PLOS GENETICS

RESEARCH ARTICLE

Estimating recent migration and population-size surfaces

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

1 Evolutionary Biology, University of Chicago, Chicago, Illinois, United States of America, **2** Department of Statistics, University of Chicago, Illinois, United States of America, **3** Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **4** Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America

* halasadi@uchicago.edu (HA-A); mstephens@uchicago.edu (MS); jnovembre@uchicago.edu (JN)

Citation: Al-Asadi H, Petkova D, Stephens M, Novembre J (2019) Estimating recent migration and population-size surfaces. PLoS Genet 15(1): e1007908. https://doi.org/10.1371/journal.pgen.1007908

Editor: Michael DeGiorgio, Pennsylvania State University, UNITED STATES

Received: August 8, 2018

Accepted: December 19, 2018

Published: January 14, 2019

Data Availability Statement: All called IBD segments can be obtained from https://github.com/petrelharp/euroibd.

Funding: This work was supported by National Institute of Health funding [U01CA198933 to 543 H.A., M.S., and J.N], [HG002585 to M.S.]; and the National Science Foundation 544 Graduate Research Fellowship to H.A. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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 maps of population sizes and migration rates associated with 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 associated with 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 using a similar method that ignores 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.

Author summary

We introduce a novel statistical method to infer migration rates and population sizes across space in recent time periods. Our approach builds upon the previously developed EEMS method, which infers effective migration rates under a dense lattice. Similarly, we infer demographic parameters under a lattice and use a (Voronoi) prior to regularize parameters of the model. However, our method differs from EEMS in a few key respects. First, we use the coalescent model parameterized by migration rates and population sizes while EEMS uses a resistance model. As another key difference, our method uses haplotype data while EEMS uses the average genetic distance. A consequence of using haplotype data is that our method can separately estimate migration rates and population sizes,

Competing interests: The authors have declared that no competing interests exist.

which in essence is done by using a recombination rate map to calibrate the decay of haplotypes over time. An additional useful feature of haplotype data is that, by varying the lengths analyzed, we can infer demography associated with different recent time periods. We call our method MAPS for estimating Migration And Population-size Surfaces. To illustrate MAPS on real data, we analyze a genome-wide SNP dataset on 2224 individuals of European ancestry.

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 [1], 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 [2]. 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 [3], conservation genetics [4], evolution [5], and human genetics [6].

Several methods have been developed to reveal spatial heterogeneity in patterns of isolation by distance [7–14]. Some methods are based on explicitly modeling the spatial structure in the data [9, 10, 12–14]; others take non-parametric approaches [7, 8]; while other methods ignore the spatial configuration of the samples and rely on researchers to make a *post hoc* geographic interpretation of the results [15, 16]. 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 [17, 18]. For example, when individuals are similar along a long segment of their chromosomes, it suggests that these segments share a recent common ancestor [19]. 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 such as principal components analysis [20–22].

Here we develop a new method to infer spatial and temporal heterogeneity in population sizes and migration rates. The method takes as input geographic coordinates for a set of individuals sampled across a spatial landscape, and a matrix of their genetic similarities as measured by sharing of lPSC segments. It then infers two maps, one representing dispersal rates across the landscape, and another representing population density. Importantly, building these maps using different lengths of lPSC segments can help reveal changes in dispersal rates and population sizes loosely associated with different recent time periods.

Our method is based on a stepping-stone model where randomly-mating subpopulations are connected to neighboring subpopulations in a grid. Such models are parameterized by a

vector of population sizes ($\vec{N}$) and a sparse migration rate matrix (**M**). Stepping-stone models with a large number of demes can approximate spatially continuous population models [23, 24], and this can be exploited to produce maps of approximate dispersal rates and population density across continuous space.

Our method builds upon a method developed for estimating effective migration surfaces (EEMS) [12]. While EEMS infers local rates of effective migration relative to a global average, here we can explicitly infer absolute parameter values by leveraging lPSC segments and modeling the recombination process [$\vec{N}$ and **M** values in the stepping-stone model, and effective spatial density function $D_e(\vec{x})$ and dispersal rate function $\sigma(\vec{x})$ in the continuous limit]. We call this method MAPS, for inferring Migration And Population-size Surfaces.

We test MAPS on coalescent simulations and apply it to a European subset of 2,224 individuals from the POPRES data [25]. In simulations, we show that MAPS can infer both time-resolved migration barriers and population sizes across the habitat. In empirical data, we infer dispersal rates $\sigma(\vec{x})$ and population densities $D_e(\vec{x})$ loosely associated with different time periods in Europe.

Overview of MAPS

MAPS estimates demography using the number of Pairwise Shared Coalescence (PSC) segments of different lengths shared between individuals. We define a PSC segment between (haploid) individuals to be a genomic segment with a single coalescent time across its length (Fig 1A). Long PSC (lPSC) segments tend to have a recent coalescent time, and so manifest themselves in genotype data as unusually long regions of high pairwise similarity, which can be detected by various software packages [26–29]. Because lPSC segments typically reflect recent coalescent events, counts of lPSC segments are especially informative for recent population structure [19, 24, 30]. And partitioning lPSC segments into different lengths bins (e.g. 2-8cM, $\geq$8cM) can help focus inference on different (recent) temporal scales. However, we caution that the historical signal that gives rise to the number of segments of in a certain length bin (e.g. 2-8cM) to strongly overlap with that has given rise to a numbers of segments subsequent length bin (e.g. $\geq$8).

The MAPS model involves two components: i) a likelihood function (Eq (7)), 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 (Fig 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 [12], 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 [23] and [24] 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 an effective spatial diffusion parameter $\sigma(\vec{x})$, often referred to as the "root mean square dispersal distance", which can be interpreted

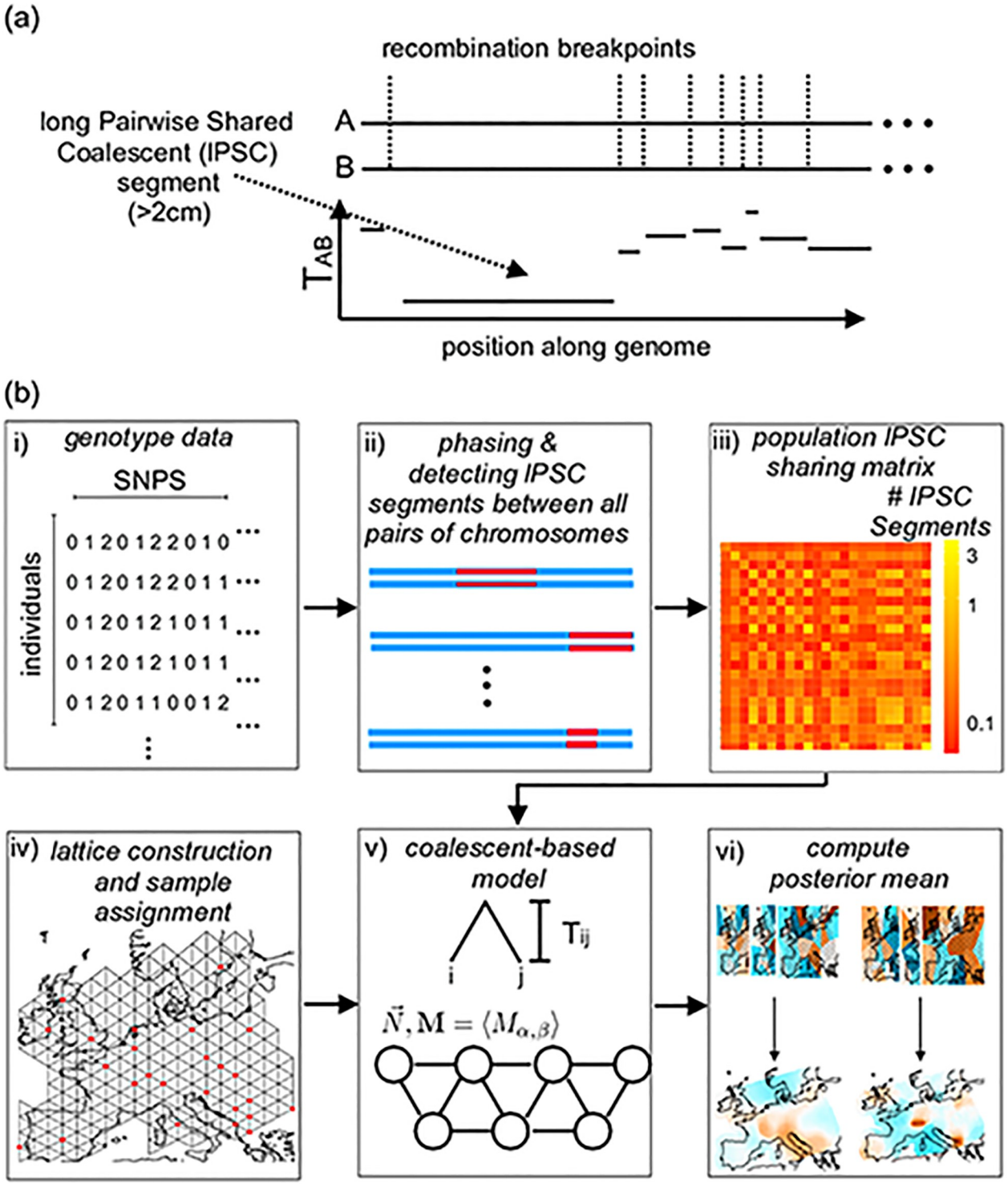

Fig 1. Schematic overview of MAPS. (a) Coalescent times between a pair of hapolotypes (A and B) will vary across the genome in discrete segments bordered by recombination breakpoints. On average, longer segments represent shorter pairwise coalescent times (T_{AB}) (b) Flow diagram of MAPS. i) We start 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, [27]); 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 Eq (6) in Methods.

https://doi.org/10.1371/journal.pgen.1007908.g001

roughly as the expected distance an individual disperses in one generation (Eq (18)); and we convert the population sizes ($\vec{N}$) to an "effective population density" $D_e(\vec{x})$, which can roughly be interpreted as the number of individuals per square kilometer (Eq (17)). These are deemed "effective" parameters because the spatial re-scaling assumes a simple approximation to the two dimensional coalescent process, see [23]. 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.

Differences from EEMS

Our MAPS approach is closely related to the EEMS method [12], 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 in 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 EEMS uses a resistance distance approximation [12, 31].

Results

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 [32] to simulate data under a coalescent stepping stone model and refinedIBD [27, 28] to identify lPSC segments. Alternatively, we could of inferred lPSC segments exactly using [32] or [33], however we found the error from refinedIBD to be negligible in our simulations. All simulations involved twenty demes, each containing 10,000 diploid individuals, and each exchanging migrants with their neighbor with a per lineage migration rate equal to 0.01 per generation. 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 S1 Appendix), however these are only the mean coalescent times and considerable variation exists in distribution of coalescent times. Results for other length bins also reflect the change in migration due to barrier (S1 & S2 Figs).

Migration rate inference. First, we simulated under a uniform (constant) migration surface with migration rate 0.01 (under a discrete model, Fig 2a), assumed to have stayed constant over time. In this case both EEMS and MAPS correctly infer uniform migration (Fig 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 generations ago, when a complete barrier to gene flow instantaneously arose (a "vicariance event", Fig 2b). In this setting EEMS again infers a uniform migration surface. This is because EEMS is based on pairwise genetic distances, which are negligibly influenced by the recent barrier. In contrast, by applying MAPS with different PSC segment lengths, we can see both the

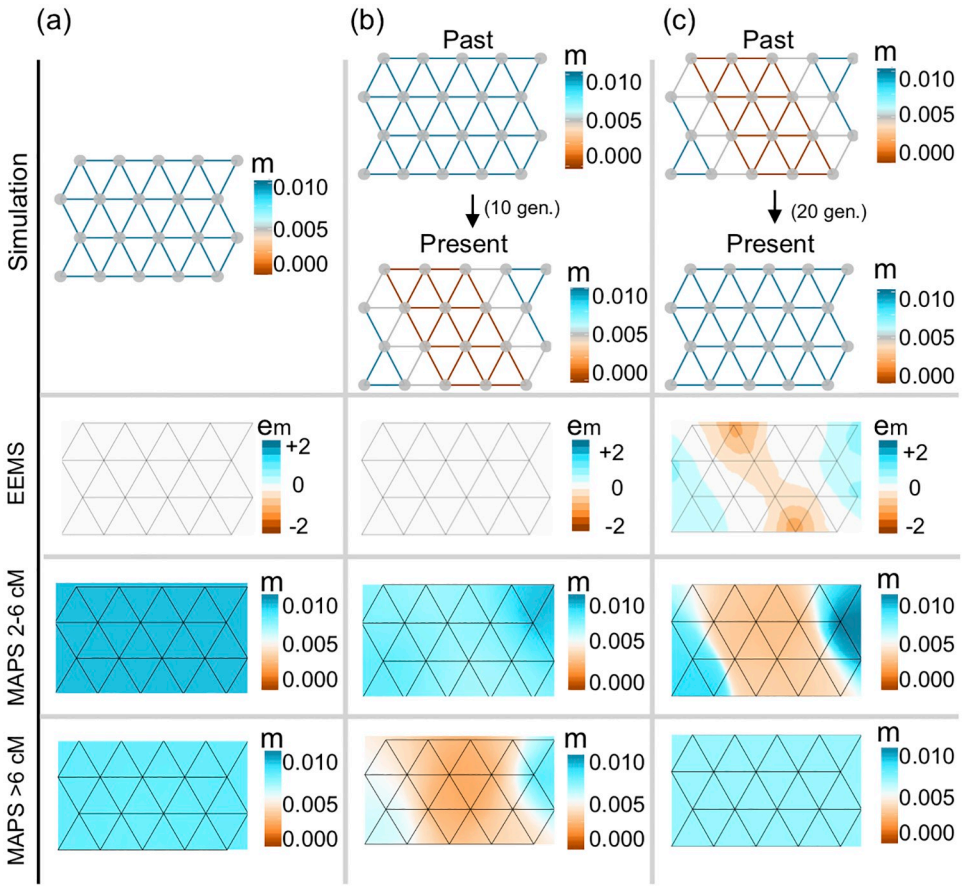

Fig 2. Simulations comparing migration rates inferred with MAPS against effective migration rates inferred with EEMS. (a) We simulated data under uniform migration rates equal to 0.01 and applied EEMS and MAPS using PSC segments in the range 2-6cM and ≥6cM. Like EEMS, MAPS correctly infers a uniform migration surface. Additionally, MAPS provides accurate estimates of the migration rates for both PSC segments 2-6cM (mean 0.01) and PSC segments ≥6cM (mean 0.0086). (b) We simulated a recent sudden migration barrier formed 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 (≥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 (≥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.

https://doi.org/10.1371/journal.pgen.1007908.g002

historically uniform migration surface (for segments 2-6cM) and the recent barrier (segments ≥6cM).

Next we consider a complementary time-varying scenario: an ancestral barrier disappeared 20 generations ago to allow uniform migration (Fig 2c). Here the EEMS results again reflect the longer-term processes, and a barrier is evident. And again, by applying MAPS with different PSC segment lengths, we can see different migration surfaces corresponding to different time scales, which are here reversed compared with the previous scenario: the historical barrier (for segments 2-6cM) and the recent uniform migration (segments ≥6cM).

Population size inference. As noted above, and as discussed in previous work, EEMS estimates an "effective" migration surface that reflects the combined effects of population sizes $\vec{N}$ and migration rates **M** [12]; consequently it cannot distinguish between variation in **M** and

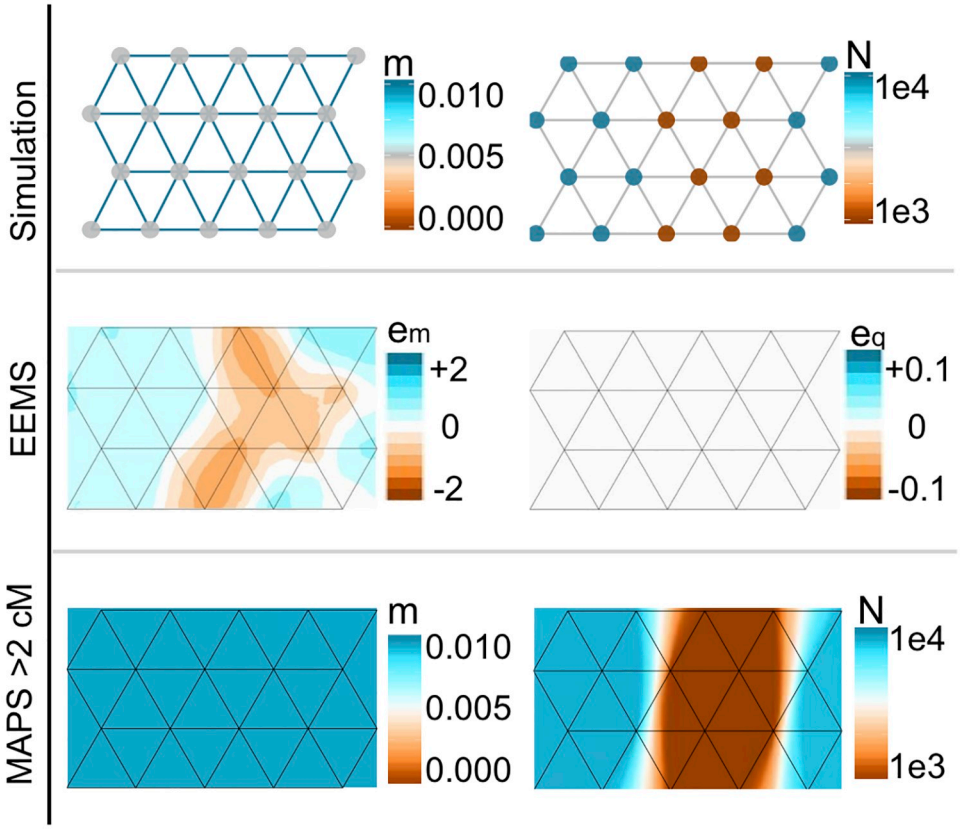

Fig 3. Simulations comparing population sizes inferred with MAPS and "diversity-rates" inferred with EEMS. We simulated uniform migration rates of 0.01 and a trough of low population sizes in the center of the habitat such that population sizes equal to 1,000 at the center and 10,000 otherwise. Under 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).

https://doi.org/10.1371/journal.pgen.1007908.g003

variation in $\vec{N}$. In contrast, MAPS has the potential to distinguish these two types of variation because MAPS utilizes the recombination rate map as an independent clock to calibrate demographic parameters.

To illustrate this difference we simulate data with a constant migration surface, and a population size surface that has a 10-fold "dip" in the middle of the habitat (deme size 1,000 vs 10,000; Fig 3). In a similar simulation, EEMS was shown to estimate an effective migration surface with an "effective barrier" in the middle, caused by the dip in population size [12]. As expected, we obtain a similar result for EEMS here. Furthermore, we examined the diversity surface inferred by EEMS [12], which reflects within-deme heterezygosity across space, please see S1 Appendix 1.4 for more on the diversity rates. We found the diversity surface to be approximately constant because within-deme deme hetereozygosity vary little in this simulation. In contrast, MAPS is able to separate the influence of migration and population sizes: the estimated migration surface is approximately constant (with mean migration rate equal to the true value 0.01) and the estimated population size surface shows a dip in the middle, with accurate estimates of deme sizes (mean 985 at the center and 9100 at the edges). Additional simulations with non-uniform migration rates reinforce these results; see S3 Fig.

Applying MAPS to the POPRES data

To illustrate MAPS on real data, we analyze a genome-wide SNP dataset on individuals of European ancestry [25]. Previous analyses of these data have shown the strong influence of geography on patterns of genetic similarity [20, 34, 35]. In particular [20] 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: 1–5cM, 5–10cM, 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, which roughly correspond to 2700 years, 675 years and 225 years if we assume 30 years per generation and a sufficiently large effective population size (see S1 Appendix). Here, we caution that these are only the *mean* coalescent times: other analyses have shown that distribution on coalescent times can have a very wide distribution and are strongly affected by the demographic history, especially in expanding populations [20].

We note that the accuracy of called PSC segments will vary across these bins: based on simulations in [20] 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, Eq 18) and population densities (population sizes scaled by grid area size, Eq 17) for each PSC length bin are shown in Fig 4.

Largely speaking, the spatial variation in inferred dispersal rates and population densities is remarkably consistent across the different time scales (Fig 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 [22]. These features are consistent with visual inspection of the raw lPSC sharing data (S4b Fig). 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 segment sharing (S4a Fig). The MAPS inferred country population sizes are also highly correlated with estimated current census population sizes from [36] and [37] (S5 Fig) which can be mainly attributed to the fact that lPSC segments are highly informative of current census population sizes (Fig 5).

The most notable variation among the estimated surfaces from different time scales is a dramatic increase in the mean estimated population density in the most recent time scale (Fig 4 and S6 Fig). Indeed, the estimated mean for the last time scale—1.4 individuals per square km—is 6-9 fold higher than those for the earlier time scales (0.16 and 0.22 respectively). This increase is consistent with the recent exponential growth of human population sizes [38]. The estimates themselves are lower than historical estimates of ≈1-30 individuals per square km based on archaeological data [39].

The dispersal surfaces show more minor changes between time periods (Fig 4 and S6 Fig). In particular, the estimated mean dispersal rates are relatively constant across time, being 73, 103 and 72 respectively (in units of km in a single generation, see S1 Appendix 1.2 on notes

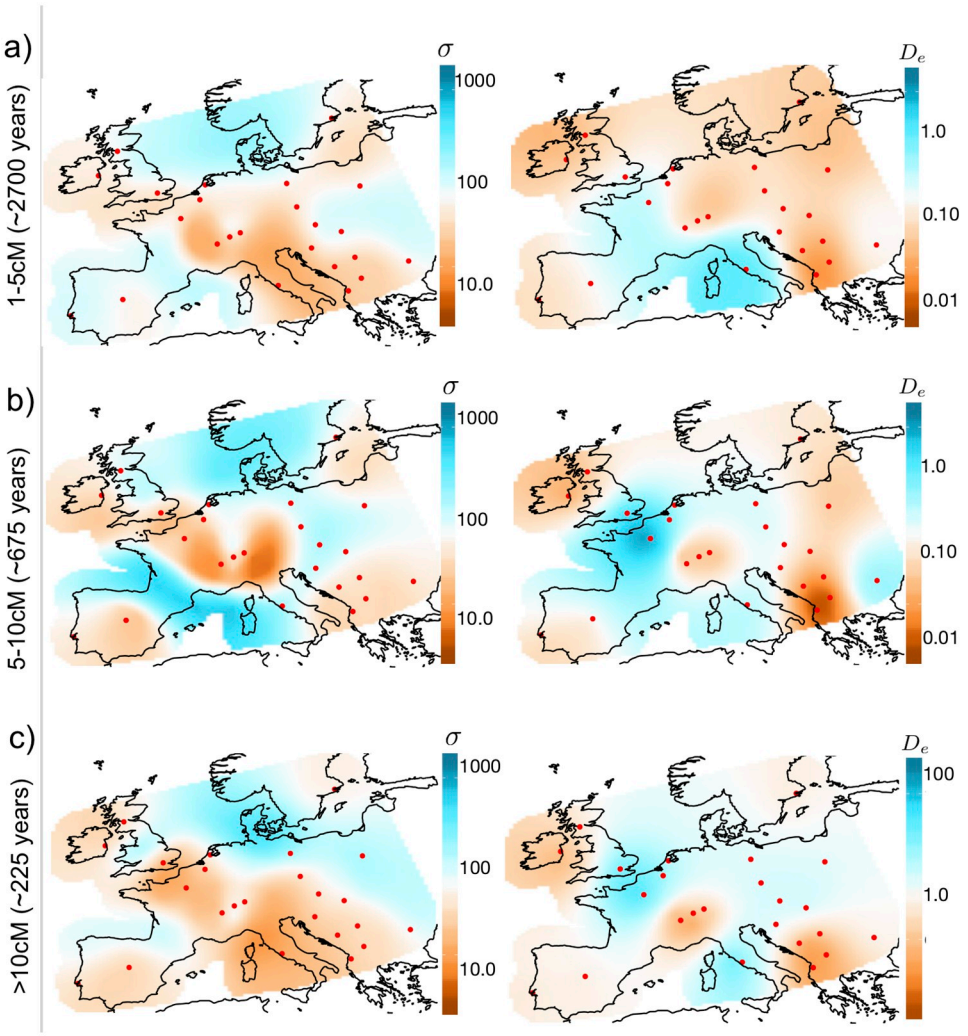

Fig 4. Inferred dispersal surfaces and population density surfaces over time for Europe. We apply MAPS to a European subset of POPRES [25] with 2,234 individuals and plot the inferred dispersal $\sigma(\vec{x})$ and population density $D_e(\vec{x})$ surfaces for PSC length bins (a) >1cM (b) 5-10cM and (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 Eqs (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.

https://doi.org/10.1371/journal.pgen.1007908.g004

about units). Our estimates are not too different from empirical estimates of 10-100 km in a single generation from [40] using pedigrees of individuals living between 1650 and 1950 AD. Although, our estimates seem to be consistently higher before the year 1800. We do note the lower estimated dispersal rates between Portugal and Spain compared to the rest of Europe in the analyses of longer PSC segments (5-10 and > 10cM), and the higher estimated dispersal rates through the Baltic Sea (> 10cM segments), possibly reflecting changing gene flow in these regions in recent history.

Comparison to Ringbauer et al. 2017. A previous study also estimate a mean dispersal rate and population density from the Eastern European subset of the data analyzed here [30]. Their estimates are based on PSC segments > 4cM, which is most comparable with our

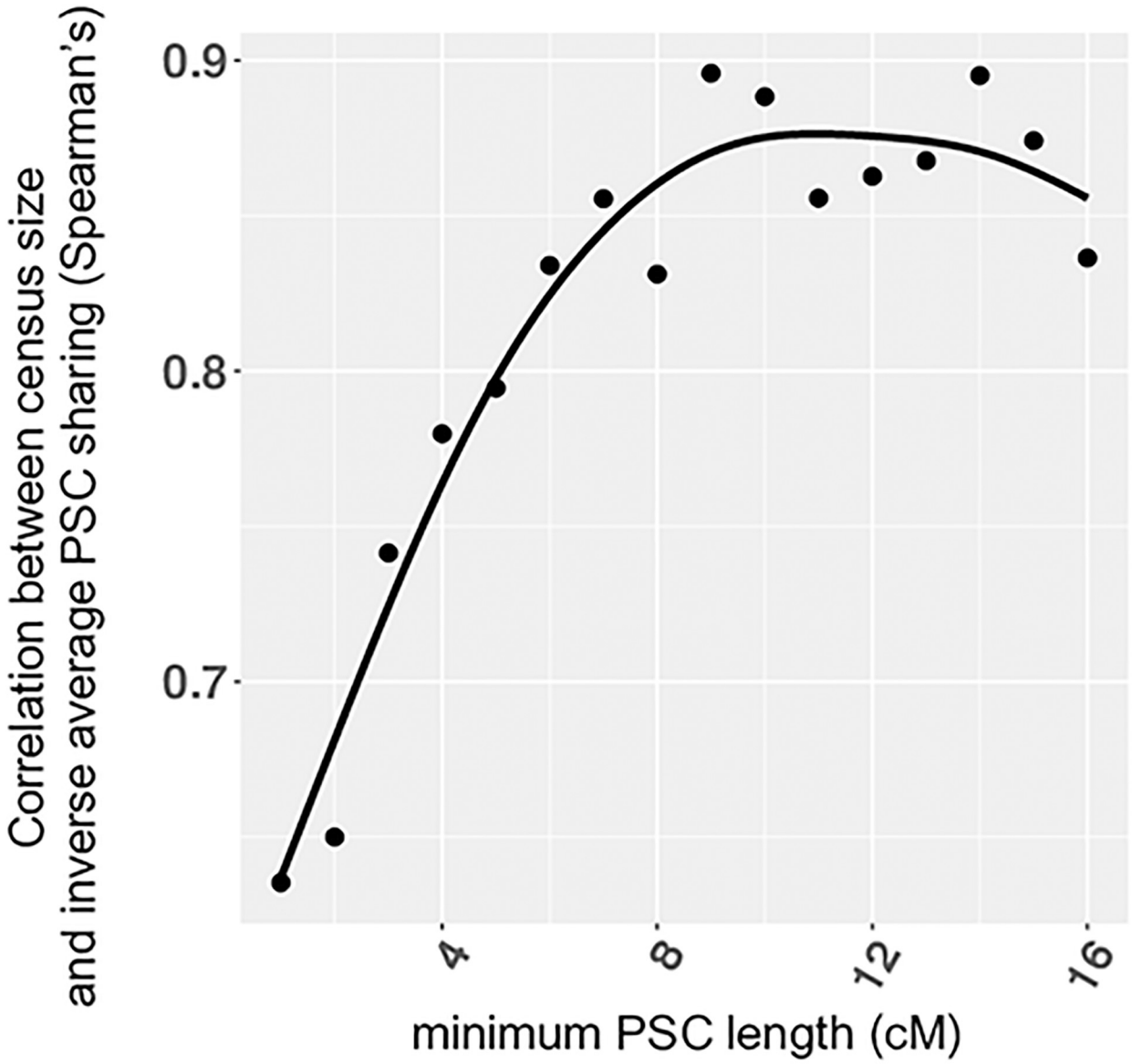

Fig 5. The correlation between census size and inverse average PSC sharing as a function of minimum PSC length considered. We computed the correlation coefficient (Spearman's) between census size and one over the average PSC sharing. We use census size compiled from the [36] and [37]. The smooth black curve denotes the loess fit. Longer PSC segments correlate more strongly with census size than shorter PSC segments.

https://doi.org/10.1371/journal.pgen.1007908.g005

analysis of 5-10cM. Unlike our analysis, their estimates are based on a spatially homogeneous model. To compare with their estimates we computed the mean of the estimated dispersal rate and population densities in Eastern Europe (but based on an analysis of the full data). For the dispersal rate this yields an estimate of 88 km in a single generation, which is consistent with the range of 50-100 given by [30]. For the population density, it yields an estimate of 0.10 individuals per square km, which is somewhat higher than the estimate of 0.05 obtained under a comparable (time-homogeneous) population model in [30]. Possibly our higher estimate partly reflects the influence of our spatial modeling approach, which will tend to shift the estimate for Eastern Europe toward the estimated mean across all of Europe (which is 0.22). In addition, the difference in length thresholds ($>$ 4cM versus 5-10cM) may also be contributing;

if segments in the analysis from [30] are on average shorter and hence older, one would expect lower density estimates, based on our results that suggest lower densities in the past (Fig 4).

Comparison with EEMS. The EEMS results for these data (S7 Fig) show non-trivial differences with the MAPS results (Fig 4a). Two potential causes are: i) differences in the summary data used (PSC segment sharing vs genetic distances) and hence sensitivity to different timescales; and ii) differences in the underlying models (e.g. composite Poisson likelihood vs Wishart likelihood, and different parameterizations/approximations to the coalescent model; see Discussion). To evaluate the impact of i) we compared the PSC segment sharing and genetic distances, and found their correlation to be only modest (Pearson's ρ = -0.38), with the most notable deviation for comparisons between countries in Eastern Europe (S8a Fig). Furthermore, most of this correlation is due to geographic distance: after controlling for geographic distance the correlation is only -0.18, which may be a more relevant metric because inferred spatial heterogeneity in gene flow (barriers and corridors) is driven by departures from simple isolation by distance.

To better assess the impact of ii) we applied EEMS on a distance matrix constructed to have the same similarity patterns as the PSC segment sharing matrix input to MAPS (1–5cM length bin). The resulting EEMS surface is more similar to the corresponding MAPS dispersal surface (S8b Fig vs Fig 4a), but there remain substantial differences. This investigation confirms what we expected *a priori*—the two surfaces should be different because the underlying models and inferred parameters of MAPS and EEMS are different. As noted before, EEMS infers the "effective migration rate" which reflects the effects of both the migration rates and population sizes, while MAPS infers them separately.

Discussion

We developed a method (MAPS) for inferring migration rates and population sizes across space and time periods from geo-referenced samples. Our method builds upon a previous method developed for estimating effective migration surfaces (EEMS) [12]. However there are several differences between MAPS and EEMS. Most fundamentally, MAPS draws inferences from observed levels of PSC sharing between samples, whereas EEMS draws inferences from the genetic distance. These two data summaries capture different information about the coalescent distributions: in essence, PSC sharing captures the frequency of recent coalescent events, whereas genetic distance captures the mean coalescent time. Consequently MAPS inferences largely reflect the recent past (mean coalescent time $\lessapprox$ 2,250 years for PSC segments $>$ 2cM), whereas EEMS inferences reflect demographic history on a longer timescale across which pairwise coalescence occurs (99% of events $>$ 6000 years old, assuming diploid N_e of 10,000 for humans, exponential coalescent time distribution).

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

population sizes. (The EEMS likelihood instead uses a "diversity rate" e_q, which reflects within-deme heterozygosity but is not explicitly a population size parameter.)

An additional useful feature of PSC segments is that, by varying the lengths analyzed, one can infer parameter values across associated with different time periods. For example, our simulations show how by contrasting shorter and longer PSC segments, the method can reveal different gene flow patterns in scenarios with recent changes (see Figs 2 and 3). Further support comes from our empirical analysis of the POPRES data-set, where we found population sizes inferred from longer PSC segments to be more correlated with census sizes than sizes inferred from shorter segments (e.g. Spearman's $\rho = 0.71$ for 1–5cM and $\rho = 0.84$ for > 10cM; see Fig 5 and S5 Fig). Also, not surprisingly, PSC segments greatly outperform using heterozygosity as an indicator of census population size (the Spearman's correlation coefficient between heterozygosity and census size was insignificant, p-value = 0.25).

Our estimates of dispersal distances and population density from the POPRES data are among the first such estimates using a spatial model for Europe (though see [30]). The features observed in the dispersal and population density surfaces are in principle discernible by careful inspection of the numbers of shared PSC segments between pairs of countries (e.g. using average pairwise numbers of shared segments, S4b Fig, as in [20]). For example, high connectivity across the North Sea is reflected in the raw PSC calls: samples from the British Isles share a relatively high number of PSC segments with those from Sweden (S4b Fig). Also the low estimated dispersal between Switzerland and Italy is consistent with Swiss samples sharing relatively few PSC segments with Italians given their close proximity (S4b Fig). However, identifying interesting patterns directly from the PSC segment sharing data is not straightforward, and one goal of MAPS (and EEMS) is to produce visualizations that point to patterns in the data that suggest deviations from simple isolation by distance.

The inferred population size surfaces for the POPRES data show a general increase in sizes through time, with small fluctuations across geography; In our results, the smallest inferred population sizes are in the Balkans and Eastern Europe more generally. This is in agreement with the signal seen previously [20]; however, taken at face value, our results suggest that high PSC sharing in these regions may be due more to consistently low population densities than to historical expansions (such as the Slavic or Hunnic expansions).

Although consistent with previous results, our estimates of dispersal and population sizes do not exactly agree with empirical estimates. For example, our estimates of population sizes are consistently lower than the census sizes (S5 Fig). This is to be expected for several reasons. First, census sizes include non-breeding individuals (juvenile and post-reproductive age) that do not impact the formation of PSC segments. Second, MAPS is fitting a single population size per location, and in a growing population the best fit population size will be an under-estimate of contemporary size. Third, in a wide class of population genetic models, the effective size, even among reproductive age individuals, is lower than the census size because of factors that inflate the variance in offspring number. Fourth, some discrepancy is expected simply because the stepping-stone population genetic model used here is only a coarse approximation to the complex spatial dynamics of human populations. Finally, there is probably cryptic relatedness in the POPRES samples which can decrease population size estimates.

Here, as in EEMS, we use a discrete stepping-stone model to approximate a process that might be more naturally modelled as continuously varying in space [12]. Recent work exploits continuous models to estimate dispersal and population density parameters from sharing of lPSC segments [24, 30]. However, these methods assume that dispersal and population density are constant across space: extending them to allow these parameters to vary across space could be an interesting avenue for future work.

Here, we infer demography given a PSC length bin. These PSC length bins correspond to very approximate time periods, and we report the mean age of the segment in the specified time period to give an idea of the approximate time period under an assumption of a large effective population size (see Lemma 5.3 in S1 Appendix). However, as mentioned previously, the variance in the distribution of ages can be very large. A major advancement would be to infer demography explicitly as a function of time. In principle, our method allows for inference of demography across time by treating PSC segments as roughly approximating independent across length bins conditional on the demography, see S1 Appendix. However, this requires fitting multiple migration/population surfaces and is computationally unfeasible with our current MCMC routine. Other computational techniques (e.g. Variational Bayes or fast optimization of the likelihood) might make this goal possible.

Methods

MAPS configuration

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

Inferring PSC segments from the data

Our pipeline to call PSC segments for simulations can be found here: https://github.com/halasadi/ibd_data_pipeline. We follow the recommendations of [27, 28] and [20] by running BEAGLE multiple times and merging shorter segments.

For the empirical data analysis, we use the PSC segments ("IBD") calls from [20], which can be found here: https://github.com/petrelharp/euroibd. The calls from [20] were obtained by running fastIBD (implemetned in BEAGLE [27]) and applying custom post-processing steps derived by simulation. We further applied a filter to retain countries with at least 5 sampled individuals, and removed Russian and Greek individuals to restrict the geographic region to a smaller spatial scale.

Model

MAPS assumes a population genetic model consisting of triangular grid of d demes (or populations) with symmetric migration. The density of the grid is pre-specified by the user with the consideration that the computational complexity is $O(d^3)$. We use Bayesian inference to estimate the MAPS parameters: the migration rates and coalescent rates M and $\underline{q}$ respectively. Its key components are the likelihood, which measures how well the parameters explain the observed data, and the prior, which captures the expectation that M and $\underline{q}$ have some spatial structure (in particular, the idea that nearby edges will tend to have similar migration rates and nearby demes have similar coalescent rates).

MAPS estimates the posterior distribution of $\Theta = M, \underline{q}$ given the data. The data used for MAPS consists of a similarity matrix $X^R = \{X^R_{i,j}\}$ which denotes the number of PSC segments in a range $R = [\mu, v]$ base-pairs shared between pairs of haploid individuals $(i, j) \in \{1, \cdots, n\} \times \{1, \cdots, n\}$ where n is the number of (haploid) individuals. Furthermore, a recombination rate map is required as input for MAPS. The likelihood is a function of the expected value of $X^R_{i,j}$ ($E[X^R_{i,j}]$). Below we describe the computation of $E[X^R_{i,j}]$ and the other key components of the

likelihood. Finally, we briefly describe the prior used and an MCMC 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. We define,

$$\lambda^{\Theta}_{\alpha,\beta} = E[X^R_{i,j}|\Theta], \tag{1}$$

which implicitly depends on R. For the marginal distribution, we can assume

$$X^R_{i,j}|\Theta \sim \text{Pois}(\lambda^{\Theta}_{\alpha,\beta}|\Theta). \tag{2}$$

See [41] for a rigorous study of the Poisson assumption. One option for computing the joint distribution of the data is to assume independence between pairs of individuals (i, j) as done previously [19, 20, 30, 42]. This assumption leads to the log-likelihood,

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

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

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

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, also studied in previous work [43]. 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. For instance, if $\alpha = \beta$, we sample n_α individuals with replacement and compute the average between all $\binom{n_\alpha}{2}$ comparisons, and repeat this process many times. Using this boostrapping procedure allows us to better adjust for the correlations between pairs of individuals for within and between-deme comparisons. The loglikelihood adjusted for correlations is given by,

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

Computing the expectation of $X^R_{i,j}|\Theta$. Next, we derive expressions to compute the expectation of the number of PSC segments of length greater than μ ($X^{R=[\mu,\infty)}_{i,j}$) conditional on the demography Θ. From results in [19], we show in S1 Appendix that

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

where G denotes the length of the genome (in base-pairs), L denotes the random length (in base-pairs) of the PSC segment between i and j containing a pre-specified position in the genome (base b say), and f_L is its probability density. Intuitively, $Gf_L(l|\Theta)$ is the expected number of base-pairs that lie in PSC segments of length l, making $\frac{Gf_L(l|\Theta)}{l}$ the expected number of PSC segments of length l. Integrating the latter quantity from μ to ∞ gives the desired result. Note, that (8) is only an approximation because we have implicitly assumed that the genome is infinitely long as in [19, 42]. A more exact formula will take account of the finite length of the genome, as in equation (6) in [20] which suggests that (8) will be off by an amount proportional to $\frac{\mu}{G}$. This correction for finite length will become more important for longer segments. For example, for segments of length 10cM, (8) is expected to be approximately 10% off.

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}:

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

$$= G \int_{\mu}^{\infty} \int_{0}^{\infty} f_{L,T_{i,j}}(l,t|\Theta)/l\, dt\, dl \tag{10}$$

$$= G \int_{0}^{\infty} f_{T_{i,j}}(t|\Theta) \int_{\mu}^{\infty} f_L(l|t)/l\, dl\, dt, \tag{11}$$

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 simplification 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 (or a Gamma Distribution with shape parameter fixed to two) [19, 42, 44] 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

$$\int_{\mu}^{\infty} f_L(l|t)/l\, dl = 2rte^{-2tr\mu}, \tag{13}$$

leading to

$$E[X^{R=[\mu,\infty)}_{i,j}|\Theta] \approx G \int_{0}^{\infty} f_{T_{i,j}}(t|\Theta) 2rte^{-2tr\mu} dt. \tag{14}$$

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

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

where demes α, β denote the deme where lineages i and j are sampled from, $q_\kappa = \frac{1}{2N_\kappa}$ is the coalescent rate in deme κ, and $M = \langle m_{\alpha,\beta} \rangle$ is the migration rate matrix between all d demes such that $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, \ldots, d\}$. Please refer to S1 Appendix 1.1 for a derivation. 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).

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

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

As mentioned previously, the units of μ, ν are in base-pairs for clarity of presentation. However, we can work with units of centiMorgans (cM) as done in [19] by making the following the transformation: $\mu_{cM} = 100\mu r$. By making this substitution, our population-genetic model becomes identical to [19] under a single population size.

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 and β are equal to $m_{\alpha,\beta} = \frac{\mathcal{M}_{c_1} + \mathcal{M}_{c_2}}{2}$ 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 S1 Appendix 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 update give support to all the parameter space, this will define an irreducible Markov chain with the correct joint posterior distribution [45]. 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 S1 Appendix 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

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

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 (in km^2), d is the number of demes, and x corresponds to the spatial position of deme α. Intuitively, (17) implies that the density multiplied by the area equals population size, i.e. $D_e(x)\Delta A \approx N_\alpha$. Eq (17) is analogous to equation 7 in [24].

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

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

where Δx is the step size of the grid (km). The dispersal distance represents the distance traveled by an individual after one generation, and sometimes is called the "root mean square distance" or "dispersal rate" [23]. Please see S1 Appendix for the derivation of (18).

Supporting information

S1 Appendix. More detailed methods.
(PDF)

S1 Fig. 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 Fig 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.
(PDF)

S2 Fig. 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 Fig 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.
(PDF)

S3 Fig. 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.
(PDF)

S4 Fig. Visualizing normalized sharing of PSC segments that are 1-5cM. The color scheme is the same as used in [20] 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 an estimate of effective population size using an equation in Appendix B of [19]. The equation 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 [20], 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.
(PDF)

S5 Fig. 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 [46]). 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.
(PDF)

S6 Fig. Plots of estimated average log10 differences in demographic parameters between adjacent time scales. (a) We plot estimates of $E\left[\log10\left(\frac{\sigma'}{\sigma}\right)\right]$ 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. $E\left[\log10\left(\frac{\sigma'}{\sigma}\right)\right] = 0$).
(PDF)

S7 Fig. EEMS applied to the POPRES dataset. We apply EEMS to the same set of individuals as used in Fig 4 (see Methods). (a) The effective migration rates (b) The effective diversity rates. Here, we ran EEMS with 200 demes (as in Fig 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.
(PDF)

S8 Fig. 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 [20] (W Western Europe, S Southern Europe, I Italian & Iberian Peninsula, 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 Fig 4.
(PDF)

Acknowledgments

We thank Dick Hudson for helpful discussion on the identifiability of demographic parameters in coalescent models; Evan Koch, Ben Peter and the Novembre & Stephens Lab for comments on the paper and helpful discussion; Peter Carbonetto for computational support and helpful discussion; and Shai Carmi for discussion on dependency between pairwise sharing of IBD segments.

Author Contributions

Conceptualization: Hussein Al-Asadi, Matthew Stephens, John Novembre.

Data curation: Hussein Al-Asadi.

Formal analysis: Hussein Al-Asadi.

Funding acquisition: Hussein Al-Asadi, Matthew Stephens, John Novembre.

Investigation: Hussein Al-Asadi.

Methodology: Hussein Al-Asadi, Desislava Petkova, Matthew Stephens, John Novembre.

Project administration: Hussein Al-Asadi.

Resources: Hussein Al-Asadi, Matthew Stephens, John Novembre.

Software: Hussein Al-Asadi, Desislava Petkova.

Supervision: Hussein Al-Asadi, Matthew Stephens, John Novembre.

Validation: Hussein Al-Asadi.

Visualization: Hussein Al-Asadi, Matthew Stephens, John Novembre.

Writing – original draft: Hussein Al-Asadi, Matthew Stephens, John Novembre.

Writing – review & editing: Hussein Al-Asadi, Matthew Stephens, John Novembre.

References

1. Wright S. Isolation by Distance. Genetics. 1943; 28(2):114. PMID: 17247074
2. Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape Genetics: Combining Landscape Ecology and Population Genetics. Trends in Ecology & Evolution. 2003; 18(4):189–197. https://doi.org/10.1016/S0169-5347(03)00008-9
3. Turner MG, Gardner RH, O'neill RV, et al. Landscape Ecology in Theory and Practice. Springer-Verlag New York; 2001.
4. Segelbacher G, Cushman SA, Epperson BK, Fortin MJ, Francois O, Hardy OJ, et al. Applications of Landscape Genetics in Conservation Biology: Concepts and Challenges. Conservation Genetics. 2010; 11(2):375–385. https://doi.org/10.1007/s10592-009-0044-5
5. Rousset F. Genetic Structure and Selection in Subdivided Populations (MPB-40). Princeton University Press New Jersey; 2004.
6. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet. 2005; 1 (6):e70. https://doi.org/10.1371/journal.pgen.0010070 PMID: 16355252
7. Womble WH. Differential Systematics. Science. 1951; 114(2961):315–322. https://doi.org/10.1126/science.114.2961.315 PMID: 14883851
8. Barbujani G, Oden NL, Sokal RR. Detecting Regions of Abrupt Change in Maps of Biological Variables. Systematic Zoology. 1989; 38(4):376–389. https://doi.org/10.2307/2992403
9. Guillot G, Mortier F, Estoup A. GENELAND: A Computer Package for Landscape Genetics. Molecular Ecology Resources. 2005; 5(3):712–715.
10. Guillot G, Leblois R, Coulon A, Frantz AC. Statistical Methods in Spatial Genetics. Molecular Ecology. 2009; 18(23):4734–4756. https://doi.org/10.1111/j.1365-294X.2009.04410.x PMID: 19878454
11. Caye K, Jay F, Michel O, Francois O. Fast Inference of Individual Admixture Coefficients Using Geographic Data. bioRxiv. 2016; p. 080291.
12. Petkova D, Novembre J, Stephens M. Visualizing Spatial Population Structure with Estimated Effective Migration Surfaces. Nat Genet. 2016; 48(1):94–100. https://doi.org/10.1038/ng.3464 PMID: 26642242
13. Bradburd GS, Ralph PL, Coop GM. A Spatial Framework for Understanding Population Structure and Admixture. PLoS Genet. 2016; 12(1):e1005703. https://doi.org/10.1371/journal.pgen.1005703 PMID: 26771578
14. Bradburd G, Coop G, Ralph P. Inferring Continuous and Discrete Population Genetic Structure across Space. bioRxiv. 2017; p. 189688.
15. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure using Multilocus Genotype Data. Genetics. 2000; 155(2):945–959. PMID: 10835412
16. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006; 2(12): e190. https://doi.org/10.1371/journal.pgen.0020190 PMID: 17194218
17. Li H, Durbin R. Inference of Human Population History from Individual Whole-Genome Sequences. Nature. 2011; 475(7357):493–496. https://doi.org/10.1038/nature10231 PMID: 21753753
18. Schraiber JG, Akey JM. Methods and Models for Unravelling Human Evolutionary History. Nature Reviews Genetics. 2015; 16(12):727. https://doi.org/10.1038/nrg4005 PMID: 26553329

19. Palamara PF, Lencz T, Darvasi A, Pe'er I. Length Distributions of Identity by Descent Reveal Fine-scale Demographic History. The American Journal of Human Genetics. 2012; 91(5):809–822. https://doi.org/10.1016/j.ajhg.2012.08.030 PMID: 23103233

20. Ralph P, Coop G. The Geography of Recent Genetic Ancestry across Europe. PLoS Biol. 2013; 11(5): e1001555. https://doi.org/10.1371/journal.pbio.1001555 PMID: 23667324

21. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. PLoS Genetics. 2012; 8(1):e1002453. https://doi.org/10.1371/journal.pgen.1002453 PMID: 22291602

22. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The Fine-Scale Genetic Structure of the British Population. Nature. 2015; 519(7543):309–314. https://doi.org/10.1038/nature14230 PMID: 25788095

23. Barton NH, Depaulis F, Etheridge AM. Neutral Evolution in Spatially Continuous Populations. Theoretical Population Biology. 2002; 61(1):31–48. https://doi.org/10.1006/tpbi.2001.1557 PMID: 11895381

24. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al. The Great Migration and African-American Genomic Diversity. PLoS Genetics. 2016; 12(5):e1006059. https://doi.org/10.1371/journal.pgen.1006059 PMID: 27232753

25. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. The American Journal of Human Genetics. 2008; 83(3):347–358. https://doi.org/10.1016/j.ajhg.2008.08.005 PMID: 18760391

26. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, et al. Whole Population, Genome-Wide Mapping of Hidden Relatedness. Genome Research. 2009; 19(2):318–326. https://doi.org/10.1101/gr.081398.108 PMID: 18971310

27. Browning BL, Browning SR. A Fast, Powerful Method for Detecting Identity-by-Descent. The American Journal of Human Genetics. 2011; 88(2):173–182. https://doi.org/10.1016/j.ajhg.2011.01.010 PMID: 21310274

28. Browning BL, Browning SR. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics. 2013; 194(2):459–471. https://doi.org/10.1534/genetics.113.150029 PMID: 23535385

29. Chiang CW, Marcus JH, Sidore C, Al-Asadi H, Zoledziewska M, Pitzalis M, et al. Population History of the Sardinian People Inferred from Whole-Genome Sequencing. bioRxiv. 2016; p. 092148.

30. Ringbauer H, Coop G, Barton NH. Inferring Recent Demography from Isolation-By-Distance of Long Shared Sequence Blocks. Genetics. 2017; 205(3):1335–1351. https://doi.org/10.1534/genetics.116.196220 PMID: 28108588

31. McRae BH. Isolation by Resistance. Evolution. 2006; 60(8):1551–1561. https://doi.org/10.1111/j.0014-3820.2006.tb00500.x PMID: 17017056

32. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. Genome Research. 2009; 19(1):136–142. https://doi.org/10.1101/gr.083634.108 PMID: 19029539

33. Palamara PF. ARGON: Fast, Whole-genome Simulation of the Discrete Time Wright-Fisher Process. Bioinformatics. 2016; 32(19):3032–3034. https://doi.org/10.1093/bioinformatics/btw355 PMID: 27312410

34. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes Mirror Geography within Europe. Nature. 2008; 456(7218):98–101. https://doi.org/10.1038/nature07331 PMID: 18758442

35. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, et al. Correlation between Genetic and Geographic Structure in Europe. Current Biology. 2008; 18(16):1241–1248. https://doi.org/10.1016/j.cub.2008.07.049 PMID: 18691889

36. The World Bank. World Development Indicators; 2016.

37. National Records of Scotland. Scotland's 2011 Census; 2011.

38. Cohen JE. Population Growth and Earth's Human Carrying Capacity. Science. 1995; 269(5222):341–346. https://doi.org/10.1126/science.7618100 PMID: 7618100

39. Zimmermann A, Hilpert J, Wendt KP. Estimations of Population Density for Selected Periods between the Neolithic and AD 1800. Human Biology. 2009; 81(3):357–380. https://doi.org/10.3378/027.081.0313 PMID: 19943751

40. Kaplanis J, Gordon A, Shor T, Weissbrod O, Geiger D, Wahl M, et al. Quantitative analysis of population-scale family trees with millions of relatives. Science. 2018;(early online):eaam9309. https://doi.org/10.1126/science.aam9309

41. Carmi S, Wilton PR, Wakeley J, Pe'er I. A Renewal Theory Approach to IBD Sharing. Theoretical Population Biology. 2014; 97:35–48. https://doi.org/10.1016/j.tpb.2014.08.002 PMID: 25149691

42. Palamara PF, Pe'er I. Inference of historical migration rates via haplotype sharing. Bioinformatics. 2013; 29(13):i180–i188. https://doi.org/10.1093/bioinformatics/btt239 PMID: 23812983

43. Carmi S, Palamara PF, Vacic V, Lencz T, Darvasi A, Pe'er I. The Variance of Identity-By-Descent Sharing in the Wright-Fisher Model. Genetics. 2013; 193(3):911–928. https://doi.org/10.1534/genetics.112.147215 PMID: 23267057

44. Hein J, Schierup M, Wiuf C. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, USA; 2004.

45. Stephens M. Bayesian Analysis of Mixture Models with an Unknown Number of Components, an Alternative to Reversible Jump Methods. Annals of Statistics. 2000; p. 40–74. https://doi.org/10.1214/aos/1016120364

46. The GeoNames Geographical Database. GeoNames;.