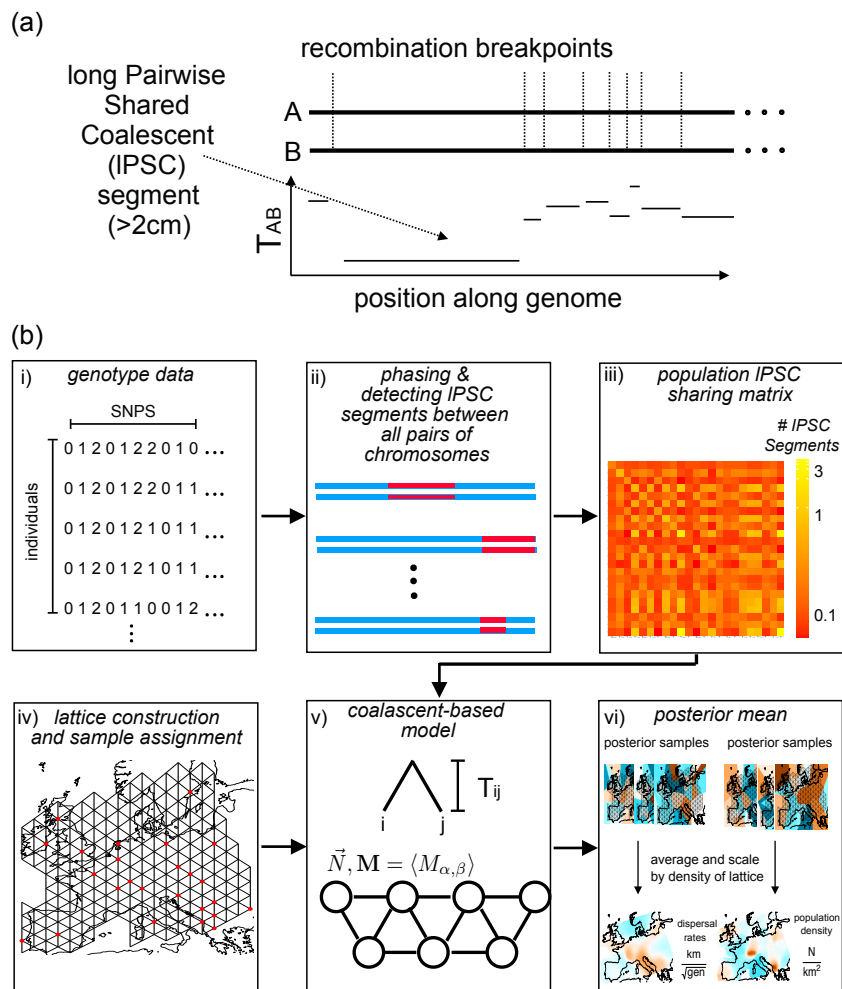
3

Figure 1: **Schematic overview of MAPS.** (a) Coalescent times between a pair of hapolo-types (A and B) will vary across the genome in discrete segments bordered by recombination breakpoints. On average, longer segments represent shorter pairwise coalescent times (T_{AB}) (b) Flow diagram of MAPS. i) We start with a matrix of called genotypes; ii) lPSC segments between all pairs of chromosomes across the genome are identified from the data using external methods (such as BEAGLE, Browning and Browning (2011)); iii) lPSC segments between pairs of individuals are aggregated at the levels of pairs of populations; iv) A grid is constructed and individuals are assigned to the most nearby node; v) The probability of the PSC sharing matrix can be computed under a stepping-stone model where each node represents a population and each edge represents symmetric migration; vi) We use an MCMC scheme to sample from the posterior distribution of migration rates and population sizes. The final MAPS output is the mean over these posterior samples, and the averaged rates can be transformed to units of dispersal rate and population density. The diagram does not show a bootstrapping step used to estimate likelihood weights to account for correlations between lPSC segments, see Equation (6) in Methods.

4

ifest themselves in genotype data as unusually long regions of high pairwise similarity, which can be detected by various software packages (Gusev et al., 2009; Browning and Browning, 2011, 2013; Chiang et al., 2016). Because lPSC segments typically reflect recent coalescent events, counts of lPSC segments are especially informative for recent population structure (Ringbauer et al., 2017; Palamara et al., 2012; Baharian et al., 2016). And partitioning lPSC segments into different lengths bins (e.g. 2-8cM, $\geq$8cM) can help focus inference on different (recent) temporal scales.

The MAPS model involves two components: i) a likelihood function, which relates the observed data (genetic similarities, as measured by sharing of lPSC segments) to the underlying demographic parameters (migration rates and population sizes); and ii) a prior distribution on the demographic parameters, which captures the idea that nearby locations will often have similar demographic parameters. The likelihood function comes from a coalescent-based "stepping-stone" model in which discrete populations (demes) arranged on a spatial grid exchange migrants with their neighbors (Figure 1b). The parameters of this model are the migration rates between neighboring demes ($M_{\alpha,\beta}$) and the population sizes within each deme (N_α). The prior distribution is similar to that from Petkova et al. (2016), and is based on partitioning the habitat into cells using Voronoi tesselations (one for migration and one for population size), and assuming that migration rates (or population sizes) are constant in each cell. We use an MCMC scheme to sample from the posterior distribution on the model parameters (migration rates, population sizes, and Voronoi cell configurations). We can summarize these results by surfaces showing the posterior means of demographic parameters across the habitat.

The inferred migration rates and population sizes will depend on the density of the grid used. However, using ideas from Barton et al. (2002) and Baharian et al. (2016) we convert them to corresponding parameters in continuous space, whose interpretation is independent of the grid for suitably dense grids. Specifically, we convert the migration rates to a spatial diffusion parameter $\sigma(\vec{x})$, often referred to as the "root mean square dispersal distance", which can be interpreted roughly as the expected distance an individual disperses in one generation; and we convert the population sizes ($\vec{N}$) to an "effective population density" $D_e(\vec{x})$ which can be interpreted as the number of individuals per square kilometer. Similar to the original grid-based demographic parameters, we can summarize MAPS results by surfaces showing the posterior means of $\sigma(\vec{x})$ and $D_e(\vec{x})$ across the habitat.

5

2.2 Differences from EEMS

Our MAPS approach is closely related to the EEMS method from Petkova et al. (2016), but there are some important differences. First, the MAPS likelihood is based on lPSC sharing, rather than a simple average genetic distance across markers. This was primarily motivated by the fact that, by considering lPSC segments in different length bins, MAPS can interrogate demographic parameters across different recent time periods. However, this change also allows MAPS, in principle, to estimate absolute values for the parameters $\mathbf{M}$ and $\vec{N}$, whereas EEMS can estimate only "effective" parameters which represent the combined effects of $\mathbf{M}$ and $\vec{N}$. This ability of MAPS to estimate absolute values stems from its use of a known recombination map, which acts as an independent clock to calibrate the decay of PSC segments. Finally, MAPS uses a coalescent model, whereas Petkova et al. (2016) uses a resistance distance approximation (McRae, 2006).

2.3 Evaluation of performance under a stepping-stone coalescent model

We assess the performance of MAPS with several simulations, and compare and contrast the results with EEMS. We used the program MACS (Chen et al., 2009) to simulate data under a coalescent stepping stone model and refinedIBD (Browning and Browning, 2011, 2013) to identify lPSC segments. All simulations involved twenty demes, each containing 10,000 diploid individuals, and each exchanging migrants with their neighbors. We analyzed each simulated data set using PSC segments of length 2-6cM and $\geq$6cM, which correspond to time-scales of approximately 50 generations and 12.5 generations respectively (see Lemma 5.3 in the Supplementary Note). Results for other length bins are qualitatively similar (Supplementary Figure S1 & S2).

Migration Rate Inference

First, we simulated under a uniform (constant) migration surface with migration rate 0.01 (Figure 2a), assumed to have stayed constant over time. In this case both EEMS and MAPS correctly infer uniform migration (Figure 2a), and MAPS provides accurate estimates of the migration rate (posterior mean 0.010 when using segments 2-6cM and 0.0086 using segments $\geq$6cM). As noted earlier, EEMS does not estimate the absolute migration rate; it estimates only the *relative* (effective) migration rates.

Next, we considered a scenario where the migration surface changed across time. Specifically the migration surface matches the constant migration scenario (above) until 10 genera-
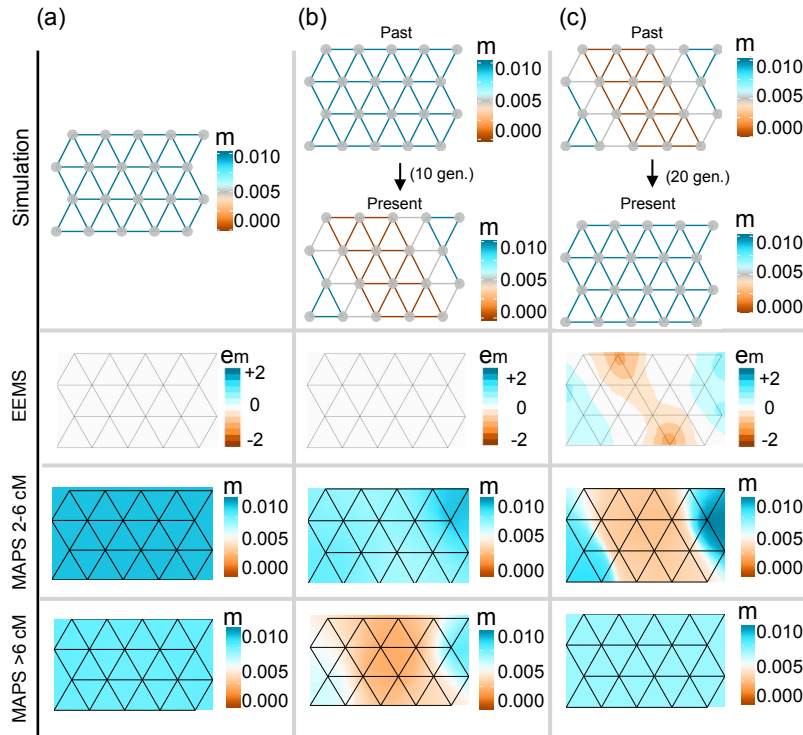
Figure 2: **Simulations comparing migration rates inferred with MAPS against effective migration rates inferred with EEMS.** (a) We simulated data under uniform migration rates equal to 0.01 and applied EEMS and MAPS using PSC segments in the range 2-6cM and $\geq$6cM. Like EEMS, MAPS correctly infers a uniform migration surface. Additionally, MAPS provides accurate estimates of the migration rates for both PSC segments 2-6cM (mean 0.01) and PSC segments $\geq$6cM (mean 0.0086). (b) We simulated a recent sudden migration barrier formation 10 generations ago. Here, EEMS is unable to infer a barrier, while MAPS correctly infers the historical uniform surface (2-6cM) and a barrier in the more recent time scale ($\geq$6cM). (c) We simulated a long-standing migration barrier that recently dissipated 20 generations ago. EEMS infers a barrier, while MAPS correctly infers both the historical migration barrier (2-6cM) and the uniform migration surface in the more recent time scale ($\geq$6cM). In all cases shown here, we simulated a 20 deme stepping stone model such that the population sizes all equal to 10,000, and 10 diploid individuals were sampled at each deme.

7

162 tions ago, when a complete barrier to gene flow instantaneously arose (a "vicariance event",
163 Figure 2b). In this setting EEMS again infers a uniform migration surface. This is because
164 EEMS is based on pairwise genetic distances, which are negligibly influenced by the recent
165 barrier. In contrast, by applying MAPS with different PSC segment lengths, we can see
166 both the historically uniform migration surface (for segments 2-6cM) and the recent barrier
167 (segments ≥6cM).

168 Next we consider a complementary time-varying scenario: an ancestral barrier disap-
169 peared 20 generations ago to allow uniform migration (Figure 2c). Here the EEMS results
170 again reflect the longer-term processes, and a barrier is evident. And again, by applying
171 MAPS with different PSC segment lengths, we can see different migration surfaces cor-
172 responding to different time scales, which are here reversed compared with the previous
173 scenario: the historical barrier (for segments 2-6cM) and the recent uniform migration (seg-
174 ments ≥6cM).

Population Size Inference

176 As noted above, and discussed in (Petkova et al., 2016), EEMS estimates an "effective"
177 migration surface that reflects the combined effects of population sizes $\vec{N}$ and migration
178 rates $\mathbf{M}$; consequently it cannot distinguish between variation in $\mathbf{M}$ and variation in $\vec{N}$. In
179 contrast, MAPS has the potential to distinguish these two types of variation.

180 To illustrate this difference we simulate data with a constant migration surface, and a
181 population size surface that has a 10-fold "dip" in the middle of the habitat (deme size 1,000
182 vs 10,000; Figure 3). Petkova et al. (2016) performed a similar simulation, and showed that
183 EEMS estimated an effective migration surface with an "effective barrier" in the middle,
184 caused by the dip in population size. As expected, we obtain a similar result for EEMS here.
185 Further, the EEMS inferred diversity surface is also approximately constant, because the
186 diversity surface reflects changes in within-deme heterozygosity, and these vary little in this
187 simulation. In contrast, MAPS is able to separate the influence of migration and population
188 sizes: the estimated migration surface is approximately constant (with mean migration rate
189 equal to the true value 0.01) and the estimated population size surface shows a dip in the
190 middle, with accurate estimates of deme sizes (mean 985 at the center and 9100 at the
191 edges). Additional simulations with non-uniform migration rates reinforce these results; see
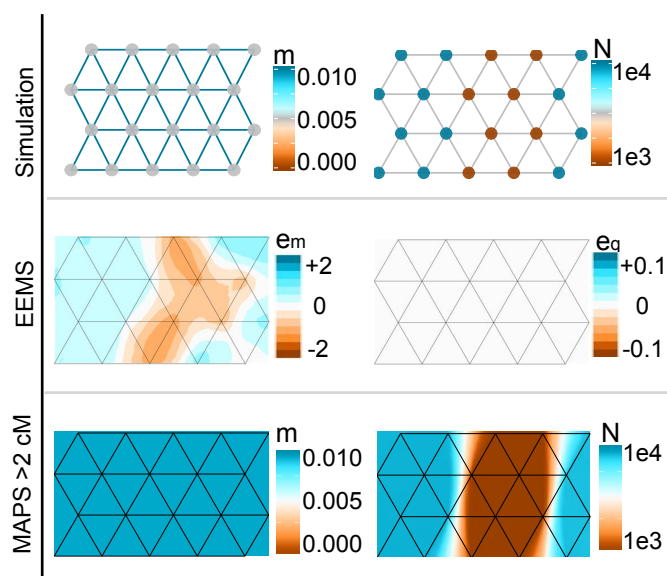192 Supplementary Figure S3.

Figure 3: **Simulations comparing population sizes inferred with MAPS and "diversity-rates" inferred with EEMS.** We simulated uniform migration rates of 0.01 and a trough of low population sizes in the center of the habitat such that population sizes equal to 1,000 at the center and 10,000 otherwise. Under these simulations, EEMS infers a barrier in effective migration and infers uniform diversity rates. However, MAPS correctly infers a uniform migration surface (mean 0.01) and provides accurate estimates of deme sizes (mean 985 at the center and 9100 at the edges)

9

2.4 Applying MAPS to the POPRES data

To illustrate MAPS on real data, we analyze a genome-wide SNP dataset on individuals of European ancestry (the "POPRES" study Nelson et al., 2008). Previous analyses of these data have shown the strong influence of geography on patterns of genetic similarity (Novembre et al., 2008; Lao et al., 2008; Ralph and Coop, 2013). In particular Ralph and Coop (2013) analyzed spatial patterns in the sharing of PSC segments across Europe. To facilitate comparison with their results, we use their PSC segment calls, focusing on a subset of 2224 individuals after filtering (see Methods).

We applied MAPS to these data using three different PSC segment length bins: $1-5$cM, $5-10$cM, and > 10cM. The longer bins correspond to more recent demography because as PSC lengths increase, the average coalescent times decrease. Indeed, the average coalescent times for each of these three length bins is inferred to be 90, 23 and 7.5 generations respectively (Supplementary Note), which correspond to 2700 years, 675 years and 225 years if we assume 30 years per generation.

We note that the accuracy of called PSC segments will vary across these bins: based on simulations in Ralph and Coop (2013) PSC segment calls in the smallest bin (1-5cM) will likely suffer from both false positives and false negatives, whereas for the longer bins PSC calls should be very reliable. Nonetheless, even in the smallest bin, closely-related individuals will still tend to show higher PSC sharing, and so the estimated MAPS surfaces should provide a useful qualitative summary of spatial patterns of variation even if quantitative estimates may be less reliable.

Inferring dispersal and population density surfaces

The inferred MAPS dispersal rates (migration rates scaled by grid step size) and population densities (population sizes scaled by grid area size) for each PSC length bin are shown in Figure 4.

Largely speaking, the spatial variation in inferred dispersal rates and population densities is remarkably consistent across the different time scales (Figure 4). In the MAPS dispersal surfaces, several regions with consistently low estimated dispersal rates coincide with geographic features that would be expected to reduce gene flow, including the English Channel, Adriatic Sea and the Alps. In addition we see consistently high dispersal across the region between the UK and Norway, which may reflect the known genetic effects of the Viking expansion (e.g Leslie et al., 2015). The MAPS population density surfaces consistently show lowest density in Ireland, Switzerland, Iberia, and the southwest region of the Balkans. This is consistent with samples within each of these areas having among the highest PSC seg-

10
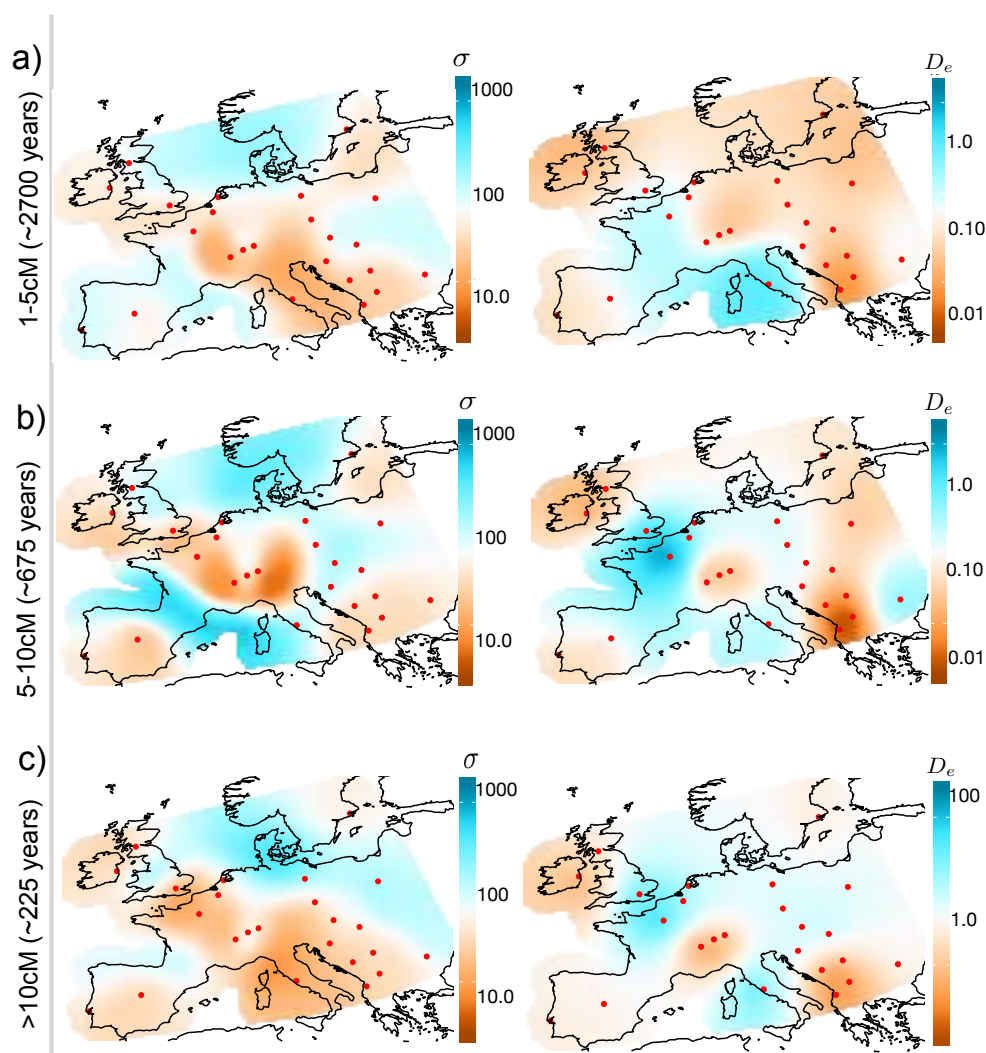
Figure 4: **Inferred Dispersal Surfaces and Population Density Surfaces over time for Europe.** We apply MAPS to a European subset of POPRES Nelson et al. (2008) with 2,234 individuals and plot the inferred dispersal $\sigma(\vec{x})$ and population density $D_e(\vec{x})$ surfaces for PSC length bins (a) > 1cM (b) 5-10cM and (c) >10cM. We transform estimates of $\vec{N}$ and $\mathbf{M}$ to estimates of $\sigma(\vec{x})$ and $D_e(\vec{x})$ by scaling the migration rates and population sizes by the grid step-size and area (see Equations (17) and (18)). Generally, we observe the patterns of dispersal to be relatively constant over time periods, however, we see a sharp increase in population density in the most recent time scale (>10cM). Note the wider plotting limits in inferred densities in the most recent time scale.

11

227 ment sharing (Supplementary Figure S4a). The MAPS inferred country population sizes are
228 also highly correlated with estimated current census population sizes from The World Bank
229 (2016) and National Records of Scotland (2011) (Supplementary Figure S6).

230   The most notable variation among the estimated surfaces from different time scales is a
231 dramatic increase in the mean estimated population density in the most recent time scale
232 (Figure 4 and Supplementary Figure S7). Indeed, the estimated mean for the last time
233 scale – 1.4 individuals per square km – is 6-9 fold higher than those for the earlier time
234 scales (0.16 and 0.22 respectively). This increase is consistent with the recent exponential
235 growth of human population sizes (Cohen, 1995). The estimates themselves are lower than
236 historical estimates of $\approx$1-30 individuals per square km based on archaeological data (e.g.
237 Zimmermann et al., 2009).

238   The dispersal surfaces show more minor changes between time periods (Figure 4 and
239 Supplementary Figure S7). In particular, the estimated mean dispersal rates are relatively
240 constant across time, being 73, 103 and 72 respectively (in units of km in a single genera-
241 tion). These mean estimates are consistent with empirical estimates of 10-100 km in a single
242 generation compiled by Kaplanis et al. (2018) using pedigrees of individuals living between
243 1650 and 1950 AD. We do note the lower estimated dispersal rates between Portugal and
244 Spain in the analyses of longer PSC segments (5-10 and $>$ 10cM), and the higher estimated
245 dispersal rates through the Baltic Sea ($>$ 10cM segments), possibly reflecting changing gene
246 flow in these regions in recent history.

247 Comparison to Ringbauer et al. (2017)

248 Ringbauer et al. (2017) also estimate a mean dispersal rate and population density from
249 the Eastern European subset of the data analyzed here. Their estimates are based on
250 PSC segments $>$ 4cM, which is most comparable with our analysis of 5-10cM. Unlike our
251 analysis, their estimates are based on a spatially homogeneous model. To compare with their
252 estimates we computed the mean of the estimated dispersal rate and population densities
253 in Eastern Europe (but based on an analysis of the full data). For the dispersal rate this
254 yields an estimate of 88 km in a single generation, which is consistent with the range of
255 50-100 given by (Ringbauer et al., 2017). For the population density, it yields an estimate
256 of 0.10 individuals per square km, which is somewhat higher than the estimate of 0.05
257 obtained under a comparable (time-homogeneous) population model in (Ringbauer et al.,
258 2017). Possibly our higher estimate partly reflects the influence of our spatial modeling
259 approach, which will tend to shift the estimate for Eastern Europe toward the estimated
260 mean across all of Europe (which is 0.22). In addition, the difference in length thresholds
261 ($>$ 4cM versus 5-10cM) may also be contributing; if segments in the Ringbauer et al. (2017)

262 analysis are on average shorter and hence older, one would expect lower density estimates,
263 based on our results that suggest lower densities in the past (Figure 4).

Comparison with EEMS

265 The EEMS results for these data (Figure S8) show non-trivial differences with the MAPS
266 results (Figure 4a). Two potential causes are: i) differences in the summary data used (PSC
267 segment sharing vs genetic distances) and hence sensitivity to different timescales; and ii)
268 differences in the underlying models (e.g. composite Poisson likelihood vs Wishart likelihood,
269 and different parameterizations/approximations to the coalescent model; see Discussion). To
270 evaluate the impact of i) we compared the PSC segment sharing and genetic distances, and
271 found their correlation to be only modest (Pearson's ρ = -0.38), with the most notable devi-
272 ation for comparisons between countries in Eastern Europe (Figure S9a). Furthermore, most
273 of this correlation is due to geographic distance: after controlling for geographic distance the
274 correlation is only -0.18, which may be a more relevant metric because inferred spatial het-
275 erogeneity in gene flow (barriers and corridors) is driven by departures from simple isolation
276 by distance.

277 To better assess the impact of ii) we applied EEMS on a distance matrix constructed
278 to have the same similarity patterns as the PSC segment sharing matrix input to MAPS
279 (1−5cM length bin). The resulting EEMS surface is more similar to the corresponding MAPS
280 dispersal surface (Supplementary Figure S9b vs Figure 4a), but there remain substantial
281 differences. This investigation confirms what we expected *a priori* — the two surfaces should
282 be different because the underlying models and inferred parameters of MAPS and EEMS
283 are different. As noted before, EEMS infers the "effective migration rate" which reflects the
284 effects of both the migration rates and population sizes, while MAPS infers them separately.

3 Discussion

286 We developed a method (MAPS) for inferring migration rates and population sizes across
287 space and time periods from geo-referenced samples. Our method builds upon a previous
288 method developed for estimating effective migration surfaces (EEMS) (Petkova et al., 2016).
289 However there are several differences between MAPS and EEMS. Most fundamentally, MAPS
290 draws inferences from observed levels of PSC sharing between samples, whereas EEMS draws
291 inferences from the genetic distance. These two data summaries capture different information
292 about the coalescent distributions: in essence, PSC sharing captures the frequency of recent
293 coalescent events, whereas genetic distance captures the mean coalescent time. Consequently
294 MAPS inferences largely reflect the recent past ($\lesssim$ 1000 years for human recombination

13

²⁹⁵ rates and generation times with PSC segments > 2cM), whereas EEMS inferences reflect
²⁹⁶ demographic history on a longer timescale across which pairwise coalescence occurs (99% of
²⁹⁷ events > 6000 years old, assuming diploid N_e of 10,000 for humans, exponential coalescent
²⁹⁸ time distribution).

²⁹⁹   Another consequence of modelling PSC sharing, rather than genetic distance, is that
³⁰⁰ MAPS can separately estimate demographic parameters related to migration rates ($\mathbf{M}$) and
³⁰¹ population sizes ($\vec{N}$), as in Figure 3 for example. In essence MAPS does this by using the
³⁰² known recombination map as an additional piece of information to help calibrate inferences.
³⁰³ In contrast EEMS, which makes no use of recombination maps, cannot separate $\mathbf{M}$ and $\vec{N}$.
³⁰⁴ Instead EEMS infers a compound parameter referred to as the "effective migration rate",
³⁰⁵ which is influenced by changes in both $\mathbf{M}$ and $\vec{N}$; see Figure 3. In principle, if applied
³⁰⁶ to sequence data instead of genotype data at ascertained SNPs, the genetic distances used
³⁰⁷ by EEMS could perhaps also separately estimate $\mathbf{M}$ and $\vec{N}$ by exploiting known mutation
³⁰⁸ rates to calibrate inferences. However, this would require non-trivial additional changes
³⁰⁹ to the current EEMS likelihood, which was designed to be applicable to ascertained SNPs
³¹⁰ and does not explicitly model variation in population sizes. (The EEMS likelihood instead
³¹¹ uses a "diversity rate" e_q, which reflects within-deme heterozygosity but is not explicitly a
³¹² population size parameter.)

³¹³   An additional useful feature of PSC segments is that, by varying the lengths analyzed,
³¹⁴ one can infer parameter values across different time scales. For example, our simulations
³¹⁵ show how by contrasting shorter and longer PSC segments, the method can reveal different
³¹⁶ gene flow patterns in scenarios with recent changes (see Figures 2 and 3). Further support
³¹⁷ comes from our empirical analysis of the POPRES data-set, where we found population sizes
³¹⁸ inferred from longer PSC segments to be more correlated with census sizes The World Bank
³¹⁹ (2015 census 2016) and National Records of Scotland (2011, 2011 census) than sizes inferred
³²⁰ from shorter segments (e.g. Spearman's $\rho = 0.71$ for $1 - 5$cM and $\rho = 0.84$ for > 10cM; see
³²¹ Supplementary Figures S5 and S6). Also, not surprisingly, PSC segments greatly outperform
³²² using heterozygosity as an indicator of census population size (the Spearman's correlation
³²³ coefficient between heterozygosity and census size was insignificant, p-value $= 0.25$).

³²⁴   Our estimates of dispersal distances and population density from the POPRES data are
³²⁵ among the first such estimates using a spatial model for Europe (though see (Ringbauer
³²⁶ et al., 2017)). The features observed in the dispersal and population density surfaces are in
³²⁷ principle discernible by careful inspection of the numbers of shared PSC segments between
³²⁸ pairs of countries (e.g. using average pairwise numbers of shared segments, Supplementary
³²⁹ Figure S4b, as in Ralph and Coop (2013)). For example, high connectivity across the North
³³⁰ Sea is reflected in the raw PSC calls: samples from the British Isles share a relatively high

331 number of PSC segments with those from Sweden (Supplementary Figure S4b). Also the low
332 estimated dispersal between Switzerland and Italy is consistent with Swiss samples sharing
333 relatively few PSC segments with Italians given their close proximity (Supplementary Figure
334 S4b). However, identifying interesting patterns directly from the PSC segment sharing data
335 is not straightforward, and one goal of MAPS (and EEMS) is to produce visualizations that
336 point to patterns in the data that suggest deviations from simple isolation by distance.

337 Our results suggest that several features of dispersal in Europe have been relatively stable
338 over the last ∼3000 years, whereas the population sizes have been increasing. The relative
339 stability of the gene flow patterns is perhaps surprising given ancient DNA results suggest
340 a continually dynamic history of population movements. One possibility is that much of
341 European population structure may have been established by the end of the Bronze Age
342 (4,000 years ago), with relatively more stable patterns in the intervening period that is
343 reflected in lPSC segments. Nonetheless, the dispersal is not completely stable– our results
344 suggest changes in Iberia, the Baltic, and to minor degrees in other areas.

345 The inferred population size surfaces for the POPRES data show a general increase in
346 sizes through time, with small fluctuations across geography; for instance, Polish samples
347 have a relatively larger population size in inferred values from the largest length scale ($>$
348 10cM). In our results, the smallest inferred population sizes are in the Balkans and Eastern
349 Europe more generally. This is in agreement with the signal seen by Ralph and Coop (2013);
350 however, taken at face value, our results suggest that high PSC sharing in these regions may
351 be due more to consistently low population densities than to historical expansions (such as
352 the Slavic or Hunnic expansions).

353 Although consistent with previous results, our estimates of dispersal and population sizes
354 do not exactly agree with empirical estimates. For example, our estimates of population
355 sizes are consistently lower than the census sizes (Supplementary Figure S6). This is to be
356 expected for several reasons. First, census sizes include non-breeding individuals (juvenile
357 and post-reproductive age) that do not impact the formation of PSC segments. Second,
358 MAPS is fitting a single population size per location, and in a growing population the best
359 fit population size will be an under-estimate of contemporary size. Third, in a wide class
360 of population genetic models, the effective size, even among reproductive age individuals, is
361 lower than the census size because of factors that inflate the variance in offspring number.
362 Fourth, some discrepancy is expected simply because the stepping-stone population genetic
363 model used here is only a coarse approximation to the complex spatial dynamics of human
364 populations. Finally, recombination rate mis-specification can bias the inferred parameters.
365 Furthermore, we caution that our results must be interpreted in the light of the fact that we
366 have limited spatial sampling across Europe, and only very coarse geographical origin data

367 (country of origin).

368 Here, as in Petkova et al. (2016) we use a discrete stepping-stone model to approximate
369 a process that might be more naturally modelled as continuously varying in space. Recent
370 work (Ringbauer et al., 2017; Baharian et al., 2016) exploits continuous models to estimate
371 dispersal and population density parameters from sharing of lPSC segments. However, these
372 methods assume that dispersal and population density are constant across space: extending
373 them to allow these parameters to vary across space could be an interesting avenue for future
374 work.

375 A major achievement in method development in population genetics would be to jointly
376 infer migration rates and population sizes across both space and time. MAPS is a step
377 towards this goal. However, we do not infer demography explicitly as a function of time and
378 instead infer surfaces in time blocks defined by PSC length bins. In principle, our method
379 allows for inference of demography across time by treating PSC segments as independent
380 across length bins, see Equation (S27) in Supplementary Methods. However, this requires
381 fitting multiple migration/population surfaces and is computationally unfeasible with our
382 current MCMC routine. Other computational techniques (e.g. Variational Bayes or fast
383 optimization of the likelihood) might make this goal possible.

384 4 Methods

385 4.1 MAPS configuration

386 For the empirical data analysis, we ran MAPS with 200 demes. The MAPS output was
387 obtained by averaging over 20 independent replicates (the number of MCMC iterations in
388 each replicate was to set 5e6, number of burn-in iterations set to 2e6, and we thinned every
389 2000 iterations). We provide the the MAPS here: https://github.com/halasadi/MAPS,
390 and the plotting scripts here: https://github.com/halasadi/plotmaps.

391 4.2 Inferring PSC segments from the data

392 Our pipeline to call PSC segments for simulations can be found here: https://github.com/
393 halasadi/ibd_data_pipeline. We follow the recommendations of Browning and Browning
394 (2011, 2013) and Ralph and Coop (2013) by running BEAGLE multiple times and merging
395 shorter segments.

396 For the empirical data analysis, we use the PSC segments ("IBD") calls from Ralph and
397 Coop (2013), which can be found here: https://github.com/petrelharp/euroibd. We

16

398 further applied a filter to retain countries with at least 5 sampled individuals, and removed

399 Russian and Greek individuals to restrict the habitat to a smaller spatial scale

4.3 Model

401 MAPS assumes a population genetic model consisting of triangular grid of d demes (or

402 populations) with symmetric migration. The density of the grid is pre-specified by the

403 user with the consideration that the computational complexity is $O(d^3)$. We use Bayesian

404 inference to estimate the MAPS parameters: the migration rates and coalescent rates M

405 and q respectively. Its key components are the likelihood, which measures how well the

406 parameters explain the observed data, and the prior, which captures the expectation that

407 M and $\underline{q}$ have some spatial structure (in particular, the idea that nearby edges will tend to

408 have similar migration rates and nearby demes have similar coalescent rates).

409 MAPS estimate the posterior distribution of $\Theta = M, \underline{q}$ given the data. The data used

410 for MAPS consists of a similarity matrix $X^R = \{X_{i,j}^R\}$ which denotes the number of PSC

411 segments in a range $R = [u, v]$ base-pairs shared between pairs of haploid individuals $(i, j) \in$

412 $\{1, \cdots, n\} \times \{1, \cdots, n\}$ where n is the number of (haploid) individuals. Furthermore, a

413 recombination rate map is required as input for MAPS. The likelihood is a function of the

414 expected value of $X_{i,j}^R$ ($E[X_{i,j}^R]$). Below we describe the computation of $E[X_{i,j}^R]$ and the other

415 key components of the likelihood. Finally, we briefly describe the prior used and an MCMC

416 scheme to sample from the posterior distribution of Θ.

The likelihood function

Let α, β denote the demes that (haploid) individuals i and j are sampled in, we define,

$$\lambda_{\alpha,\beta}^{\Theta} = E[X_{i,j}^R | \Theta]. \tag{1}$$

418 For the marginal distribution, we assume

$$X_{i,j}^R | \Theta \sim \text{Pois}(\lambda_{\alpha,\beta}^{\Theta} | \Theta), \tag{2}$$

419 and one option for computing the joint distribution of the data is to assume independence

420 between pairs of individuals (i, j) as done previously (Palamara et al., 2012; Palamara and

421 Peer, 2013; Ralph and Coop, 2013; Ringbauer et al., 2017). This assumption leads to the

422 log-likelihood,

17

$$log\mathcal{L}(\Theta; \bar{X}) = \sum_{\alpha \leq \beta} n_{\alpha,\beta}\left(\bar{X}_{\alpha,\beta}\log(\lambda_{\alpha,\beta}^{\Theta}) - \lambda_{\alpha,\beta}^{\Theta} \right), \tag{3}$$

where $\bar{X} = \{\bar{X}_{\alpha,\beta}\}$ such that $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, \cdots, d\}$ and d is the number of demes. Furthermore

$$\bar{X}_{\alpha,\beta} = \begin{cases} \frac{1}{n_\alpha n_\beta} \sum_{i \in d_\alpha, j \in d_\beta} X_{ij}^{R} & \text{if } \alpha \neq \beta \\ \frac{1}{\binom{n_\alpha}{2}} \sum_{i \in d_\alpha, i < j} X_{ij}^{R} & \text{if } \alpha = \beta \end{cases}, \tag{4}$$

where n_α is the number of sampled individuals in deme α, d_α is the set of all individuals in deme α, and

$$n_{\alpha,\beta} = \begin{cases} n_\alpha n_\beta & \text{if } \alpha \neq \beta \\ \binom{n_\alpha}{2} & \text{if } \alpha = \beta \end{cases}. \tag{5}$$

However, we found that there were significant correlations in lPSC segments between individuals. To deal with this, we down-weighted the likelihood function to reflect the "effective" number of samples ($e_{\alpha,\beta}$) instead of the number of pairs ($n_{\alpha,\beta}$). The effective number of samples between demes α,β is given by,

$$e_{\alpha,\beta} = \frac{\bar{X}_{\alpha,\beta}}{\text{Var}[\bar{X}_{\alpha,\beta}]}. \tag{6}$$

In the case of independence, $\text{Var}[\bar{X}_{\alpha,\beta}] \approx \frac{\bar{X}_{\alpha,\beta}}{n_{\alpha,\beta}}$. However, because of correlations in the data, the actual variance is significantly larger than the variance computed under an independence model. Here, we estimate $\text{Var}[\bar{X}_{\alpha,\beta}]$ by bootstrapping individuals with replacement. This way, we model the correlations between pairs of individuals for within and between-deme comparisons. The loglikelihood adjusted for correlations is given by,

$$log\mathcal{L}(\Theta; \bar{X}) = \sum_{\alpha \leq \beta} e_{\alpha,\beta}\left(\bar{X}_{\alpha,\beta}\log(\lambda_{\alpha,\beta}^{\Theta}) - \lambda_{\alpha,\beta}^{\Theta} \right). \tag{7}$$

Computing the expectation of $X_{i,j}^{R}|\Theta$

Next, we derive expressions to compute the expectation of the number of PSC segments of length greater than u ($X_{i,j}^{R=[\mu,\infty)}$) conditional on the demography Θ. From results in Palamara et al. (2012) it is easy to show that

$$E[X_{i,j}^{R=[\mu,\infty)}|\Theta] \approx G \int_{u}^{\infty} f_L(l|\Theta)/l \, dl, \tag{8}$$

433 where G denotes the length of the genome (in base-pairs), L denotes the random length (in
434 base-pairs) of the PSC segment between i and j containing a pre-specified position in the
435 genome (base b say), and f_L is its probability density. Intuitively, $Gf_L(l|\Theta)$ is the expected
436 number of base-pairs that lie in PSC segments of length l, making $\frac{Gf_L(l|\Theta)}{l}$ the expected
437 number of PSC segments of length l. Integrating the latter quantity from μ to ∞ gives the
438 desired result.

To help compute (8) we introduce T_{ij} to denote the (random) coalescent time in generations between i and j at base b, with density $f_{T_{ij}}(t|\Theta)$. Then (8) can be written as an integral over T_{ij}:

$$E[X_{i,j}^{R=[\mu,\infty)}|\Theta] \approx G \int_\mu^\infty f_L(l|\Theta)/l\,dl \tag{9}$$

$$= G \int_\mu^\infty \int_0^\infty f_{L,T_{i,j}}(l,t|\Theta)/l\,dt\,dl \tag{10}$$

$$= G \int_0^\infty f_{T_{i,j}}(t|\Theta) \int_\mu^\infty f_L(l|t)/l\,dl\,dt, \tag{11}$$

439 using the relation that $f_{L,T_{i,j}}(l,t|\Theta) = f_L(l|t,\Theta)f_{T_{i,j}}(t|\Theta) = f_L(l|t)f_{T_{i,j}}(t|\Theta)$. A key simplifi-
440 cation here comes from the fact that, given T_{ij}, L is conditionally independent of Θ.

It can be shown that the conditional distribution of L given T_{ij} is an erlang-2 distribution (Palamara et al., 2012; Palamara and Peer, 2013; Hein et al., 2004) with density

$$f_L(l|t) = 4r^2t^2le^{-2trl}, \tag{12}$$

where r is the recombination rate per base-pair. Substituting this into the inner integral of (11) and integrating analytically yields

$$\int_u^\infty f_L(l|t)/l\,dt = 2rte^{-2tru}, \tag{13}$$

leading to

$$E[X_{i,j}^{R=[\mu,\infty)}|\Theta] \approx G \int_0^\infty f_{T_{i,j}}(t|\Theta)2rte^{-2tru}dt. \tag{14}$$

Here, we assume the probability density of $T_{i,j}$ is given by,

$$f_{T_{i,j}}(t|\Theta) \approx \sum_\kappa q_k(e^{-Mt})_{\alpha,\kappa}(e^{-Mt})_{\beta,\kappa}, \tag{15}$$

441 where demes α,β denote the deme where lineages i and j are sampled from, $q_\kappa = \frac{1}{2N_\kappa}$ is
442 the coalescent rate in deme κ, and $M = \langle m_{\alpha,\beta}\rangle$ is the migration rate matrix between all d

19

demes such that $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, ..., d\}$. We compute the matrix exponential by first diagonalizing the matrix $M = PDP^T$ and compute $e^{-Mt} = Pe^{-Dt}P^T$.

Having computed all individual components of $\int_0^\infty f_{T_{i,j}}(t|\Theta)2rte^{-2tru}dt$, we are left to evaluate a one-dimensional integral which we do by Gaussian quadrature (with 50 weights).

To compute the expected number of PSC segments in a range $R = (\mu, \nu)$

$$E[X_{i,j}^{R=[\mu,\nu]}] = E[X_{i,j}^{R=[\mu,\infty)}] - E[X_{i,j}^{R=[\nu,\infty)}]. \tag{16}$$

As mentioned previously, the units of μ, ν are in base-pairs. However, we can transform to units of centiMorgans (cM) by : $\mu_{cM} = 100\mu r$.

The Prior

MAPS uses a hierarchical prior parameterized by Voronoi tessellation (similar to EEMS). The Voronoi tessellation partitions the habitat into C cells. Given a Voronoi tessellation of the habitat, each cell $c \in \{1, \cdots, C\}$ is associated with a migration rate $(\mathcal{M}_c)$ and population size $(\mathcal{N}_c)$. Demes (α) that fall into cell c will have population size $N_\alpha = \mathcal{N}_c$, and similarly, migration rates between demes α, β equal $m_{\alpha,\beta} = \frac{\mathcal{M}_{c_1} + \mathcal{M}_{c_2}}{2}$ if demes α, β fall into cells c_1 and c_2. We use an MCMC to integrate over the distribution on partitions of Voronoi cells. See Supplementary Notes section 5.4 for more information.

The MCMC

We break up the MCMC updates into updating a series of conditionally independent distributions. Provided the conditional posterior distributions for each part give support to all the parameter space, this will define an irreducible Markov chain with the correct joint posterior distribution Stephens (2000). We use Metropolis-Hastings to update all parameters, and random-walk proposals for most updates, with exception of birth-death updates for updating the number of Voronoi cells. See Supplementary Notes section 5.5 for more information.

Transformation of parameters to continuous space

Given an inferred population size at a particular deme α and a grid with uniform spacing, the transformation from population size to population density is given by

$$D_e(x) = \frac{N_\alpha}{\Delta A}, \tag{17}$$

20

466 where $\Delta A = \frac{\mathcal{A}_H}{d}$ is the area covered per deme such that $\mathcal{A}_H$ is the area of the habitat
467 (in km^2), d is the number of demes, and x corresponds to the spatial position of deme α.
468 Intuitively, (17) implies that the density multiplied by the area equals population size, i.e.
469 $D_e(x)\Delta A \approx N_\alpha$. Equation (17) can is analogous to equation 7 in (Baharian et al., 2016).

470   Given a migration rate (m), the transformation to dispersal distances is given by,

$$\sigma = \sqrt{m}\Delta x, \tag{18}$$

471 where Δx is the step size of the grid (km). The dispersal distance represents the distance
472 traveled by an individual after one generation, and sometimes is called the "root mean square
473 distance" or "dispersal rate" (Barton et al., 2002). Please see Supplementary Note section
474 5.2 for the derivation of (18).

Funding

Acknowledgements

5 Supplementary Note

5.1 The model

The coalescent process for two samples under a multi-deme model can be described by a continuous time Markov chain (CTMC) (Bahlo and Griffiths, 2001). Let i, j represent sampled lineages and α, β their locations, respectively, d is the number of demes (or populations) and $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, ..., d\}$. Let c denote the coalescent state. The infinitesimal rate matrix R of this CTMC is

$$
\begin{aligned}
R_{(\alpha,\beta),(\gamma,\beta)} &= m_{\alpha,\gamma} \quad \beta = 1, ..., d, \ \gamma \neq \alpha \\
R_{(\alpha,\beta),(\alpha,\gamma)} &= m_{\beta,\gamma} \quad \alpha = 1, ..., d, \ \gamma \neq \beta \\
R_{(\alpha,\alpha),(c)} &= q_\alpha \\
R_{(\alpha,\beta),(\alpha,\beta)} &= -(m_{\alpha_+} + m_{\beta_+}) - \delta_{\alpha\beta} q_\alpha \\
R_{(c),(c)} &= 0 \\
R_{(\alpha,\beta),(\gamma,\kappa)} &= 0 \quad \gamma, \ \kappa = 1, ..., d, \ \gamma \neq \alpha, \ \kappa \neq \beta,
\end{aligned}
\tag{S19}
$$

where $M = \langle m_{\alpha,\beta} \rangle$ denotes the migration rate matrix, and $m_{\alpha,\beta}$ is the migration rate between demes α, β and $q_\alpha = \frac{1}{2N_\alpha}$ is the coalescent rate of deme α which is proportional to the inverse of the population size at deme α (N_k). Let $T_{i,j}$ denote the (random) coalescent time between the pair of sampled lineages, and $f_{T_{i,j}}(t)$ denote the probability density of a coalescent event at time t. Here, we derive $f_{T_{i,j}}(t)$ by conditioning on the position of the two lineages.

Lemma 5.1 Let $(X_i(t), X_j(t)) \in \{1, \cdots, d\} \times \{1, ..., d\}$ denote the position of lineage i and lineage j at time t respectively. The probability density $f_{T_{i,j}}(t)$ that lineage i and j coalesce at time t is given by $\sum_{\kappa=1}^{d} q_\kappa P(X_i(t) = \kappa, X_j(t) = \kappa)$.

For $\Delta t \approx 0$,

$$
P(T_{i,j} \in [t, t + \Delta t]) \tag{S20}
$$

$$
\approx \sum_{\kappa=1}^{d} P(T_{i,j} \in [t, t + \Delta t] | X_i(t) = \kappa, X_j(t) = \kappa) P(X_i(t) = \kappa, X_j(t) = \kappa) \tag{S21}
$$

$$
\approx \sum_{\kappa=1}^{d} q_\kappa \Delta t P(X_i(t) = \kappa, X_j(t) = \kappa). \tag{S22}
$$

Taking the limit $\Delta t \to 0$, we arrive at the density

23

$$f_{T_{i,j}}(t) = \lim_{\Delta t \to 0} \; P(T_{i,j} \in [t, t + \Delta t])/\Delta t = \sum_{\kappa=1}^{d} q_\kappa P(X_i(t) = \kappa, X_j(t) = \kappa). \qquad (S23)$$

The random walk approximation to the coalescent

Here, we introduce an approximation,

$$P(X_i(t) = \kappa, X_j(t) = \kappa) \approx P(X_i(t) = \kappa)P(X_j(t) = \kappa). \qquad (S24)$$

The intuition is that probability that lineage i and j coalesce before time t is extremely small such that the two lineages approximately behave like two independently moving particles. Each lineage can be modeled by a random walk with transition matrix M. These assumptions were also made in the context of continuous spatial diffusion models for haplotype sharing Baharian et al. (2016); Ringbauer et al. (2017) , and even further back, as a general approximation to the two-dimensional continuous-space coalescent process (Barton et al., 2002; Wilkins, 2004; Blum et al., 2004; Novembre and Slatkin, 2009; Robledo-Arnuncio and Rousset, 2010).

This approximation implies that

$$f_{T_{i,j}}(t) \approx \sum_{\kappa} q_k (e^{-Mt})_{\alpha,\kappa} (e^{-Mt})_{\beta,\kappa}, \qquad (S25)$$

where lineages i, j are initially sampled in deme α, β. Or equivalently in matrix form,

$$f_{T_{i,j}}(t) \approx \left(e^{-Mt} Q e^{-Mt} \right)_{i,j}, \qquad (S26)$$

where $Q = diag(q_1, ..., q_d)$.

Varying migration rates and population sizes across time

Corollary 5.1.1 *Let time slice k be defined by the interval $t_{k-1} < t < t_k$, M_k denote the migration rate matrix in time slice k, and $Q_k = diag(q_1^k, ..., q_d^k)$ where q_α^k denotes the coalescent rate in deme α at time slice k. Let $T_{i,j}$ denote the coalescent time between lineage i, j sampled in demes α, β, then under the independence assumption, for $t \in (t_{K-1}, t_K)$,*

$$f_{T_{i,j}}(t) \approx \left(G_K(t) Q_K G_K(t) \right)_{\alpha,\beta}, \qquad (S27)$$

24

515 *where* $G_K(t) = \exp\left(-\sum_{k=1}^{K-1}(t_k - t_{k-1})M_k - (t - t_K)M_K\right).$

Expected number of lPSC segments given the demography Θ

517 **Lemma 5.2** *Let $X_{i,j}^\mu$ denote the number of PSC segment greater than μ basepairs shared*
518 *between haploid individuals i, j, Θ denote the demographic model, G the size of the genome,*
519 *L denotes the random length (in base-pairs) of the PSC segment between i and j containing*
520 *a pre-specified position in the genome, then $E[X_{i,j}^\mu|\theta] \approx G \int_u^\infty f_L(l|\Theta)/l\,dl$.*

521 Let $E[\mathcal{F}^\mu|\Theta]$ denote the expected fraction of the genome between i, j that lies in PSC seg-
522 ments greater than μ, and $E[s^\mu|\Theta]$ the expected size of a PSC segment conditional on it
523 being at least length μ. According to equations 9-14 from (Palamara et al., 2012),

$$E[X_{i,j}^\mu|\theta] \approx \frac{G\,E[\mathcal{F}^\mu|\theta]}{E[s^\mu|\Theta]}, \tag{S28}$$

$$E[\mathcal{F}^\mu|\Theta] = \int_\mu^\infty f_L(l|\Theta)dl, \tag{S29}$$

$$E[s^\mu|\Theta] = \frac{\int_\mu^\infty f_L(l|\Theta)dl}{\int_\mu^\infty f_L(l|\Theta)/l\,dl}. \tag{S30}$$

524 We obtain the desired result by substituting (S29) and (S30) into (S28) and canceling like-
525 terms.

Expected age of a segment

527 We choose PSC segment lengths based on their expected age which is derived below.

528 **Lemma 5.3** *The expected coalescent time (t, in generations) of an PSC segment between*
529 *between length L_1 centiMorgans and L_2 centiMorgans is approximately $\frac{300}{4}\left(\frac{1}{L_1} + \frac{1}{L_2}\right)$ if the*
530 *effective population size (N) is sufficiently large.*

531 We choose to work in units of basepairs, and will convert back to units of morgans at the
532 end. We convert L_1 into units of base-pairs with the transformation: $\mu = \frac{L_1}{100r}$ and similarly
533 $\nu = \frac{L_2}{100r}$.

Let us denote $T|l, N$ as the random coalescent time of a PSC segment that is at least
length l under a single-deme demography model with population size N. The expected

coalescent time of an PSC segment longer than μ base-pairs can be expressed as

$$E[T|l \geq \mu, N] = \int_0^\infty t f_T(t|l \geq \mu, N) dt = \int_0^\infty t \frac{f_L(l \geq \mu|t) f_T(t|N)}{f_L(l \geq \mu|N)} dt$$
$$= \frac{\int_0^\infty t f_L(l \geq \mu|t) f_T(t|N) dt}{\int_0^\infty f_L(l \geq \mu|t) f_T(t|N) dt},$$

(S31)

where $f_L(l|t) = 4r^2 t^2 l e^{-2trl}$ denotes the probability density that a PSC segment is of length l given it has a common ancestor event at time t, $f_T(t|N)$ denotes the probability density that a coalescent event occurs at time t under the demography model with population size N.

Next, we expand a key term in equation (S31)

$$f_L(l \geq \mu|t) = \int_\mu^\infty f_L(l|t) dl = (2rt\mu + 1) \exp\left(-2rt\mu + 1 \right)$$

(S32)

and assume,

$$f_T(t|N) = \frac{e^{-t/N}}{N}.$$

(S33)

Putting everything together,

$$E[T|l \geq \mu, N] = \frac{N(1 + 6Nr\mu)/(1 + 2Nr\mu)^3}{(1 + 4Nr\mu)/(1 + 2Nr\mu)^2} = \frac{N(1 + 6Nr\mu)}{1 + 6Nr\mu + 8N^2(r\mu)^2}.$$

(S34)

We can remove the dependence of N by taking $\lim_{N\to\infty}$ as done similarly in Baharian et al. (2016),

$$\lim_{N\to\infty} E[T|l \geq \mu, N] = \frac{3}{4r\mu}$$

(S35)

Now that we have derived the expected age of PSC segment longer than μ, it is quite simple to expand the equation for PSC segments between μ and ν base-pairs,

$$E[T|\mu \leq l \leq \nu] = \frac{\int_0^\infty t f_L(\mu \leq l \leq \nu|t) f_{T|N}(t) dt}{\int_0^\infty f_L(\mu \leq l \leq \nu|t) f_{T|N}(t) dt} = \frac{\int_0^\infty t \left(f_L(l \geq \nu|t) - f_L(l \geq \mu|t) \right) f_{T|N}(t) dt}{\int_0^\infty \left(f_L(l \geq \nu|t) - f_L(l \geq \mu|t) \right) f_{T|N}(t) dt}$$
$$= \frac{3}{4} \left(\frac{1}{r\mu} + \frac{1}{r\nu} \right)$$

(S36)

26

541 We transform back to units of centimorgans: let $L_1 = 100r\mu$ and $L_2 = 100r\nu$ be in units of
542 centiMograns, and taking the limit, we get the desired result

$$\lim_{N \to \infty} E[t | \mu \leq l \leq \nu] = \frac{300}{4}\left(\frac{1}{L_1} + \frac{1}{L_2}\right). \tag{S37}$$

543 5.2 Transformation of migration rates to dispersal rates

544 Migration rates inferred under a discrete model can be transformed to dispersal distances
545 representing parameters in continuous space. Here, we derive the transformation.

546 **Lemma 5.4** *Consider a random walk on a 2D grid, where steps are taken according to a*
547 *Poisson process with rate m, and let $\underline{X}(t)$ be a vector denoting the coordinates of the particle*
548 *at time t. The distribution of $\underline{X}(t)$ approximately only depends on the compound parameter*
549 *$m(\Delta x)^2$ (or equivalently $\sqrt{m}\Delta x$).*

$$\underline{X}(t) = \sum_{i=1}^{N(t)} \underline{Z}_i, \tag{S38}$$

550 where $N(t)$ is the number of steps taken by time t, and $\underline{Z}_i$ is a random variable representing
551 the direction and magnitude taken at step i. Since $X(t)$ is a sum of iid variables, a form of
552 the central limit theorem applies here and $X(t)$ converges to the normal distribution (Rényi,
553 1960).
554 In a random walk on a triangular grid, a particle can move in one of the 6 directions
555 (upper-right, right, lower-right, left, upper-left, and lower-left):

$$
\begin{aligned}
\underline{Z}_i \\
&= (1/2, \Delta x\sqrt{3}/2)^T \text{ with } p = 1/6 \\
&= (\Delta x, 0)^T \text{ with } p = 1/6 \\
&= (\Delta x/2, -\Delta x\sqrt{3}/2)^T \text{ with } p = 1/6 \\
&= (-\Delta x, 0)^T \text{ with } p = 1/6 \\
&= (-\Delta x/2, \Delta x\sqrt{3}/2)^T \text{ with } p = 1/6 \\
&= (-\Delta x/2, -\Delta x\sqrt{3}/2)^T \text{ with } p = 1/6
\end{aligned}
$$

556 where Δx represents the step size in the grid (i.e. edge length). The mean and variance are
557 given by,

27

$$E[\underline{X}(t)] = 0 \tag{S39}$$

and,

$$Var[\underline{X}(t)] = \frac{mt(\Delta x)^2}{2} I_2. \tag{S40}$$

where I_2 is the identity matrix. Under normality, the mean and variance are sufficient statistics. Note that (S39) and (S40) also hold for square grids.

Interpretation of the migration diffusion parameter $m(\Delta x)^2$

In addition, we provide a physical interpretation to $(\Delta x)^2$ in terms of the squared distance from the origin per generation. Let the distance $d = \|\underline{X}(t)\| = \sqrt{x_1^2 + x_2^2}$, then

$$E[d^2]/t = E[x_1^2 + x_2^2]/t = E[x_1^2]/t + E[x_2^2]/t = \frac{m(\Delta x)^2}{2} + \frac{m(\Delta x)^2}{2} = m(\Delta x)^2. \tag{S41}$$

$\sqrt{\frac{E[d^2]}{t}} = \sqrt{m}\Delta x$ can be interpreted as the distance traveled by an individual after one generation, and sometimes is called the "dispersal" distance or the "root mean square distance".

5.3 Diversity rates versus coalescent rates

For computational efficiency, the EEMS software uses a combination of the resistance distance model and within-deme "diversity rates" to approximate expected pairwise coalescent times, in which,

$$E[\hat{T}_{\alpha,\beta}] = \begin{cases} \frac{R_{\alpha,\beta}}{4} + \frac{e_{q_\alpha} + e_{q_\beta}}{2} & \text{if } \alpha \neq \beta \\ e_{q_\alpha} & \text{if } \alpha = \beta \end{cases}. \tag{S42}$$

where $E[\hat{T}_{\alpha,\beta}]$ is the resistance distance approximation to the expected coalescent time between deme α and deme β, e_{q_α} is the "diversity rate" in deme α, and $R_{\alpha,\beta}$ is the resistance distance between demes α, β (Petkova et al., 2016). The diversity rates have no simple expression in terms of population-genetic parameters under the multi-deme coalescent model. As an alternative, diversity rates can be interpreted as reflecting average within deme heterozygosity since $e_q = E[\hat{T}_w] \propto H_\alpha$ where the heterozygosity for deme α (H_α) is defined

28

as,

$$H_\alpha = \frac{1}{\binom{n_\alpha}{2}} \sum_{i<j, i\in\alpha, j\in\alpha} D_{i,j}, \tag{S43}$$

where $D_{i,j}$ is the average number of differences between (haploid) individuals i and j.

Migration and population sizes are identifiable in MAPS

MAPS models the recombination process using rates estimated from a recombination rate map. In this model, population sizes and migration rates can be inferred separately rather than as a joint parameter. Intuitively, the recombination rate serves an independent clock to calibrate estimates.

More formally, a statement of identifiability is a statement regarding the likelihood. MAPS models the expected number of lPSC segments shared between pairs of (haploid) individuals, and can be computed with an integral (14). The integral can be broken up into a product of two functions: a function describing the decay of PSC segments as a function of time ("recombination rate clock"), and the coalescent time probability density $f_{T_{i,j}}(t)$ (15). The migration rates and population sizes only appear in $f_{T_{i,j}}(t)$, and cannot be cannot be factored into parameters involving combinations of the migration rates and population sizes.

5.4 The prior

The structure of the prior closely resembles the prior in the EEMS method Petkova et al. (2016). The tessellation for the migration rates (T_m) is encoded by a list $(\underline{l}^m, \underline{m}, c_m, \mu_m)$ where $\underline{l}^m$ are the locations of each cell, $\underline{m}$ the rates of each cell, and are vectors of length c_m (i.e. number of Voronoi cells), and μ_m is the overall mean migration rate. The Voronoi tessellation for the coalescent rates is $T_q = (\underline{l}^q, \underline{q}, c_q, \mu_q)$.

The location of each (unordered) Voronoi cell is distributed uniformly across the habitat,

$$l_c^m \overset{iid}{\sim} U(H), \tag{S44}$$

and the number of cells (a-priori) are drawn from a negative binomial distribution,

$$c_m \sim \text{NegBi}(r_m, p_m). \tag{S45}$$

The effects of each Voronoi cell is normally distributed with variance ω^2.

$$\log 10(m_i) \overset{iid}{\sim} N(\mu_m, \omega_m^2) \tag{S46}$$

29

$$\log 10(q_i) \overset{iid}{\sim} N(\mu_q, \omega_q^2) \tag{S47}$$

The probability of a particular (unordered) cell configuration is,

$$p(\underline{m}|c_m) = c_m! \prod_{i=1}^{c_m} N(m_i|\mu_m, \omega_m^2) \tag{S48}$$

We assume,

$$\log 10(\omega_m) \sim U(-3, \log 10(1.5)) \tag{S49}$$

$$\log 10(\omega_q) \sim U(-3, log10(1)) \tag{S50}$$

We set $\log 10(2)$ as the upper bound for $\log 10(\omega_m)$ so the m so the probability that it is within 3 orders of magnitude from the mean is 0.95 a priori, and we set $\log 10(1)$ as the upper bound for $\log 10(\omega_q)$ to restrict the population sizes so to be within 2 orders of magnitude from the mean with probability 0.95 a priori.

We assume,

$$\mu_m \sim U(-10, 4) \tag{S51}$$

$$\mu_q \sim U(-10, 4). \tag{S52}$$

We place a uniform prior on the log of the mean rates to reflect that we are uncertain about the order of magnitude. Here, the data is highly informative of the mean, as a result, we can allow the support of the prior to vary by many orders of magnitude.

5.5   MCMC

Re-parameterization

We re-parameterize the model to improve mixing of the MCMC. We decouple the migration (or coalescent) rates from the mean rate (μ), and variance (ω) by introducing a new variable e_i,

$$e_i \overset{iid}{\sim} N(0, 1) \tag{S53}$$

and the cell specific migration (or coalescent) rates are computed as,

$$\log 10(m_i) = e_i\omega + \mu, \tag{S54}$$

which allows us to update the magnitude of the parameters (μ) and the variance scale (ω) separately.

We add MH joint random-walk updates to μ and e_i to ensure that $\bar{e} = \frac{\sum_i e_i}{c} \approx 0$. To do this, we jointly update μ and e_i by,

$$\mu' = \mu + \epsilon \tag{S55}$$

$$e_i' = e_i - \frac{\epsilon}{\omega} \tag{S56}$$

where $\epsilon \sim N(0,1)$. We do this for both the migration rates and population sizes.

Updating the number of cells

The number of cells change the dimension of the likelihood, and a result, we must use a Reversible Jump MCMC step so that the ratio of densities in the Metropolis-Hastings acceptance ratio is well-defined (Green, 1995). We choose to update the number of cells with a birth-death update (Stephens, 2000). Fortunately, in such a case, the updates reduce to standard Metropolis-Hastings because the dimension matching constant (i.e. the "Jacobian") equals one (Petkova et al., 2016; Stephens, 2000). See equations S31 and S32 in Petkova et al. (2016) for formulas regarding the birth-death update. Here, we use nearly identical updates (with a slight modification).

When increasing the number of cells from c to $c+1$ (i.e. a birth-update), we randomly choose a location uniformly across the habitat, and the new migration is proposed from a standard normal because our cell effects are standardized. In contrast, EEMS proposes cell effects migration to be normally distributed around a cell effect at a randomly chosen point in the habitat. Here we set, p(birth) = p(death) = 0.5 if the number cells ≥ 1, otherwise p(birth) = 1.

The acceptance ratio for a birth update (going from c cells to $c+1$ cells) is

$$\alpha(x, x') = min(1, \frac{p(\text{death})}{p(\text{birth})} \frac{l(x')p(x')\frac{1}{c+1}}{l(x)p(x)N(e_{c+1}|0,1)}), \tag{S57}$$

where x denotes the current state of the MCMC, x' the proposed state, e_{c+1} is the proposed cell effect drawn from a standard normal. Conversely, in a death-update, we randomly choose one cell uniformly to kill. In this case, the acceptance ratio for a death proposal (going from

31

$c + 1$ cells to c cells) is

$$\alpha(x, x') = min(1, \frac{p(\text{birth})}{p(\text{death})} \frac{l(x')p(x')N(e_c|0, 1)}{l(x)p(x)\frac{1}{c+1}}).$$
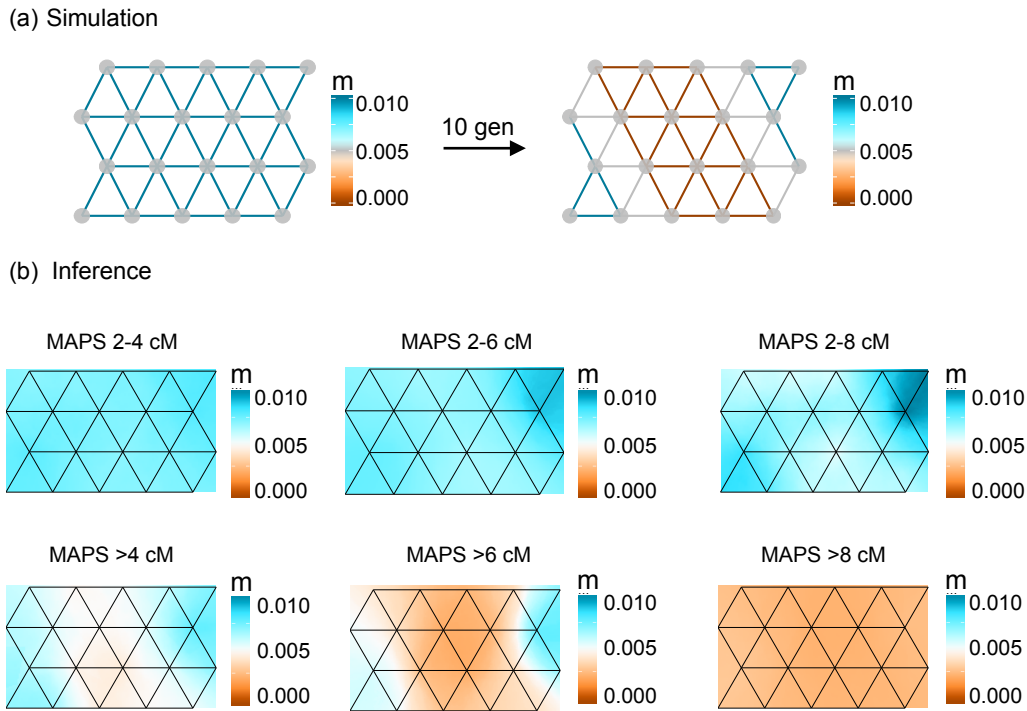
(S58)

6 Supplementary Figures

Figure S1: **The performance of MAPS on a recent barrier scenario under different PSC length bins.** Here, we investigate the ability of MAPS to detect a recent barrier (< 10 generations) for various PSC length bins (a) Simulation scenario. Population sizes were set to 10,000 per deme and 10 diploids were sampled per deme, replicating the conditions in Figure 2b. (b) Results for different PSC length bins. Length bins that encompass shorter segments (2-4cM 2-6cM 2-8cM) recover the higher uniform migration surface; while length bins with longer segments (>4, >6, >8) recover the recent ancestral barrier. For the last length scale (> 8cM), the signature of low migration extends across the habitat. The variation in migration rates is missed presumably because of the small number of shared segments at this length scale.

(a) Simulation

(b) Inference
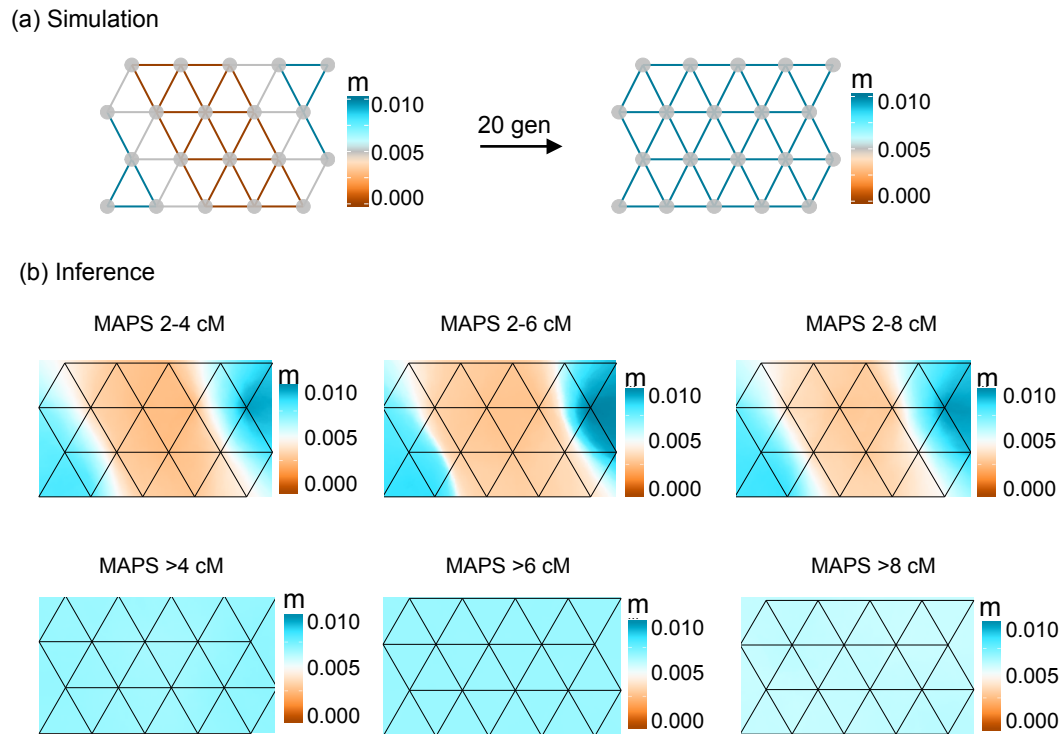
Figure S2: **The performance of MAPS on a past barrier scenario under different PSC length bins.** a) Simulation scenario. Population sizes were set to 10000 per deme and 10 diploids were sampled per deme, replicating the conditions in Figure 2c. (b) Results for different PSC length bins. Length bins that encompass shorter segments (2-4cM, 2-6cM, 2-8cM) recover the ancestral barrier; while length bins with longer segments (>4, >6, >8) recover the recent constant migration surface.
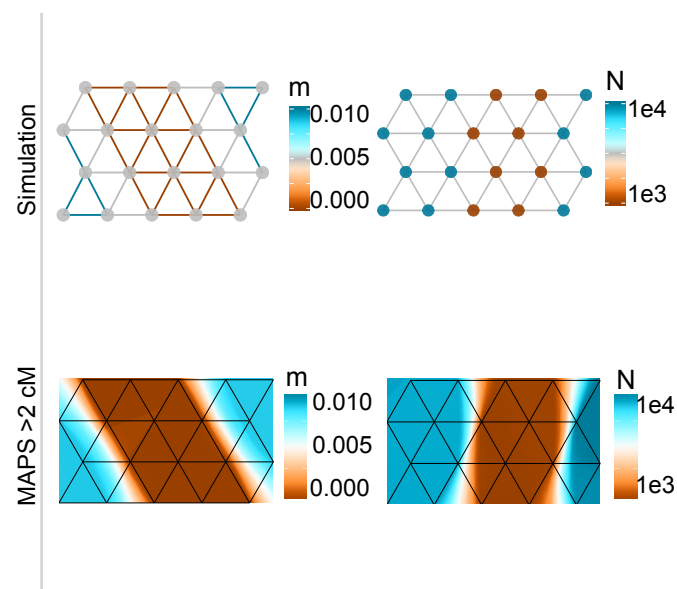
Figure S3: **The performance of MAPS under a jointly heterogeneous migration rate and population size surface.** a) Simulation Scenario. Heterogeneous population-sizes and migration rates (as shown) were simulated, and 10 diploid individuals were sampled per deme. (b) Results for PSC segments greater than 2cM are shown.
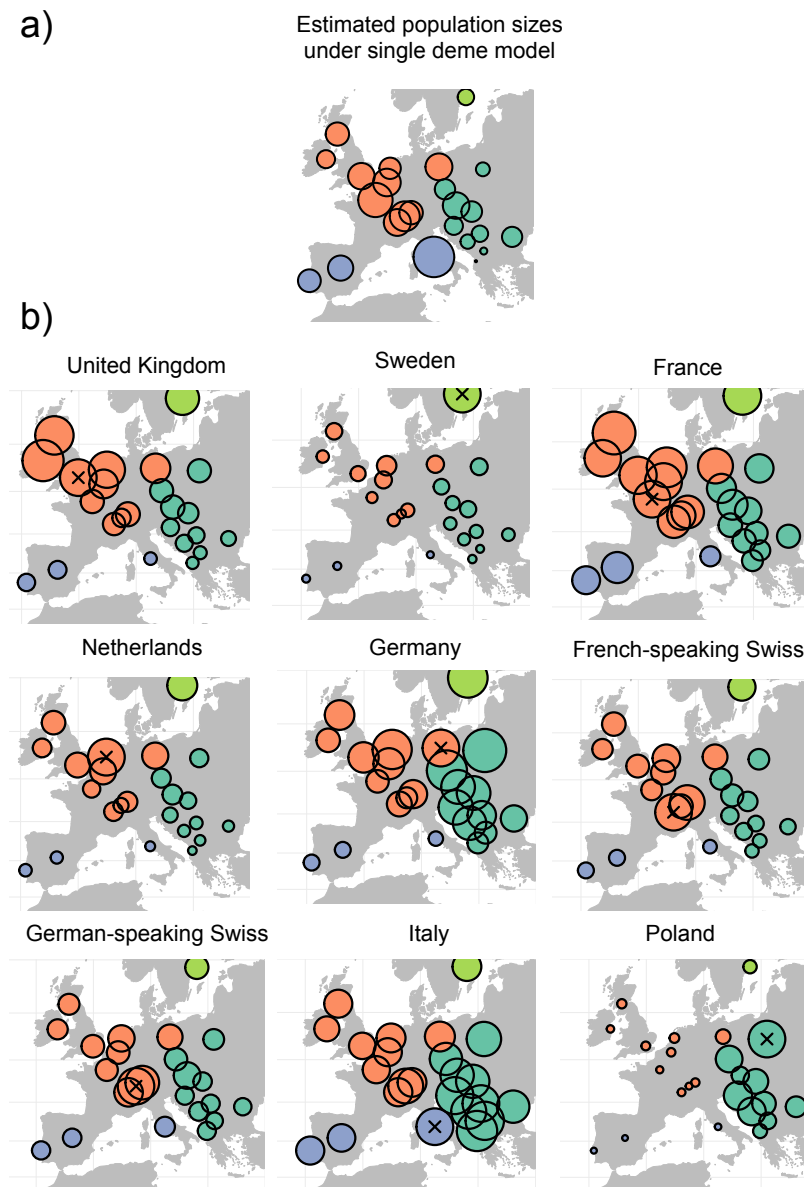
Figure S4: **Visualizing normalized sharing of PSC segments that are 1-5cM.** The color scheme is the same as used in Ralph and Coop (2013) where the colors give categories based on the regional groupings: W Western Europe, S Southern Europe, and E Eastern Europe (a) The average sharing within each sample locale is transformed to population sizes using the simple single deme estimator by Palamara et al. (2012). This transformation can be roughly summarized as to say that $N_\alpha \propto \frac{1}{\bar{x}_{\alpha,\alpha}}$ where N_α is the effective population size in deme α and $\bar{x}_{\alpha,\alpha}$ is the average pairwise PSC sharing between individuals in deme α. (b) Similar to Ralph and Coop (2013), for each focal population (marked with an x), we plot the normalized average pairwise sharing between that population and all others (normalized by the average sharing within the focal population), i.e. if α is the focal population, we show $\frac{\bar{x}_{\alpha,\beta}}{\bar{x}_{\alpha,\alpha}}$ for each other country β.
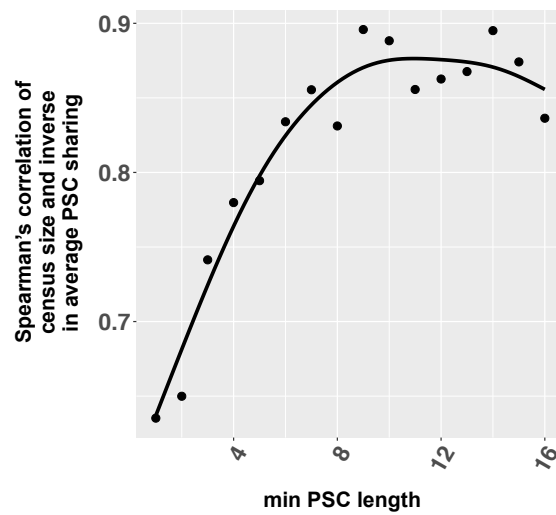
Figure S5: **The correlation between census size and inverse average PSC sharing as a function of minimum PSC length considered.** We use census size compiled from the The World Bank (2016) and National Records of Scotland (2011). The smooth black curve denotes the loess fit. Longer PSC segments correlate more strongly with census size than shorter PSC segments
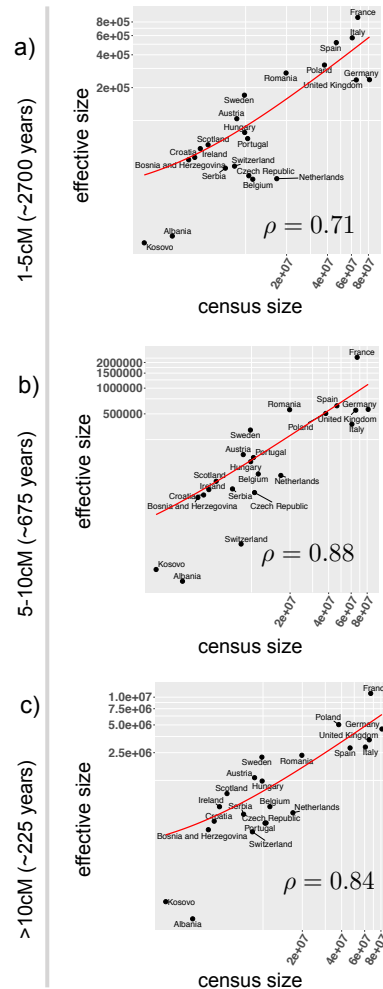
Figure S6: **Census size versus MAPS estimated population sizes.** Using the MAPS output, we estimate a total size per population by summing the estimated deme-level sizes across the area of each respective country (whether's a deme's location falls within a country was determined by querying The GeoNames Geographical Database). Finally, we plot the results on a log10 scale for different length scales (a) 1-5cM, (b) 5-10cM, and (c) >10cM. The red curve denotes the linear fit on the absolute scale. Note Kosovo and Albania as candidate outliers possibility because of cryptic relatedness artificially decreasing population sizes.
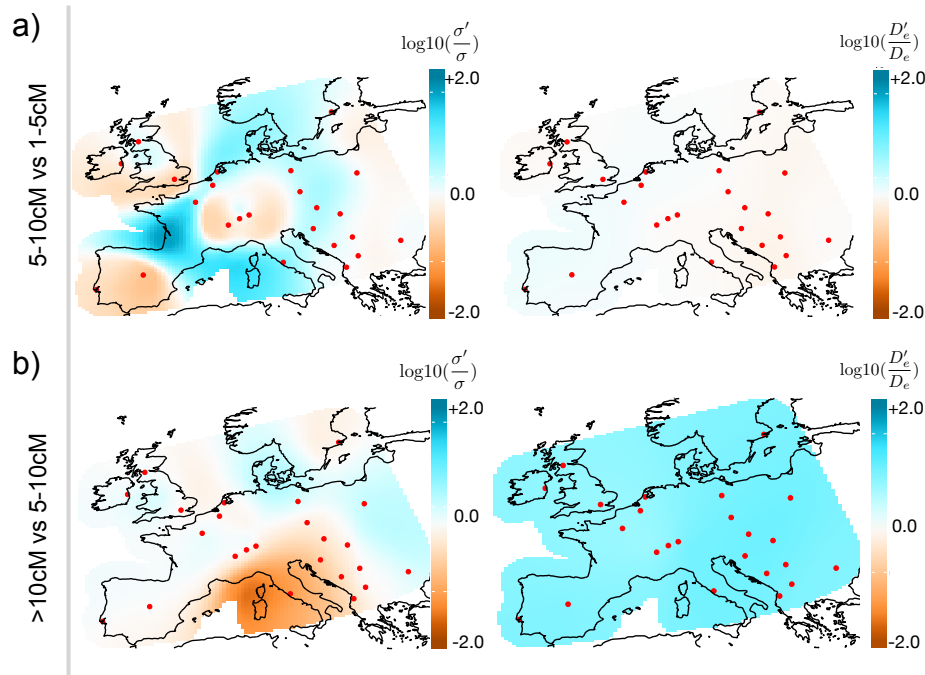
Figure S7: **Plots of estimated average log10 differences in demographic parameters between adjacent time scales**. (a) We plot estimates of $E[\log10(\frac{\sigma'}{\sigma})]$ and $E[\log10(\frac{D_e'}{D_e})]$ across the spatial habitat where σ' (D_e') denotes the dispersal rates (population densities) in the 5-10cM length bin and σ (D_e) denotes the dispersal rates (population densities) in the 1-5cM length bin. (b) The results here are similarly plotted as above, however, the adjacent length scales are given by: 5-10cM and >10cM. The log10 differences are estimated in such a way so that the mean log10 difference is shrunk to zero. For example, for estimating dispersal in 5-10cM, we assume $\log10(\sigma') = E[\log10(\sigma)] + \epsilon$ where $E[\log10(\sigma)]$ is estimated using PSC segments 1-5cM and $\epsilon \sim N(0, \omega^2)$ is estimated from PSC segments 5-10cM. Consequently, the log ratio between dispersal rates from the two lengths bins is constructed to have mean zero *apriori* (i.e. $E[\log10(\frac{\sigma'}{\sigma})] = 0$).
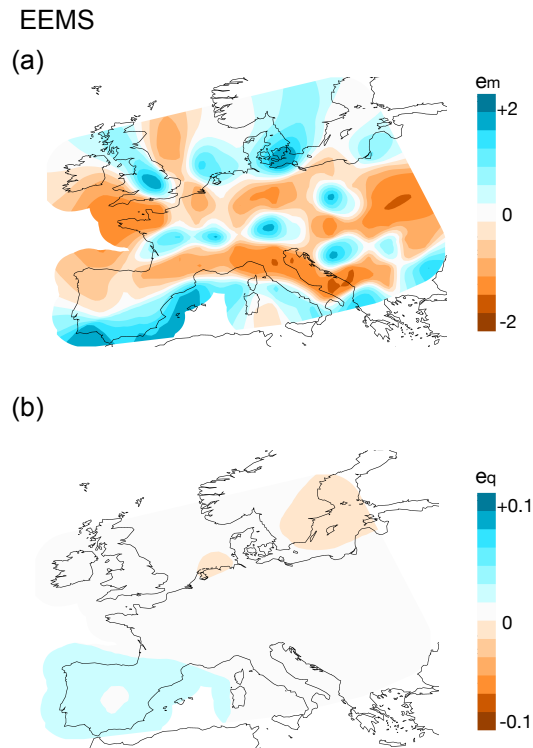
Figure S8: **EEMS applied to the POPRES dataset.** We apply EEMS to the same set of individuals as used in Figure 4 (see Methods). (a) The effective migration rates (b) The effective diversity rates. Here, we ran EEMS with 200 demes (as in Figure 4) with default parameters and averaged over 10 independent replicate chains. Each chain ran with 50e6 MCMC iterations, 25e6 set as burn-in, and we thinned every 5000 iterations.
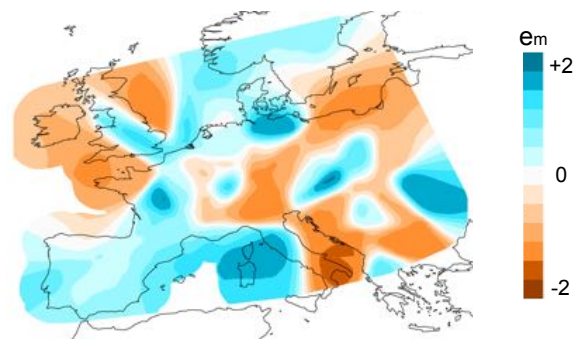
(a)
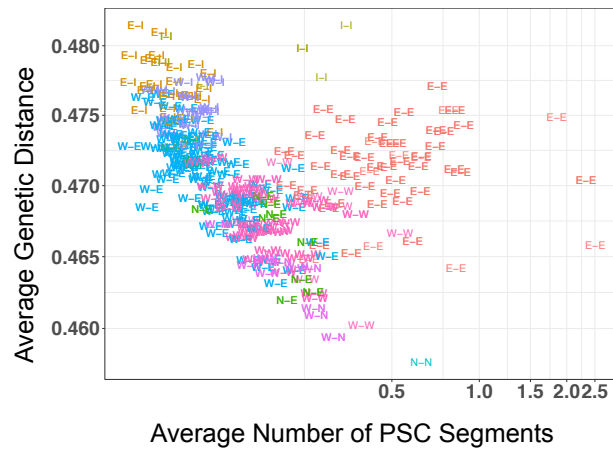
(b)

Figure S9: **Genetic distance vs PSC sharing** (a) The averaged genetic distance (as used in EEMS) is plotted against the average number of PSC segments (> 1cM) for each pair of populations. Each point denotes a pair, the symbols represent groupings from Ralph and Coop (2013) (W Western Europe, S Southern Europe, and E Eastern Europe), and the colors represent the pair of regions. We see a negative correlation between the two summary statistics (Pearson's ρ = -0.38, p-value = 7e-11), with the largest deviations occurring in comparisons between Eastern European populations. (b) EEMS results on PSC data transformed to a distance matrix. First, we encoded the PSC sharing statistics into a similarity matrix S such that $S_{i,j}$ is the number of shared PSC segments between samples i and j and $S_{i,i}$ is the maximum number of shared segments in the dataset (which we denote as c) to ensure S is a similarity matrix. Next, we transformed S to a genetic distance matrix D such that $D = c11^T - S + E$ where $E \approx 0$ is a random genetic distance matrix of normal vectors with mean 0 and standard deviation of 0.01 added to ensure D is full rank. Finally, we applied EEMS to the distance matrix D. Though this procedure is heuristic, we see shared features between this surface and the MAPS dispersal surface shown in Figure 4.

References

Soheil Baharian, Maxime Barakatt, Christopher R Gignoux, Suyash Shringarpure, Jacob Errington, William J Blot, Carlos D Bustamante, Eimear E Kenny, Scott M Williams, Melinda C Aldrich, et al. The Great Migration and African-American Genomic Diversity. *PLoS Genetics*, 12(5):e1006059, 2016.

Melanie Bahlo and Robert C Griffiths. Coalescence Time for Two Genes from a Subdivided Population. *Journal of Mathematical Biology*, 43(5):397–410, 2001.

Guido Barbujani, Neal L Oden, and Robert R Sokal. Detecting regions of abrupt change in maps of biological variables. *Systematic Zoology*, 38(4):376–389, 1989.

Nick H Barton, Frantz Depaulis, and Alison M Etheridge. Neutral Evolution in Spatially Continuous Populations. *Theoretical Population Biology*, 61(1):31–48, 2002.

Michael GB Blum, Christophe Damerval, Stephanie Manel, and Olivier François. Brownian Models and Coalescent Structures. *Theoretical Population Biology*, 65(3):249–261, 2004.

Gideon Bradburd, Graham Coop, and Peter Ralph. Inferring Continuous and Discrete Population Genetic Structure across Space. *bioRxiv*, page 189688, 2017.

Gideon S Bradburd, Peter L Ralph, and Graham M Coop. A Spatial Framework for Understanding Population Structure and Admixture. *PLoS Genet*, 12(1):e1005703, 2016.

Brian L Browning and Sharon R Browning. A Fast, Powerful Method for Detecting Identity-by-Descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

Brian L Browning and Sharon R Browning. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics*, 194(2):459–471, 2013.

Kevin Caye, Flora Jay, Olivier Michel, and Olivier Francois. Fast Inference of Individual Admixture Coefficients Using Geographic Data. *bioRxiv*, page 080291, 2016.

Gary K Chen, Paul Marjoram, and Jeffrey D Wall. Fast and flexible simulation of DNA sequence data. *Genome Research*, 19(1):136–142, 2009.

Charleston WK Chiang, Joseph H Marcus, Carlo Sidore, Hussein Al-Asadi, Magdalena Zoledziewska, Maristella Pitzalis, Fabio Busonero, Andrea Maschio, Giorgio Pistis, Maristella Steri, et al. Population History of the Sardinian People Inferred from Whole-Genome Sequencing. *bioRxiv*, page 092148, 2016.

Joel E Cohen. Population Growth and Earth's Human Carrying Capacity. *Science*, 269 (5222):341–346, 1995.

Peter J Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732, 1995.

Gilles Guillot, Frédéric Mortier, and Arnaud Estoup. GENELAND: A Computer Package for Landscape Genetics. *Molecular Ecology Resources*, 5(3):712–715, 2005.

Gilles Guillot, Raphael Leblois, Aurelie Coulon, and Alain C Frantz. Statistical Methods in Spatial Genetics. *Molecular Ecology*, 18(23):4734–4756, 2009.

Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole Population, Genome-Wide Mapping of Hidden Relatedness. *Genome Research*, 19(2):318–326, 2009.

Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, USA, 2004.

Joanna Kaplanis, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, et al. Quantitative analysis of population-scale family trees with millions of relatives. *Science*, (early online): eaam9309, 2018.

Oscar Lao, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A Bindoff, David Comas, et al. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16):1241–1248, 2008.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1):e1002453, 2012.

Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Daniel J Lawson, et al. The Fine-Scale Genetic Structure of the British Population. *Nature*, 519(7543):309–314, 2015.

Heng Li and Richard Durbin. Inference of Human Population History from Individual Whole-Genome Sequences. *Nature*, 475(7357):493–496, 2011.

Stéphanie Manel, Michael K Schwartz, Gordon Luikart, and Pierre Taberlet. Landscape Genetics: Combining Landscape Ecology and Population Genetics. *Trends in Ecology & Evolution*, 18(4):189–197, 2003.

Brad H McRae. Isolation by Resistance. *Evolution*, 60(8):1551–1561, 2006.

National Records of Scotland. Scotland's 2011 census, 2011. census size retrieved from, http://www.scotlandscensus.gov.uk/statistical-bulletins.

Matthew R Nelson, Katarzyna Bryc, Karen S King, Amit Indap, Adam R Boyko, John Novembre, Linda P Briley, Yuka Maruyama, Dawn M Waterworth, Gérard Waeber, et al. The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. *The American Journal of Human Genetics*, 83(3): 347–358, 2008.

John Novembre and Montgomery Slatkin. Likelihood-Based Inference in Isolation-by-Distance Models Using the Spatial Distribution of Low-Frequency Alleles. *Evolution*, 63(11):2914–2925, 2009.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes Mirror Geography within Europe. *Nature*, 456(7218):98–101, 2008.

Pier Francesco Palamara and Itsik Peer. Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188, 2013.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Peer. Length Distributions of Identity by Descent Reveal Fine-scale Demographic History. *The American Journal of Human Genetics*, 91(5):809–822, 2012.

Nick Patterson, Alkes L Price, and David Reich. Population Structure and Eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.

Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing Spatial Population Structure with Estimated Effective Migration Surfaces. *Nat Genet*, 48(1):94–100, Jan 2016. doi: 10.1038/ng.3464.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000.

Peter Ralph and Graham Coop. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol*, 11(5):e1001555, 2013.

A Rényi. On the Central Limit Theorem for the Sum of a Random Number of Independent Random Variables. *Acta Mathematica Hungarica*, 11(1-2):97–102, 1960.

Harald Ringbauer, Graham Coop, and Nicholas H Barton. Inferring Recent Demography from Isolation-By-Distance of Long Shared Sequence Blocks. *Genetics*, 205(3):1335–1351, 2017.

JJ Robledo-Arnuncio and Rousset. Isolation by Distance in a Continuous Population under Stochastic Demographic Fluctuations. *Journal of Evolutionary Biology*, 23(1):53–71, 2010.

Noah A Rosenberg, Saurabh Mahajan, Sohini Ramachandran, Chengfeng Zhao, Jonathan K Pritchard, and Marcus W Feldman. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet*, 1(6):e70, 2005.

François Rousset. *Genetic Structure and Selection in Subdivided Populations (MPB-40)*. Princeton University Press New Jersey, 2004.

Joshua G Schraiber and Joshua M Akey. Methods and Models for Unravelling Human Evolutionary history. *Nature Reviews Genetics*, 16(12):727, 2015.

Gernot Segelbacher, Samuel A Cushman, Bryan K Epperson, Marie-Josée Fortin, Olivier Francois, Olivier J Hardy, Rolf Holderegger, Pierre Taberlet, Lisette P Waits, and Stéphanie Manel. Applications of Landscape Genetics in Conservation Biology: Concepts and Challenges. *Conservation Genetics*, 11(2):375–385, 2010.

Matthew Stephens. Bayesian Analysis of Mixture Models with an Unknown Number of Components, an Alternative to Reversible Jump Methods. *Annals of Statistics*, pages 40–74, 2000.

The GeoNames Geographical Database. Geonames. Queried on July 2018, http://www.geonames.org/.

The World Bank. World development indicators, 2016. data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SP.POP.TOTL.

Monica G Turner, Robert H Gardner, Robert V O'neill, et al. *Landscape Ecology in Theory and Practice*. Springer-Verlag New York, 2001.

Jon F Wilkins. A Separation-of-Timescales Approach to the Coalescent in a Continuous Population. *Genetics*, 168(4):2227–2244, 2004.

William H Womble. Differential systematics. *Science*, 114(2961):315–322, 1951.

749 Sewall Wright. Isolation by Distance. *Genetics*, 28(2):114, 1943.

750 Andreas Zimmermann, Johanna Hilpert, and Karl Peter Wendt. Estimations of Population
751 Density for Selected Periods between the Neolithic and AD 1800. *Human Biology*, 81(3):
752 357–380, 2009.