Modelling the spatiotemporal spread of beneficial alleles using ancient genomes

Rasa Muktupavela[1,*], Martin Petr[1], Laure Ségurel[2], Thorfinn Korneliussen[1], John Novembre[3], Fernando Racimo[1]

[1]Lundbeck GeoGenetics Centre, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

[2]Laboratoire de Biométrie et Biologie Evolutive UMR5558, CNRS - Université Lyon 1, Université de Lyon, Villeurbanne, France

[3]University of Chicago, Department of Human Genetics, Chicago, IL, USA

[*]Corresponding author: rasa.muktupavela@gmail.com

September 9, 2021

Abstract

Ancient genome sequencing technologies now provide the opportunity to study natural selection in unprecedented detail. Rather than making inferences from indirect footprints left by selection in present-day genomes, we can directly observe whether a given allele was present or absent in a particular region of the world at almost any period of human history within the last 10,000 years. Methods for studying selection using ancient genomes often rely on partitioning individuals into discrete time periods or regions of the world. However, a complete understanding of natural selection requires more nuanced statistical methods which can explicitly model allele frequency changes in a continuum across space and time. Here we introduce a method for inferring the spread of a beneficial allele across a landscape using two-dimensional partial differential equations. Unlike previous approaches, our framework can handle time-stamped ancient samples, as well as genotype likelihoods and pseudohaploid sequences from low-coverage genomes. We apply the method to a panel of published ancient West Eurasian genomes, to produce dynamic maps showcasing the inferred spread of candidate beneficial alleles over time and space. We also provide estimates for the strength of selection and diffusion rate for each of these alleles. Finally, we highlight possible avenues of improvement for accurately tracing the spread of beneficial alleles in more complex scenarios.

Introduction

Understanding the dynamics of the spread of a beneficial allele through a population is one of the fundamental problems in population genetics (Ewens, 2012). We are often interested in knowing the location where an allele first arose and the way in which it spread through a population, but this is often unknown, particularly in natural, non-experimental settings where genetic sampling is scarce and uneven.

Patterns of genetic variation can be used to estimate how strongly natural selection has affected the trajectory of an allele and to fit the parameters of the selection process. The problem of estimating the age of a beneficial allele, for example, has yielded a rich methodological literature (Slatkin & Rannala, 2000), and recent methods have exploited fine-scale haplotype information to produce highly accurate age estimates (Mathieson & McVean, 2014; Platt *et al.*, 2019; Albers & McVean, 2020). In contrast, efforts to infer the geographic origins of beneficial mutations are scarcer. These include Novembre *et al.* (2005), who developed a maximum likelihood method to model the origin and spread of a beneficial mutation and applied it to the *CCR5*-Δ32 allele, which was, at the time, considered to have been under positive selection (Stephens *et al.*, 1998; Sabeti *et al.*, 2005; Novembre & Han, 2012). Similarly, Itan *et al.* (2009) developed an approximate Bayesian computation (ABC) approach using demic simulations, in order to find the geographic and temporal origins of a beneficial allele, based on present-day allele frequency patterns.

As ancient genome sequences become more readily available, they are increasingly being used to understand the process of natural selection (see reviews in Malaspinas *et al.* (2012); Dehasque *et al.* (2020)). However, few studies have used ancient genomes to fit spatial dynamic models of the spread of an allele over a landscape. Most spatiotemporal analyses which included ancient genomes have used descriptive modelling in order to learn the spatiotemporal covariance structure of allele frequencies (Segurel *et al.*, 2020) or hidden ancestry clusters (Racimo *et al.*, 2020b), and then used that structure to hindcast these patterns onto a continuous temporally-evolving landscape. In contrast to descriptive approaches, dynamic models have the power to infer interpretable parameters from genomic data and perhaps reveal the ultimate causes for these patterns (Wikle *et al.*, 2019).

Dynamic models can also contribute to ongoing debates about the past trajectories of phenotypically important loci. For example, the geographic origin of the rs4988235(T) allele—upstream of the *LCT* gene and associated with adult lactase persistence in most of Western Eurasia (Enattah *et al.*, 2002)—remains elusive, as is the way in which it spread (an extensive review can be found in Ségurel & Bon, 2017). The allele has been found in different populations, with frequencies ranging from 5% up to almost 100%, and its selection coefficient has been estimated to be among the highest in human populations (Bersaglieri *et al.*, 2004; Enattah *et al.*, 2008; Tishkoff *et al.*, 2007). However, the exact causes for its adaptive advantage are contested (Szpak *et al.*, 2019), and it has been suggested that the selection pressures acting on the allele may have been different in different parts of the continent (Gerbault *et al.*, 2009). Ancient DNA evidence shows that the allele was rare in Europe during the Neolithic (Burger *et al.*, 2007; Gamba *et al.*, 2014; Allentoft *et al.*, 2015; Mathieson *et al.*, 2015) and only became common in Northern Europe after the Iron Age, suggesting a rise in frequency during this period, perhaps mediated by gene flow from regions east of the Baltic where this allele was more common during the onset of the Bronze Age (Krüttli *et al.*, 2014; Margaryan *et al.*, 2020). Itan *et al.* (2009) deployed their ABC approach to model the spatial spread of the rs4988235(T) allele and estimated that it was first under selection among farmers around 7,500 years ago possibly between the central Balkans and central Europe. Others have postulated a steppe origin for the allele (Allentoft *et al.*, 2015), given that the rise in frequency appears to have occurred during and after the Bronze Age migration of steppe peoples into Western Eurasia (Haak *et al.*, 2015; Allentoft *et al.*, 2015). However, the allele is at low frequency in genomes of Bronze Age individuals associated with Corded Ware and Bell Beaker assemblages in Central Europe who have high steppe ancestry (Mathieson *et al.*, 2015; Margaryan *et al.*, 2020), complicating the story further (Ségurel & Bon, 2017).

The origins and spread dynamics of large-effect pigmentation-associated SNPs in ancient Eurasians have also been intensely studied (Ju & Mathieson, 2020). Major loci of large effect on skin, eye and hair pigmentation have been documented as having been under recent positive selection in Western Eurasian history (Voight *et al.*, 2006; Sabeti *et al.*, 2007; Pickrell *et al.*, 2009; Lao *et al.*, 2007; Mathieson *et al.*, 2015; Alonso *et al.*, 2008; Hudjashov *et al.*, 2013). These include genes *SLC45A2*, *OCA2*, *HERC2*, *SLC24A5* and *TYR*. While there is extensive evidence supporting the adaptive significance of these alleles, debates around their exact origins and spread are largely driven by comparisons of allele frequency estimates in population groups which are almost always discretized in time and/or space. Among these, selection at the *TYR* locus is thought to have occurred particularly recently, over the last 5,000 years (Stern *et al.*, 2019), driven by a recent mutation (Albers & McVean, 2020) that may have spread rapidly in Western Eurasia.

Here, we develop a method to model the spread of a recently selected allele across both space and time, avoiding artificial discretization schemes to more rigorously assess the evidence for or against a particular dispersal process. We begin with the model proposed by Novembre *et al.* (2005), and adapt it in order to handle ancient low-coverage genomic data, and explore more complex models that allow for both diffusion and advection (i.e. directional transport) in the distribution of allele frequencies over space, as well as for a change in these parameters at different periods of time. We apply the method to alleles in two of the aforementioned loci in the human genome, which have been reported to have strong evidence for recent positive selection: *LCT/MCM6* and *TYR*. We focus on Western Eurasia during the Holocene, where ancient genomes are most densely sampled, and infer parameters relevant to the spread of these alleles, including selection, diffusion and advection coefficients.

Results

Summary of model

We based our statistical inference framework on a model proposed by Novembre *et al.* (2005) to fit allele frequencies in two dimensions to present-day genotype data spread over a densely sampled map. We extend this model in several ways:

- We incorporate temporally sampled data (ancient genomes) to better resolve changes in frequency distributions over time
- We make use of genotype likelihoods and pseudohaploid genotypes to incorporate low-coverage data into the inference framework
- We permit more general dynamics by including advection parameters.
- We allow the selection, advection and diffusion parameters to be different in different periods of time. Specifically, to reflect changes in population dynamics and mobility before and after the Bronze Age (Loog *et al.*, 2017; Racimo *et al.*, 2020a), we partitioned the model fit into two time periods: before and after 5,000 years BP.

We explored the performance of two different spread models, which are extensions of the original model by Novembre *et al.* (2005), hereby called model A. This is a diffusion model containing a selection coefficient s (determining the rate of local allele frequency growth) and a single diffusion term (σ). A more general diffusion model - hereby model B - allows for two distinct diffusion parameters for latitudinal (σ_y) and longitudinal (σ_x) spread. Finally, model C is even more general and includes two advection terms (v_x and v_y), allowing the center of mass of the allele's frequency to diverge from its origin over time. The incorporation of advection is meant to account for the fact that population displacements and expansions could have led to allele frequency dynamics that are poorly explained by diffusion alone.

In order to establish a starting time point for our diffusion process, we used previously published allele age estimates obtained from a non-parametric approach leveraging the patterns of haplotype concordance and discordance around the mutation of interest (Albers & McVean, 2020). In the case of the allele in the *LCT/MCM6* region, we also used age estimates based on an approximate Bayesian computation approach (Itan *et al.*, 2009).

Performance on deterministic simulations

To characterize the accuracy of our inference method under different parameter choices we first generated deterministic simulations from several types of diffusion models. First, we produced an allele frequency surface map with a specified set of parameters from which we drew 1,040 samples matching the ages, locations and genotype calling format (diploid vs. pseudo-haploid) of the 1,040 genomes that we analyze below when studying the rs1042602(A) allele.

We generated six different simulations with different diffusion coefficients and afterwards ran our method assuming model B. The results (simulations B1-B6) are summarised in figures 1, S1, S2, S3, S4, S5 and table S1. Overall, the model is more accurate at correctly inferring the parameters for the time period before 5,000 years BP (figure 1b), with decreased performance when longitudinal diffusion is high (figure S5).

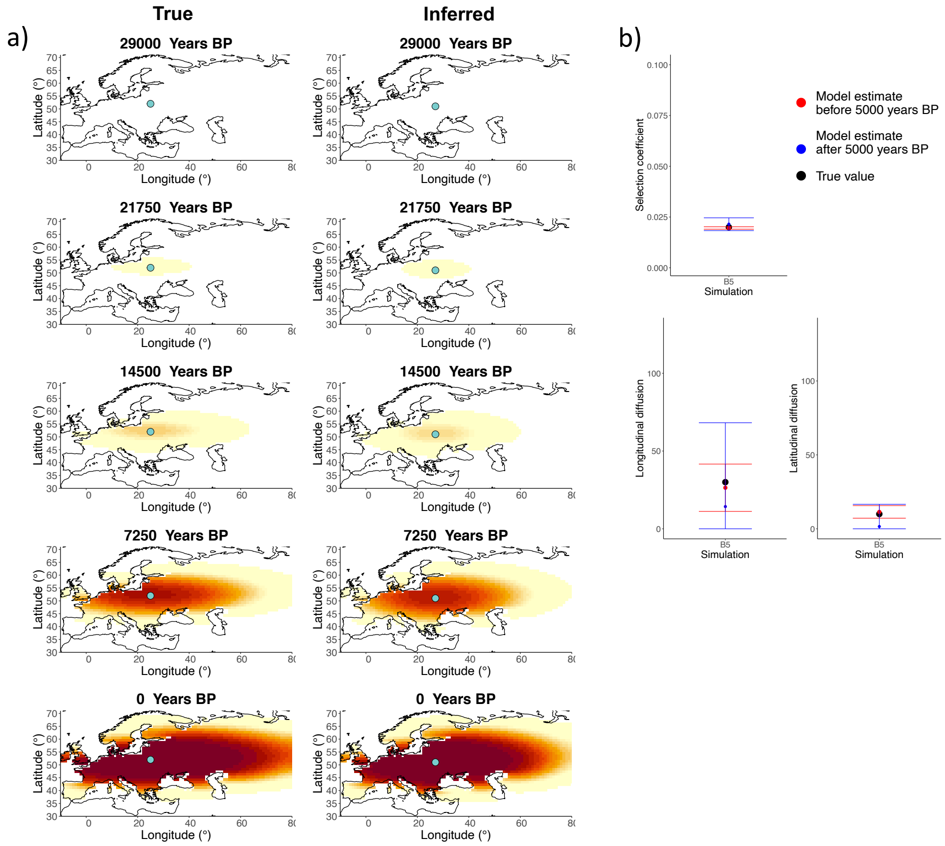


Figure 1: a) Comparison of true and inferred allele frequency dynamics for a simulation with diffusion and no advection (B5). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S1. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

Next, we investigated the performance of model C, which includes advection coefficients. We generated four different simulations including advection (simulations C1-C4: Figure 2, supplementary figures S6, S7, S8 and table S2). We found that our method is generally able to estimate the selection coefficient accurately. However, in some of the simulations, we found discrepancies between the estimated and true diffusion and advection coefficients, often occurring because of a misestimated origin forcing the other parameters to adjust in order to better fit the allele frequency distribution in later stages of the allele's spread (Figure 2). Despite the disparities between the true and inferred parameter values, the resulting surface plots become very similar as we approach the present, suggesting that different combinations of parameters can produce similar present-day allele frequency distributions.

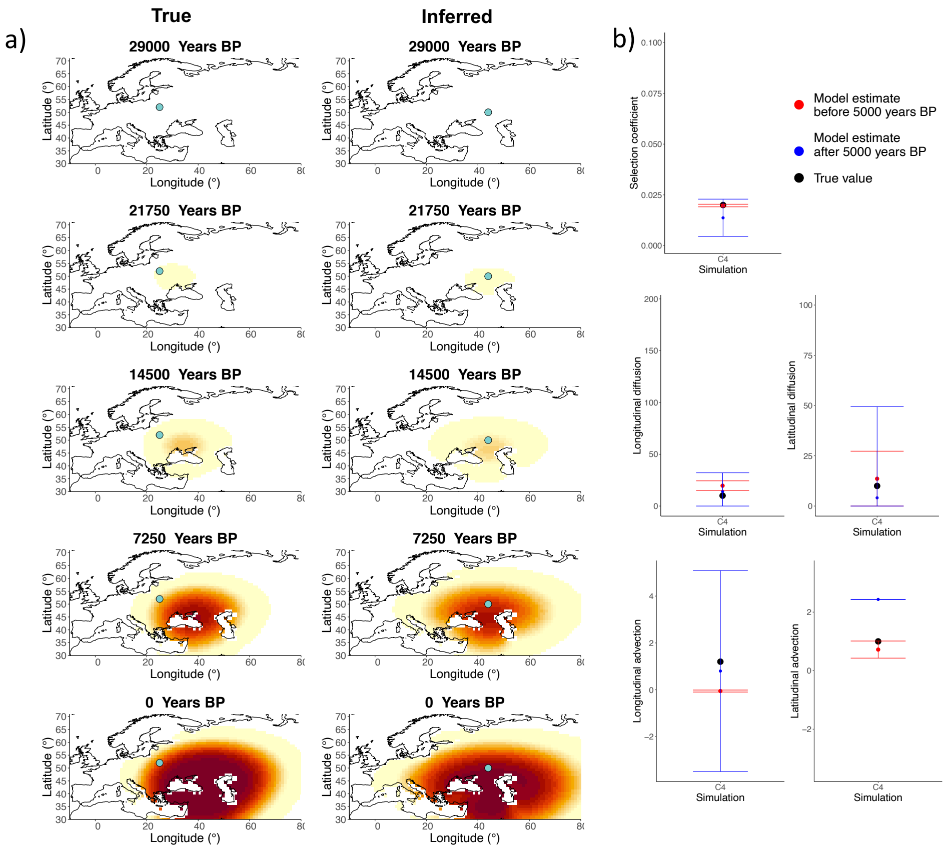


Figure 2: a) Comparison of true and inferred allele frequency dynamics for one of the simulations including advection (C4). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S2. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

Spatially-explicit forward simulations

In addition to drawing simulated samples from a diffusion model, we performed spatially explicit individual-based forward-in-time simulations of selection acting on a beneficial allele using a new simulation framework implemented in the R package *slendr* (Petr (2021)). This package makes it possible to define spatiotemporal population models in R and then feeds them into the forward population genetic simulator SLiM (Haller & Messer (2019)) for generating genotype data.

We introduced a single beneficial additive mutation in a single individual and let it evolve across the European landscape. Before applying our method on the simulated data, we sampled 1,040 individuals whose ages were log-uniformly distributed, to ensure that there were more samples closer to the present, as in the real data. We transformed the diploid genotypes to pseudohaploid genotypes by assigning a heterozygous individual an equal probability of carrying the ancestral or the derived genotype. The parameter values estimated by our model to the simulations described in this section are summarised in table S3.

We can see that the origin of the allele inferred by the model closely corresponds to the first observation of the derived allele in the simulation (figure 3). The inferred selection coefficient is only slightly higher than the true value from the simulation (0.0366 vs 0.030). In general, the model accurately captures the spread of the allele centered in central Europe, though we observe some discrepancies due to differences between the model assumed in the simulation (which, for example, accounts for local clustering of individuals, figure S9), and that assumed by our diffusion-based inference.

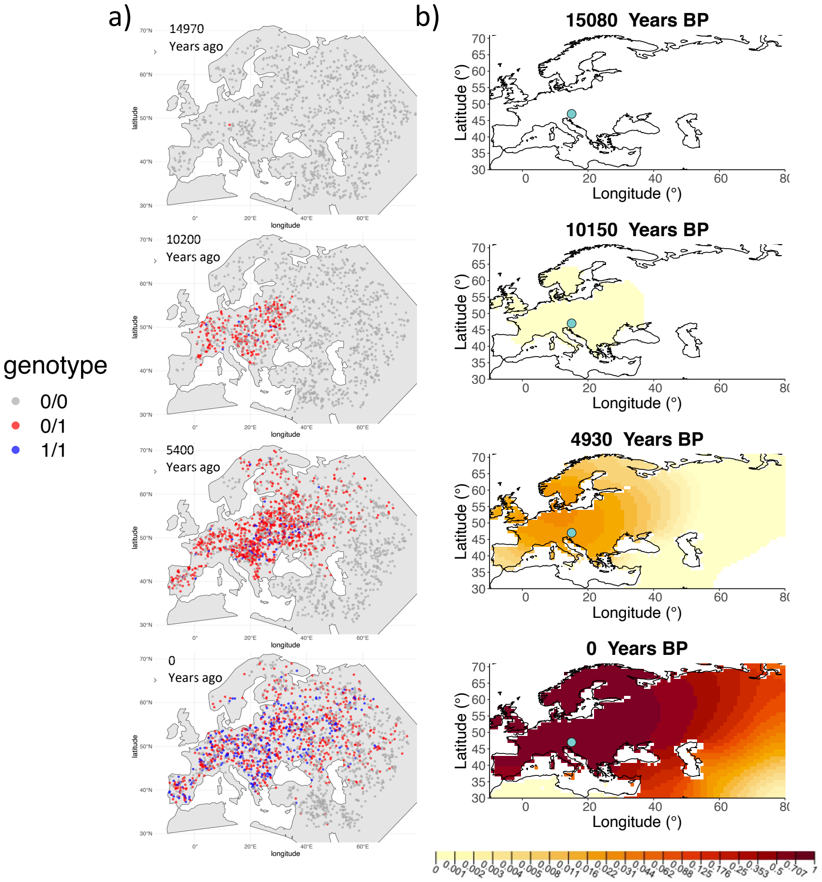


Figure 3: A) Individual-based simulation of an allele that arose in Central Europe 15,000 years ago with a selection coefficient of 0.03. Each dot represents a genotype from a simulated genome. To avoid overplotting, only 1,000 out of the total 20,000 individuals in the simulation in each time point are shown for each genotype category. B) Allele frequency dynamics inferred by the diffusion model on the individual-based simulation to the left, after randomly sampling 1,040 individuals from the simulation and performing pseudohaploid genotype sampling on them. The individuals' ages were log-uniformly distributed. The estimated parameter values of the fitted model are shown in table S3.

Dynamics of the rs4988235(T) allele

Having tested the performance of our method on simulated data, we set out to infer the allele frequency dynamics of the rs4988235(T) allele (associated with adult lactase persistence) in ancient Western Eurasia. For our analysis, we used a genotype dataset compiled by Segurel *et al.* (2020), which amounts to 1,434 genotypes from ancient Eurasian genomes individuals, and a set of 36,659 genotypes from present-day Western and Central Eurasian genomes (Ségurel & Bon, 2017; Heyer *et al.*, 2011; Marchi *et al.*, 2018; Liebert *et al.*, 2017; Gallego Romero *et al.*, 2012; Itan *et al.*, 2010; Charati *et al.*, 2019). After filtering out individuals falling outside of the range of the geographic boundaries considered in this study, we retained 1,332 ancient individuals. The locations of ancient and present-day individuals used in the analysis to trace the spread of rs4988235(T) are shown in figure 4.

We used a two-period scheme by allowing the model to have two sets of estimates for the selection coefficient and the diffusion and advection coefficients in two different periods of time: before and after 5,000 years ago, reflecting the change in population dynamics and mobility before and after the Bronze Age transition (Loog *et al.*, 2017; Racimo *et al.*, 2020a). We used two allele age estimates as input: a relatively young one (7,441 years ago) obtained from Itan *et al.* (2009), and a relatively old one (20,106 years ago) obtained from Albers & McVean (2020). The results obtained for fitting the model on rs4988235(T) are summarised in tables S4 and S5, and in figures 5b (younger age) and S12 (older age).

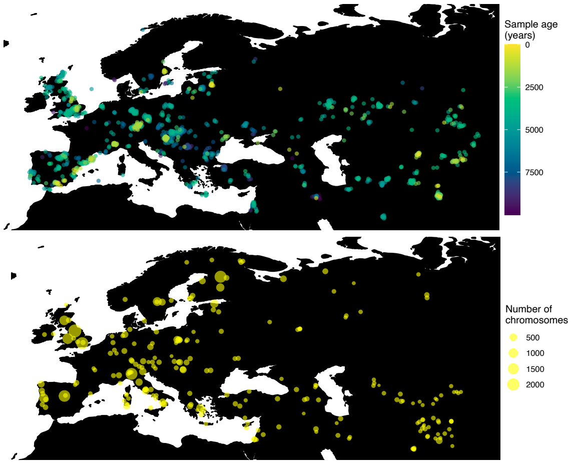


Figure 4: Locations of samples used to model the spread of the rs4988235(T) allele. The upper panel shows the spatiotemporal locations of ancient individuals, the bottom panel represents the locations of present-day individuals.

Assuming the age estimate from (Itan *et al.*, 2009), the origin of the allele is estimated to be north of the Caucasus, around what is now southwestern Russia and eastern Ukraine (Figure 5b). Given that this age is relatively young, our method fits a very strong selection coefficient (≈ 0.1) during the first period in order to accommodate the early presence of the allele in various points throughout Eastern Europe, and a weaker (but still strong) selection coefficient (≈ 0.03) in the second period. We also estimate stronger diffusion in the second period than in the first, to accommodate the rapid expansion of the allele throughout

Western Europe, and a net westward advection parameter, indicating movement of the allele frequency's center of mass to the west as we approach the present.

Assuming the older age estimate from Albers & McVean (2020), the origin of the allele is estimated to be in the Northeast of Europe (figure S12), which is at a much higher latitude than the first occurrence of the allele, in Ukraine. A comparison of the parameters related to the allele expansion inferred for the two time periods shows that the allele initially expands at a much higher rate in the latitudinal direction relative to the longitudinal direction (table S5). This difference greatly decreases in the second time period. The model appears to restrict the expansion of the allele in the region with a lower density of available aDNA data and thus avoids an overlap of the increasing allele frequencies with individuals who do not carry the derived rs4988235(T) allele (see figure 5a). The rapid expansion in the southern direction allows the model to eventually reach the sample carrying the derived variant in Ukraine. As the rs4988235(T) allele becomes more widely distributed after 5000 years BP, the longitudinal diffusion and advection parameters in the second period are higher than in the first.

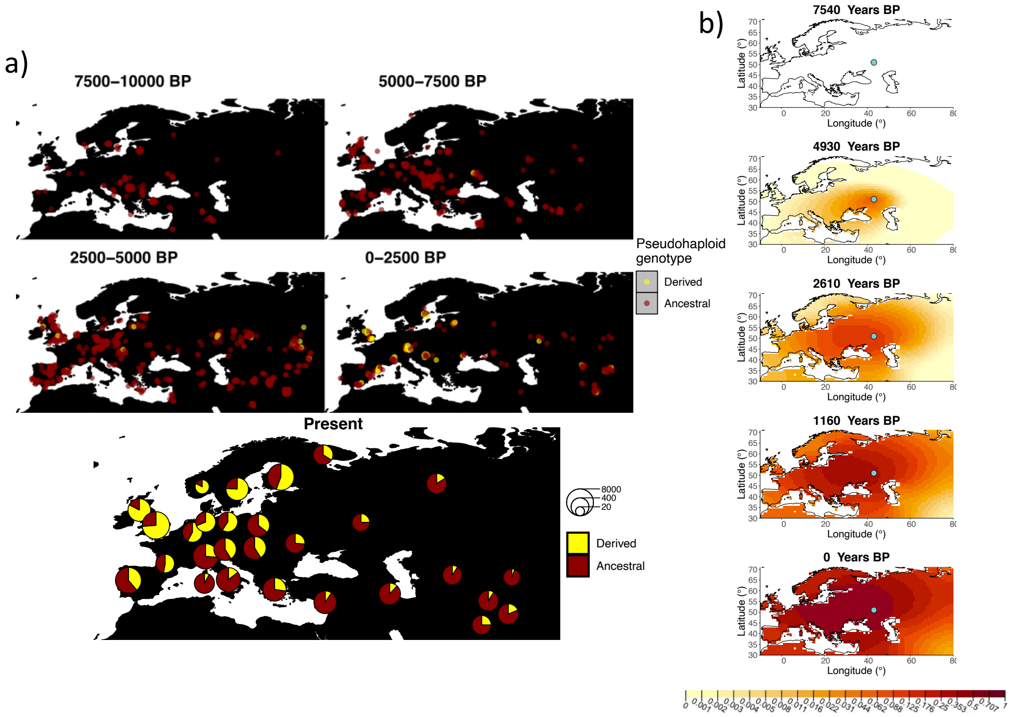


Figure 5: a) Top: Pseudohaploid genotypes of ancient samples at the rs4988235 SNP in different periods. Yellow corresponds to the rs4988235(T) allele. Bottom: allele frequencies of present-day samples represented as pie charts. The size of the pie charts corresponds to the number of available sequences in each region. b) Inferred allele frequency dynamics of rs4988235(T). The green dot indicates the inferred geographic origin of the allele.

Dynamics of the rs1042602(A) allele

Next, we investigated the spatiotemporal dynamics of the spread of an allele at a pigmentation-associated SNP in the *TYR* locus (rs1042602(A)), which has been reported to be under recent selection in Western Eurasian history (Stern *et al.*, 2019). For this purpose, we applied our method to the Allen Ancient DNA Resource data (AADR: https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data), which contains randomly sampled pseudohaploid genotypes from 1,513 published ancient Eurasian genomes (listed in Supplementary Text 1), from which we extracted those genomes that had genotype information at this locus in Western Eurasia. We merged this dataset with diploid genotype information from high-coverage present-day West Eurasian genomes from the Human Genome Diversity Panel (HGDP) (Bergström *et al.*, 2020), which resulted in a total of 1,040 individuals with genotype information at rs1042602, which were as input to our analysis. Geographic locations of individuals in the final dataset are shown in Figure 6.

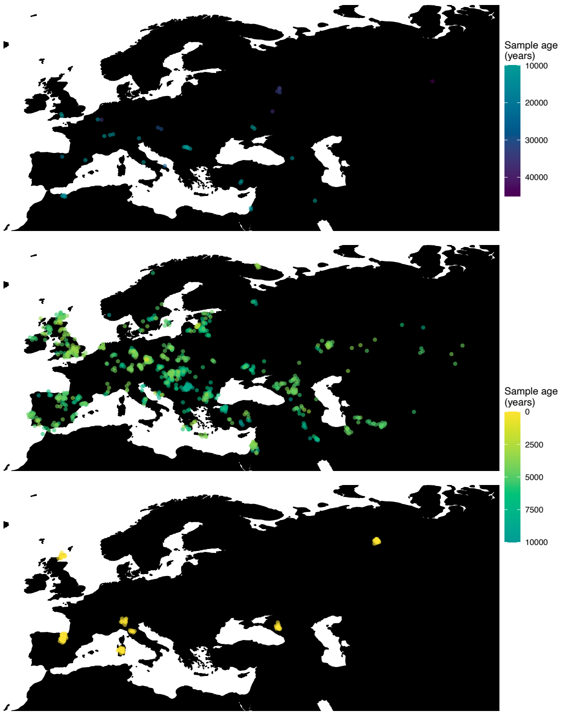


Figure 6: Spatiotemporal sampling locations of sequences used to model the rs1042602(A) allele in Western Eurasia. Upper panel: ancient individuals dated as older than 10,000 years ago. Middle panel: ancient individuals dated as younger than 10,000 years ago. Bottom panel: present-day individuals from HGDP.

Similarly to our analysis of the spread of the allele in rs4988235(T), we inferred the dynamics of the rs1042602(A) allele separately for the time periods before and after 5,000 years BP. The inferred parameters for both time periods are summarised in table S6 and the allele frequency surface maps generated using these parameters are shown in figure 7b. The origin of the rs1042602(A) corresponds closely to the region where the allele initially starts to segregate in the time period between 7,500 and 10,000 years BP as seen in figure 7a. Estimates of the selection coefficient for both time periods (0.0221 and 0.0102 for the period before and after 5000 years BP, respectively) suggest that selection acting on the allele has decreased after 5000 years BP.

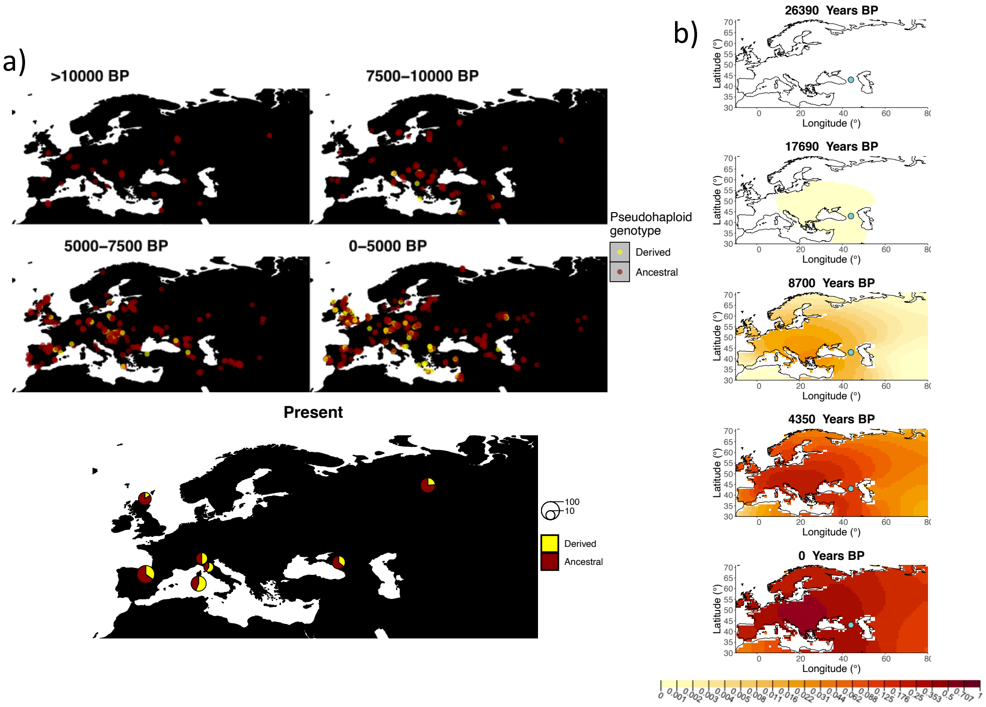


Figure 7: a) Top: Pseudohaploid genotypes of ancient samples of the rs1042602 in different periods. Yellow corresponds to the A allele. Bottom: diploid genotypes of present-day samples. b) Inferred allele frequency dynamics of rs1042602(A). The green dot corresponds to the inferred geographic origin of the allele.

Robustness of parameters to the inferred geographic origin of allele

We carried out an analysis to characterize how sensitive the selection, diffusion and advection parameters are to changes in the assumed geographic origin of the allele. For the rs4988235(T) allele, we forced the origin of the allele to be 10 degrees away from our inferred origin in each cardinal direction, while assuming the allele age from Itan *et al.* (2009) (table S7). In figures S13, S14, S15 and S16, we can see the allele frequency dynamics of these four scenarios, respectively. We also forced the allele origin to be at the geographic origin estimated in Itan *et al.* (2009) (table S8, figure S17), which is westward of our estimate. In all five cases during the period prior 5,000 years BP, the allele is inferred to expand in the direction of the first sample that is observed to carry the rs4988235(T) allele and is located in Ukraine. During the time period after 5000 years BP, the patterns produced by the model are rather similar, although the parameters associated with diffusion and advection differ, in order to account for the different starting conditions.

We also investigated how the results are affected when the estimated geographic origin of the rs1042602(A) allele is moved with respect to the initial estimate. We set the allele to be 10 degrees east, 10 degrees north and 10 degrees south of the original estimate as shown in figures S18, S19 and S20, respectively. We did not look at a scenario in which the origin of the allele is moved to the west, since it would either end up in the Black sea or more westwards than 10 degrees. The selection coefficient remains similar to the original estimate throughout all three scenarios. The way the allele spreads across the landscape is also similar in all cases and, as in the case of rs4988235(T), the model accounts for the different origins of the allele by adjusting the diffusion and advection coefficients in the time period after 5000 years BP.

Robustness of parameters to the assumed age of the allele

In order to investigate how sensitive our inferences are to the point estimates of allele ages we obtained from the literature (Albers & McVean, 2020; Itan *et al.*, 2009), we also fitted our model using the upper and lower ends of the 95% confidence intervals or credible intervals for each age estimate (depending on whether the inference procedure in the literature was via a maximum likelihood or a Bayesian approach). For the rs4988235(T) allele, the reported credible intervals for the Itan *et al.* (2009) age are 8,683 and 6,256 years BP. For the rs1042602(A) allele, the reported confidence intervals for the age are 27,315 and 25,424 years BP (Albers & McVean, 2020).

When re-fitting the model for the rs4988235(T) allele, we found that the inferred selection coefficient is slightly lower when the allele age is assumed to be at the lower bound of the 95% credible interval and slightly higher when assumed to be at the upper bound (table S4 and figures S21 and S22). This occurs because the selection intensity must be higher or lower when there is more or less time, respectively, for the allele to reach the allele frequencies observed in the data. In the case of the rs1042602(A) allele, this only affects the earlier time period (table S6). The rs4988235(T) allele's geographic distribution in the more recent time periods is also less extended geographically when the age is assumed to be young. The inferred geographic origin of both alleles slightly differs under different assumed ages (figures S23 and S24).

Discussion

A spatially explicit framework for allele frequency diffusion can provide new insights into the dynamics of selected variants across a landscape. We have shown that under the conditions of strong, recent selection, our method can infer selection and dispersal parameters, using a combination of ancient and present-day human genomic data. However, when allowing for advection, the inferred location tends to become less accurate. This suggests that migration events early in the dispersal of the selected allele could create difficulties in finding the true allele origin if net directional movement (i.e. via major migratory processes) had a large effect in this dispersal. This issue could be alleviated with the inclusion of more ancient genomes around the time of the mutational origin, perhaps in combination with a more fine-scaled division into periods where advection may have occurred in different directions.

The inferred geographic origin of the rs4988235(T) allele reflects the best guess of our framework given the constraints provided by its input, namely the previously inferred age of the allele and the observed instances of this allele throughout Western Eurasia. We are also assuming that the allele must have arisen somewhere within the bounding box of our studied map. When assuming a relatively young allele age (7,441 years ago, Itan *et al.* (2009)), the origin of the allele is placed north of the Caucasus, perhaps among steppe populations that inhabited the area at this time (Haak *et al.*, 2015; Allentoft *et al.*, 2015). This origin is further east than the geographic origin estimate from Itan *et al.* (2009), likely reflecting additional ancient DNA information that is available to us, and indicates an early presence of the allele in eastern Europe. When assuming a relatively old allele age (20,106 years ago, Albers & McVean (2020)), the age is placed in northeast Europe, perhaps among Eastern hunter-gatherer groups that inhabited the region in the early Holocene. We note that the number of available genomes for eastern and northeastern Europe during the early Holocene is scarce, so our confidence on the exact location of this origin is necessarily low. Regardless of the assumed age, we estimate a net westward displacement of the allele frequency's center of mass, and a rapid diffusion, particularly in the period after 5,000 years ago.

Various studies have estimated the selection coefficient for the rs4988235(T) allele, and these range from as low as 0.014 to as high as 0.19 (Enattah *et al.*, 2008; Mathieson & Mathieson, 2018; Mathieson, 2020; Stern *et al.*, 2019; Burger *et al.*, 2020; Peter *et al.*, 2012; Gerbault *et al.*, 2009; Itan *et al.*, 2009; Bersaglieri *et al.*, 2004). Recent papers incorporating ancient DNA estimate the selection coefficient to be as low as 0 (in certain regions of Southern Europe) and as high as 0.06 (Mathieson & Mathieson, 2018; Mathieson, 2020; Burger *et al.*, 2020). It is also likely that the selection coefficient was different for different regions of Europe, perhaps due to varying cultural practices (Mathieson, 2020). In our case, the estimated selection coefficient during the first period - before 5,000 years ago - depends strongly on the assumed allele age ($s = 0.0993$ vs. $s = 0.0285$). As in the case of the geographic origin, these estimates should be taken with caution as the number of available allele observations in the early Holocene is fairly low. The estimates for the second period - after 5,000 years ago - are more robust to the assumed age: $s = 0328$ (95% CI: 0.0327–0.0329) if we assume the younger allele age (7,441 years ago) and $s = 0.0255$ (95% CI: 0.0252–0.0258) if we assume the older allele age (20,106 years ago). These estimates are also within the range of previous estimates.

In the case of the rs1042602(A) allele, our estimated selection coefficients of 0.0221 (95% CI: 0.0216-0.0227) and 0.0102 (95% CI: 0.0083-0.0120) for the time periods before and after 5000 years BP, respectively, are generally in agreement with previous results. Wilde *et al.* (2014) used a forward simulation approach to infer a point estimate of 0.026. Another study using an approximate Bayesian computation framework (Nakagome *et al.*, 2019) estimated the strength of selection acting on rs1042602 to be 0.013 (0.002–0.029). Although both studies utilized ancient DNA data, the estimates were obtained without explicitly modelling the spatial dimension of the selection process.

Our estimates of the longitudinal advection parameter are negative for both the SNPs in the *TYR* and *LCT* loci: the mutation origins are always to the east of the center of mass of the allele frequency distribution seen in present-day data. This perhaps reflects common migratory processes, like the large-scale Neolithic and Bronze Age population movements from east to west, affecting the allele frequencies at these loci across the Eurasian landscape (Allentoft *et al.*, 2015; Haak *et al.*, 2015). As a form of regularization, we kept the range of explored values for the advection parameters to be small (-2.5 to 2.5 km per generation), while allowing the diffusion parameters to be explored over a much wider range of values. In certain cases, like the second period of the rs4988235(T) spread when the allele age is assumed to be young (table S4), we find that the advection parameters are fitted at the boundary of the explored range, because the allele needs to spread very fast across the landscape to fit the data. A future improvement to our method could include other forms of regularization that better account for the joint behavior of the advection and diffusion processes, or the use of priors for these parameters under a Bayesian setting, which could be informed by realistic assumptions about the movement of individuals on a landscape.

When investigating the robustness of the geographic origin of both rs4988235(T) and rs1042602(A), we found that parameters related to the beneficial allele's expansion change in response to different assumed origins of the allele. The resulting allele frequency surface plots, however, appear very similar throughout the later stages of the process, showing that the model tends to adjust the diffusion and advection coefficients in a way such that the allele will end up expanding into the same areas regardless of the origin.

As we apply these methods to longer time scales and broader geographic areas, the assumptions of spatiotemporal homogeneity of the parameters seem less plausible. There may be cases where the allele may have been distributed over a wide geographic area but remained at low frequencies for an extended period of time, complicating the attempts to pinpoint the allele's origin. In our study, we estimated diffusion and selection coefficients separately for two time periods before and after 5000 years ago to account for changes in mobility during the Neolithic transition, but this approach may still be hindered by uneven sampling, especially when the allele in question exists at very low frequencies. Notably, our results for the spread of the rs4988235(T) allele during the older time period should be interpreted with caution, since they may be affected by sparse sampling in the early Holocene.

Potential future extensions of our method could incorporate geographic features and historical migration events that create spatially or temporally varying moderators of gene flow. An example of this type of processes is the retreat of glaciers after the last Glacial maximum, which allowed migration of humans into Scandinavia (Günther *et al.*, 2018). These changing geographic features could lead to changes in the rate of advection or diffusion across time or space. They could also serve to put more environmentally-aware constraints on the geographic origin of the allele, given that it cannot have existed in regions uninhabitable by humans, and to extend our analyses beyond the narrow confines of the Western Eurasian map chosen for this study. One could also envision incorporating variation in population densities over time, or known migration processes in the time frames and regions of interest. These might have facilitated rapid, long-range dispersal of beneficial alleles (Bradburd *et al.*, 2016; Hallatschek & Fisher, 2014) or caused allelic surfing on the wave of range expansions (Klopfstein *et al.*, 2006). Additional information like this could come, for example, from previously inferred spatiotemporal demographic processes (e.g. Racimo *et al.* (2020b)).

As described above, our model only accounts for diffusion in two directions. Further extension of our model could therefore incorporate anisotropic diffusion (Othmer *et al.*, 1988; Painter & Hillen, 2018). Another possibility could be the introduction of stochastic process components, in order to convert the partial differential equations into stochastic differential equations (Brown *et al.*, 2000). Stochastic components could serve to induce spatial autocorrelation and capture local patterns of allele frequency covariance in space that might not be well modeled by the deterministic PDEs (Cressie & Wikle, 2015). They could also serve to induce stochasticity in allele frequency changes over time as a consequence of genetic drift (Crow *et al.*, 1970), allowing one to model the dynamics of more weakly selected variants, where drift plays an important role. Eventually, one could perhaps combine information across loci to jointly model the spatiotemporal frequency surfaces at multiple loci associated with the same trait. This could help clarify the dynamics of polygenic adaptation and negative selection on complex traits (Irving-Pease *et al.*, 2021), and perhaps hindcast the genetic value of traits across a landscape.

The availability of hundreds of ancient genomes (Marciniak & Perry, 2017) and the increasing interest in spatiotemporal method development (Bradburd & Ralph, 2019), such as the one described in this manuscript, will likely lead researchers to posit new questions and hypotheses about the behavior of natural selection. In the case of a beneficial allele spreading on a landscape, new ontologies and vocabulary for describing positive selection in time and space will be needed. Abundant terms exists to classify the initial conditions and dynamics of a selective sweep in a single population (hard sweep, multiple origin soft sweep, single origin soft sweep, partial sweep) (Hermisson & Pennings, 2005; Pritchard & Di Rienzo, 2010; Hermisson & Pennings, 2017). In contrast, there is a lack of vocabulary for distinguishing between a scenario of strong selection that is locally constrained in space from a scenario of widespread selection extended over a landscape, or a model of neutral diffusion in space followed by parallel non-neutral increases in frequency at multiple locations. For example, Ralph & Coop (2010) showed how multiple localized hard sweeps may be seen as a soft sweep at a larger population-wide scale. Existing vocabulary for spatiotemporal genetic processes is clearly not enough, limiting the types of questions or hypotheses we can pose about them.

Population genetic models that explicitly account for space and time are an important area of future methodological development (Bradburd & Ralph, 2019). We believe that methods such as the one described in this study show great promise at broadening the horizon of our understanding of natural selection across space and time in humans and other species. As in the case of demographic reconstruction (Ray & Excoffier, 2009), spatiotemporal information can greatly help improve our knowledge of how natural selection operated in the past.

Methods

The model

To describe the allele frequency dynamics in time and space, we first begin by using a deterministic model based on a two-dimensional partial differential equation (PDE) (Fisher, 1937; Kolmogorov *et al.*, 1937; Novembre *et al.*, 2005). This PDE represents the distribution $p(x, y, t)$ of the allele frequency across a two dimensional (x, y) landscape at time t:

$$\frac{\partial p}{\partial t} = \frac{1}{2}\sigma^2 \frac{\partial^2 p}{\partial x^2} + \frac{1}{2}\sigma^2 \frac{\partial^2 p}{\partial y^2} + \gamma(p, s, d) \tag{1}$$

where

$$\gamma(p, s, d) = p(1 - p)(pd + s(1 - 2p)). \tag{2}$$

Here, σ is the diffusion coefficient, s is the selection coefficient, and d is the dominance coefficient (Novembre *et al.*, 2005). We assumed an additive model and fixed $d = 2s$ in all analyses below. We call this "model A", but we also evaluated the fit of our data under more complex models which are more flexible, and are described below.

Model B is a more general diffusion-reaction model, which incorporates distinct diffusion terms in the x and y axes (σ_x and σ_y, respectively):

$$\frac{\partial p}{\partial t} = \frac{1}{2}\sigma_x^2 \frac{\partial^2 p}{\partial x^2} + \frac{1}{2}\sigma_y^2 \frac{\partial^2 p}{\partial y^2} + \gamma(p, s, d) \tag{3}$$

Model C is a generalization of model B that incorporates advection terms in the x and y directions (see e.g. Cantrell & Cosner (2004) for a motivation of this type of model in the context of spatial ecology):

$$\frac{\partial p}{\partial t} = \frac{1}{2}\sigma_x^2 \frac{\partial^2 p}{\partial x^2} + \frac{1}{2}\sigma_y^2 \frac{\partial^2 p}{\partial y^2} + v_x \frac{\partial p}{\partial x} + v_y \frac{\partial p}{\partial y} + \gamma(p, s, d) \tag{4}$$

Here, v_x and v_y represent the coefficients for advective velocity along the x and y axes respectively.

In the Appendix, we motivate the construction of these equations using model C as an example, and show that equation 4 can be obtained by taking an infinitesimal limit of a random walk on a two-dimensional lattice, after including a reaction term due to selection. Models A and B are then shown to be special cases of model C.

For evaluating the likelihood of the observed data, we use a binomial genotype sampling model. Let $g_i \in 0, 1, 2$ be the genotype of individual i at the locus of interest, let a_i be the number of reads carrying ancestral alleles, let d_i be the number of reads carry derived reads. Let (x_i, y_i) be the coordinates of the location from which individual i was sampled, and t_i its estimated age (e.g. from radiocarbon dating). Then, the likelihood for individual i can be computed as follows:

$$L(d_i, a_i) = \sum_{h=0}^{2} P[d_i, a_i | g_i = h] P[g_i = h | p(x_i, y_i, t_i)] \tag{5}$$

Here, $p(x_i, y_i, t_i)$ is the solution to one of the partial differential equations described above (equations (1), (2) or (4), depending on the process model chosen), evaluated at location (x_i, y_i) and time t_i. In turn, $P[d_i, a_i|g_i = h]$ is the likelihood for genotype i. Furthermore, $P[g_i = h|p(x_i, y_i, t_i)]$ is a binomial distribution, where n represents the ploidy level, which in this case is 2:

$$P[g_i = h|p(x_i, y_i, t_i)] = \binom{n}{h} p(x_i, y_i, t_i)^h (1 - p(x_i, y_i, t_i))^{n-h} \tag{6}$$

Then, the likelihood of the entire data can be computed as

$$L(\mathbf{d}, \mathbf{a}) = \prod_{i=1}^{M} L(x_i, y_i, t_i) \tag{7}$$

where M is the total number of individuals for which we have data, **d** is the vector containing the derived read count for each individual and **a** is the vector containing the ancestral read count for each individual. We computed genotype likelihoods directly on the BAM file read data, using the SAMtools genotype model (Li, 2011) implemented in the software ANGSD (Korneliussen *et al.*, 2014).

When only randomly sampled pseudohaploid allele counts are available, we used a Bernoulli sampling likelihood (conditional on the genotype g_i) on the left-hand side of equation 5 instead. Briefly, assuming that the probability of an individual having genotype g at a particular locus given the underlying allele frequency p follows a binomial distribution and that the probability of sampling a read given the genotype of an individual follows a Bernoulli distribution with probability of success $\frac{1}{2}g$, then the probability of sampling a read given the genotype follows a Bernoulli distribution with probability of success p.

Map

We restricted the geographic area explored by our model fit to be between 30°N to 75°N, and between 10°W and 80°E. For numerical calculations, we used a grid constructed using a resolution of approximately 1 grid cell per latitude and longitude. We used Harvesine functions in order to transform the distance from degrees to kilometers between two geographic points. The diffusion of the allele frequency was disallowed in the map regions where the topology is negative (i.e. regions under water), based on ETOPO5 data (NOAA (1988)). For this reason we added land bridges between the European mainland and Sardinia, and between the mainland and Great Britain, in order to allow the allele to diffuse in these regions (see figure S10).

Parameter search

Parameter optimization was done via maximum likelihood estimation with a two-layer optimization set-up. The first layer consists of a simulated annealing approach (Bélisle (1992)) starting from 50 random points in the parameter space. The initial 50 points are sampled using latin hypercube sampling to ensure an even spread across the parameter space. The output of this fit was then fed to the L-BFGS-B algorithm to refine the parameter estimates around the obtained maximum and obtain confidence intervals for the selection, diffusion and advection parameters (Byrd *et al.* (1995)).

The parameters optimised were:

- the selection coefficient (s), restricted to the range 0.001-0.1
- two dispersal parameters σ_x and σ_y in the longitudinal and latitudinal directions respectively, restricted to the range of 1-100 square-kilometers per generation
- the longitudinal and latitudinal advection coefficients v_x and v_y respectively. As a form of regularization, we set the range of explored values to be narrowly centered around zero: -2.5 to 2.5 kilometers per generation
- the geographic origin of the allele, which is randomly initialized to be any of the 28 spatial points shown in Figure S11 at the start of the optimization process

The latitude and longitude are discretized in our model in order to solve the differential equations numerically, thus the origin of a mutation is measured in terms of discrete units. For this reason, when using the L-BFGS-B algorithm, we fixed the previously estimated origin of the allele, and did not explore it during this second optimization layer. Time was measured in generations, assuming 29 years per generation. During the optimization we scaled the time and the parameters by a factor of 10, which allowed us to decrease the execution time of the model.

We initialized the grid by setting the initial allele frequency to be p_0 in a grid cell where the allele originates and 0 elsewhere. p_0 was calculated as $1/(2*D*A)$, where D is the population density and is equal to 2.5 inhabitants per square-kilometer, which is the estimated population density in Europe in 1000 B.C. (Colin McEvedy, 1978; Novembre *et al.*, 2005). In the equation, D is multiplied by 2 because we assume that the allele originated in a single chromosome in a diploid individual. A is the area in square-kilometers of the grid cell where the allele emerged.

Asymptotic 95% confidence intervals for a given parameter θ_j were calculated using equation

$$\hat{\theta_j} \pm 1.96\sqrt{(F(\boldsymbol{\theta})^{-1})_{jj}}$$

where $F(\boldsymbol{\theta})$ is an estimate of the observed Fisher information matrix (Fisher, 1922; Efron & Hastie, 2016; Casella & Berger, 2021).

Implementation

The above described model was implemented in R version 3.6. To numerically solve the differential equations and obtain maximum likelihood estimates, we used the libraries *deSolve* (Soetaert *et al.*, 2010), *ReacTran* (Soetaert & Meysman, 2012) and *bbmle* (Bolker & R Development Core Team, 2020). Scripts containing the code used in this paper are available on github:
https://github.com/RasaMukti/stepadna

Indvidual-based simulations

For the individual-based spatiotemporal forward simulations, we first defined a spatial boundary for a population spread across a broad geographic region of Europe. In order to ensure a reasonably uniform distribution of individuals across this spatial range throughout the course of the simulation, we set the maximum distance for spatial competition and mating choice between individuals to 250 km (translated, on a SLiM level, to the interaction parameter *maxDistance*), and the standard deviation of the normal distribution governing the spread of offspring from their parents at 25 km (leveraged in SLiM's *modifyChild()* callback function) (Haller & Messer, 2019). We note that we have chosen the values of these parameters merely to ensure a uniform spread of individuals across a simulated landscape. They are not intended to represent realistic estimates for these parameters at any time in human history.

After defining the spatial context of the simulations and ensuring the uniform spread of individuals across their population boundary, we introduced a single beneficial additive mutation in a single individual. In order to test how accurately our model can infer the parameters of interest, we simulated a scenario in which the allele appeared in Central Europe 15,000 years ago with the selection coefficient of the beneficial mutation set to 0.03. Over the course of the simulation, we tracked the position of each individual that ever lived together with its location on a two-dimensional map, as well as its genotype (i.e. zero, one, or two copies of the beneficial allele). We then used this complete information about the spatial distribution of the beneficial allele in each time point to study the accuracy of our model in inferring the parameters of interest.

Acknowledgments

We thank Graham Gower, Evan Irving-Pease, Montgomery Slatkin and the members of the Racimo group for helpful comments and advice. FR and RM were funded by a Villum Fonden Young Investigator award to FR (project no. 00025300). FR was also supported by a Lundbeckfonden grant (R302-2018-2155) and a Novo Nordisk Fonden grant (NNF18SA0035006) to the GeoGenetics Centre. TSK was funded by a Carlsberg grant (CF19-0712). JN was funded by NIH grant R01 GM132383.

Competing interests

The authors declare that they have no conflict of interest.

References

Albers PK & McVean G (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS biology*, **18(1)**:e3000586

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, *et al.* (2015). Population genomics of bronze age eurasia. *Nature*, **522(7555)**:167–172

Alonso S, Izagirre N, Smith-Zubiaga I, Gardeazabal J, Díaz-Ramón JL, Díaz-Pérez JL, Zelenika D, Boyano MD, Smit N, & De la Rúa C (2008). Complex signatures of selection for the melanogenic loci tyr, tyrp1 and dct in humans. *BMC evolutionary biology*, **8(1)**:1–14

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, *et al.* (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, **367(6484)**

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, & Hirschhorn JN (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, **74(6)**:1111–1120

Bolker B & R Development Core Team (2020). *bbmle: Tools for General Maximum Likelihood Estimation.* R package version 1.0.23.1

Bradburd GS & Ralph PL (2019). Spatial population genetics: it's about time. *Annual Review of Ecology, Evolution, and Systematics*

Bradburd GS, Ralph PL, & Coop GM (2016). A spatial framework for understanding population structure and admixture. *PLoS genetics*, **12(1)**:e1005703

Brown PE, Roberts GO, Kåresen KF, & Tonellato S (2000). Blur-generated non-separable space–time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62(4)**:847–860

Burger J, Kirchner M, Bramanti B, Haak W, & Thomas MG (2007). Absence of the lactase-persistence-associated allele in early neolithic europeans. *Proceedings of the National Academy of Sciences*, **104(10)**:3736–3741

Burger J, Link V, Blöcher J, Schulz A, Sell C, Pochon Z, Diekmann Y, Žegarac A, Hofmanová Z, Winkelbach L, *et al.* (2020). Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years. *Current Biology*, **30(21)**:4307–4315.e13. 10.1016/j.cub.2020.08.033

Byrd RH, Lu P, Nocedal J, & Zhu C (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, **16(5)**:1190–1208. 10.1137/0916069. Publisher: Society for Industrial and Applied Mathematics

Bélisle CJP (1992). Convergence Theorems for a Class of Simulated Annealing Algorithms on Rd. *Journal of Applied Probability*, **29(4)**:885–895. 10.2307/3214721. Publisher: Applied Probability Trust

Cantrell RS & Cosner C (2004). *Spatial ecology via reaction-diffusion equations*. John Wiley & Sons

Casella G & Berger RL (2021). *Statistical inference*. Cengage Learning

Charati H, Peng MS, Chen W, Yang XY, Jabbari Ori R, Aghajanpour-Mir M, Esmailizadeh A, & Zhang YP (2019). The evolutionary genetics of lactase persistence in seven ethnic groups across the Iranian plateau. *Human Genomics*, **13(1)**:7. 10.1186/s40246-019-0195-5

Colin McEvedy RJ (1978). *Atlas of World Population History*. Penguin Books Lyd. and Allen Lane, Great Britain

Cressie N & Wikle CK (2015). *Statistics for spatio-temporal data*. John Wiley & Sons

Crow JF, Kimura M, *et al.* (1970). An introduction to population genetics theory. *An introduction to population genetics theory.*

Dehasque M, Ávila-Arcos MC, Díez-del Molino D, Fumagalli M, Guschanski K, Lorenzen ED, Malaspinas AS, Marques-Bonet T, Martin MD, Murray GG, *et al.* (2020). Inference of natural selection from ancient dna. *Evolution Letters*, **4(2)**:94–108

Efron B & Hastie T (2016). *Computer age statistical inference*, volume 5. Cambridge University Press

Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, *et al.* (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, **82(1)**:57–72

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, & Järvelä I (2002). Identification of a variant associated with adult-type hypolactasia. *Nature genetics*, **30(2)**:233–237

Ewens WJ (2012). *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media

Fisher RA (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **222(594-604)**:309–368

Fisher RA (1937). The wave of advance of advantageous genes. *Annals of eugenics*, **7(4)**:355–369

Gallego Romero I, Basu Mallick C, Liebert A, Crivellaro F, Chaubey G, Itan Y, Metspalu M, Eaaswarkhanth M, Pitchappan R, Villems R, *et al.* (2012). Herders of Indian and European Cattle Share Their Predominant Allele for Lactase Persistence. *Molecular Biology and Evolution*, **29(1)**:249–260. 10.1093/molbev/msr190

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboróczki L, Kővári I, Pap I, Anders A, *et al.* (2014). Genome flux and stasis in a five millennium transect of european prehistory. *Nature communications*, **5(1)**:1–9

Gerbault P, Moret C, Currat M, & Sanchez-Mazas A (2009). Impact of selection and demography on the diffusion of lactase persistence. *PLoS One*, **4(7)**:e6369

Günther T, Malmström H, Svensson EM, Omrak A, Sánchez-Quinto F, Kılınç GM, Krzewińska M, Eriksson G, Fraser M, Edlund H, *et al.* (2018). Population genomics of mesolithic scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS biology*, **16(1)**:e2003703

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, *et al.* (2015). Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, **522(7555)**:207–211

Hallatschek O & Fisher DS (2014). Acceleration of evolutionary spread by long-range dispersal. *Proceedings of the National Academy of Sciences*, **111(46)**:E4911–E4919

Haller BC & Messer PW (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, **36(3)**:632–637. 10.1093/molbev/msy228

Hermisson J & Pennings PS (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169(4)**:2335–2352

Hermisson J & Pennings PS (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution*, **8(6)**:700–716

Heyer E, Brazier L, Ségurel L, Hegay T, Austerlitz F, Quintana-Murci L, Georges M, Pasquet P, & Veuille M (2011). Lactase persistence in central Asia: phenotype, genotype, and evolution. *Human Biology*, **83(3)**:379–392. 10.3378/027.083.0304

Hudjashov G, Villems R, & Kivisild T (2013). Global patterns of diversity and selection in human tyrosinase gene. *PLoS One*, **8(9)**:e74307

Irving-Pease EK, Muktupavela R, Dannemann M, & Racimo F (2021). Quantitative paleogenetics: what can ancient dna tell us about complex trait evolution? *arXiv preprint arXiv:2105.02754*

Itan Y, Jones BL, Ingram CJ, Swallow DM, & Thomas MG (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evolutionary Biology*, **10(1)**:36. 10.1186/1471-2148-10-36

Itan Y, Powell A, Beaumont MA, Burger J, & Thomas MG (2009). The Origins of Lactase Persistence in Europe. *PLoS Computational Biology*, **5(8)**:e1000491. 10.1371/journal.pcbi.1000491

Ju D & Mathieson I (2020). The evolution of skin pigmentation associated variation in west eurasia. *bioRxiv*

Karlin S & Taylor H (1975). *A first course in Stochastic Processes*. Academic Press, New York

Klopfstein S, Currat M, & Excoffier L (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular biology and evolution*, **23(3)**:482–490

Kolmogorov A, Petrovskii I, & Piskunov N (1937). A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem. *Byull. Moskov. Univ. Ser. AMat. Mekh*, **1(6)**:1–26

Korneliussen TS, Albrechtsen A, & Nielsen R (2014). Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, **15(1)**:356

Krüttli A, Bouwman A, Akgül G, Della Casa P, Rühli F, & Warinner C (2014). Ancient dna analysis reveals high frequency of european lactase persistence allele (t-13910) in medieval central europe. *PLoS One*, **9(1)**:e86251

Lao O, De Gruijter J, van Duijn K, Navarro A, & Kayser M (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of human genetics*, **71(3)**:354–369

Li H (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27(21)**:2987–2993

Liebert A, López S, Jones BL, Montalva N, Gerbault P, Lau W, Thomas MG, Bradman N, Maniatis N, & Swallow DM (2017). World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Human Genetics*, **136(11-12)**:1445–1453. 10.1007/s00439-017-1847-y

Loog L, Lahr MM, Kovacevic M, Manica A, Eriksson A, & Thomas MG (2017). Estimating mobility using sparse data: Application to human genetic variation. *Proceedings of the National Academy of Sciences*, **114(46)**:12213–12218

Malaspinas AS, Malaspinas O, Evans SN, & Slatkin M (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics*, **192(2)**:599–607

Marchi N, Mennecier P, Georges M, Lafosse S, Hegay T, Dorzhu C, Chichlo B, Ségurel L, & Heyer E (2018). Close inbreeding and low genetic diversity in Inner Asian human populations despite geographical exogamy. *Scientific Reports*, **8(1)**:9397. 10.1038/s41598-018-27047-3

Marciniak S & Perry GH (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, **18(11)**:659

Margaryan A, Lawson DJ, Sikora M, Racimo F, Rasmussen S, Moltke I, Cassidy LM, Jørsboe E, Ingason A, Pedersen MW, *et al.* (2020). Population genomics of the viking world. *Nature*, **585(7825)**:390–396

Mathieson I (2020). Estimating time-varying selection coefficients from time series data of allele frequencies. preprint, Genetics. 10.1101/2020.11.17.387761

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, *et al.* (2015). Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, **528(7583)**:499–503

Mathieson I & McVean G (2014). Demography and the age of rare variants. *PLoS Genet*, **10(8)**:e1004528

Mathieson S & Mathieson I (2018). FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, **35(12)**:2957–2970. 10.1093/molbev/msy180

Nakagome S, Hudson RR, & Rienzo AD (2019). Inferring the model and onset of natural selection under varying population size from the site frequency spectrum and haplotype structure. *The Royal Society*, **286(1896)**:8

NOAA BC National Geophysical Data Center (1988). Data Announcement 88-MGG-02, , Digital relief of the Surface of the Earth. Publisher: U.S. Department of Commerce

Novembre J, Galvani AP, & Slatkin M (2005). The geographic spread of the CCR5 δ32 HIV-resistance allele. *PLoS Biol*, **3(11)**:e339

Novembre J & Han E (2012). Human population structure and the adaptive response to pathogen-induced selection pressures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367(1590)**:878–886

Okubo A *et al.* (1980). Diffusion and ecological problems: mathematical models

Othmer HG, Dunbar SR, & Alt W (1988). Models of dispersal in biological systems. *Journal of mathematical biology*, **26(3)**:263–298

Painter KJ & Hillen T (2018). From random walks to fully anisotropic diffusion models for cell and animal movement. In *Cell Movement*, pp. 103–141. Springer

Peter BM, Huerta-Sanchez E, & Nielsen R (2012). Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLOS Genetics*, **8(10)**:e1003011. 10.1371/journal.pgen.1003011. Publisher: Public Library of Science

Petr M (2021). bodkan/slendr. Original-date: 2021-02-18T15:07:15Z

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, *et al.* (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, **19(5)**:826–837

Platt A, Pivirotto A, Knoblauch J, & Hey J (2019). An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS genetics*, **15(8)**:e1008340

Pritchard JK & Di Rienzo A (2010). Adaptation–not by sweeps alone. *Nature Reviews Genetics*, **11(10)**:665–667

Racimo F, Sikora M, Vander Linden M, Schroeder H, & Lalueza-Fox C (2020a). Beyond broad strokes: sociocultural insights from the study of ancient genomes. *Nature Reviews Genetics*, **21(6)**:355–366

Racimo F, Woodbridge J, Fyfe RM, Sikora M, Sjögren KG, Kristiansen K, & Vander Linden M (2020b). The spatiotemporal spread of human migrations during the european holocene. *Proceedings of the National Academy of Sciences*, **117(16)**:8989–9000

Ralph P & Coop G (2010). Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*, **186(2)**:647–668

Ray N & Excoffier L (2009). Inferring past demography using spatially explicit population genetic models. *Human Biology*, **81(3)**:141–157

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449(7164)**:913–918

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N, *et al.* (2005). The Case for Selection at CCR5-32. *PLoS Biology*, **3(11)**. 10.1371/journal.pbio.0030378

Ségurel L & Bon C (2017). On the evolution of lactase persistence in humans. *Annual review of genomics and human genetics*, **18**

Segurel L, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, Fabre A, Hegay T, & Heyer E (2020). Why and when was lactase persistence selected for? insights from central asian herders and ancient dna. *PLoS Biology*, **18(6)**:e3000742

Slatkin M & Rannala B (2000). Estimating allele age. *Annual review of genomics and human genetics*, **1(1)**:225–249

Soetaert K & Meysman F (2012). Reactive transport in aquatic ecosystems: Rapid model prototyping in the open source software r. *Environmental Modelling Software*, **32**:49–60

Soetaert K, Petzoldt T, & Setzer RW (2010). Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, **33(9)**:1–25. 10.18637/jss.v033.i09

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, *et al.* (1998). Dating the origin of the CCR5-δ32 AIDS-resistance allele by the coalescence of haplotypes. *The American Journal of Human Genetics*, **62(6)**:1507–1515

Stern AJ, Wilton PR, & Nielsen R (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genetics*, **15(9)**:e1008384. 10.1371/journal.pgen.1008384

Szpak M, Xue Y, Ayub Q, & Tyler-Smith C (2019). How well do we understand the basis of classic selective sweeps in humans? *FEBS letters*, **593(13)**:1431–1448

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, *et al.* (2007). Convergent adaptation of human lactase persistence in africa and europe. *Nature genetics*, **39(1)**:31–40

Voight BF, Kudaravalli S, Wen X, & Pritchard JK (2006). A map of recent positive selection in the human genome. *PLoS Biol*, **4(3)**:e72

Wikle CK, Zammit-Mangion A, & Cressie N (2019). *Spatio-temporal Statistics with R*. CRC Press

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, *et al.* (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*, **111(13)**:4832–4837. 10.1073/pnas.1316513111

Appendix

Here, we motivate the construction of model C as a large scale limit of a random walk model on a lattice (Karlin & Taylor, 1975; Cantrell & Cosner, 2004). We think of the allele frequency as a variable p that can increase in magnitude due to its inherent advantage (selection), spread across a landscape (diffusion) or move directionally as a consequence of migration (advection). We imagine a lattice composed of small square cells of size Δx x Δy, where a certain amount of allele frequency p can occur at a given time point t. At each small time step (of duration Δt), inflow and outflow of p can occur in the x-direction with probability h or in the y-direction with probability 1-h, and the magnitude of these flows depend on the amount of p present in neighboring cells. If flow of p is along the x-axis, it does so in the positive direction with probability α and in the negative direction with probability $1-\alpha$. If flow of p is along the y-axis, it does so in the positive direction with probability β and in the negative direction with probability $1-\beta$. The allele frequency can also increase in magnitude locally, via a function $\gamma()$ that depends on its dominance (d), selection coefficient (s) and current magnitude ($p(x,y,t)$). Then, we obtain:

$$\begin{aligned} p(x,y,t+\Delta t) = {} & p(x,y,t) + \gamma(p(x,y,t),s,d)\Delta t + \\ & h\alpha p(x-\Delta x,y,t) + h(1-\alpha)p(x+\Delta x,y,t) + \\ & (1-h)\beta p(x,y-\Delta y,t) + (1-h)(1-\beta)p(x,y+\Delta,y,t) \end{aligned} \tag{8}$$

We can also write this as:

$$\begin{aligned} p(x,y,t+\Delta t) - p(x,y,t) = {} & h\left(\frac{1}{2}-\alpha\right)[p(x+\Delta x,y,t) - p(x-\Delta x,y,t)] + \\ & (1-h)\left(\frac{1}{2}-\beta\right)[p(x,y+\Delta y,t) - p(x,y-\Delta y,t)] + \\ & h\frac{1}{2}[p(x+\Delta x,y,t) - 2p(x,y,t) + p(x+\Delta x,y,t)] + \\ & (1-h)\frac{1}{2}[p(x,y+\Delta y,t) - 2p(x,y,t) + p(x,y+\Delta y,t)] + \\ & \gamma(p(x,y,t),s,d)\Delta t \end{aligned} \tag{9}$$

If we divide both sides by Δt and take the limit of infinitesimally small Δx, Δy and Δt, while assuming that, in this limit, $\frac{\Delta x^2}{\Delta t}$ and $\frac{\Delta y^2}{\Delta t}$ are finite (Okubo *et al.*, 1980), we obtain:

$$\frac{\partial p}{\partial t} = \frac{1}{2}h\lambda_x\frac{\partial^2 p}{\partial x^2} + \frac{1}{2}(1-h)\lambda_y\frac{\partial^2 p}{\partial y^2} + h(1-2\alpha)u_x\frac{\partial p}{\partial x} + (1-h)(1-2\beta)u_y\frac{\partial p}{\partial y} + \gamma(p(x,y,t),s,d) \tag{10}$$

where $u_x = \frac{\Delta x}{\Delta t}$, $u_y = \frac{\Delta y}{\Delta t}$, $\lambda_x = \frac{\Delta x^2}{\Delta t}$, $\lambda_y = \frac{\Delta y^2}{\Delta t}$.

If we let $\sigma_x^2 = h\lambda_x$, $\sigma_y^2 = (1-h)\lambda_y$, $v_x = h(1-2\alpha)u_x$, $v_y = (1-h)(1-2\beta)u_y$, then we obtain equation 4. Thus, we can see that the squared diffusion coefficient σ_x^2 depends on the square of the length of the cells in the x-axis relative to the duration of a time step (λ_x), and on the probability that flows occurs in the x-axis at a given time step (h). Similarly, the squared diffusion coefficient σ_y^2 depends on the square of the length of the cells in the y-axis relative to the duration of a time step (λ_y), and on the probability that flows occurs in the y-axis at a given time step $(1 - h)$. The advection coefficient v_x depends on the advective velocity along the x-axis (u_x) as well as on the probability of flow occurring along the x-axis (h) and the directional bias $1 - 2\alpha$, which depends on the probability that flow occurs in the positive x-direction (α). Finally, the advection coefficient v_y depends on the advective velocity along the y-axis (u_y) as well as on the probability of flow occurring along the y-axis $(1 - h)$ and the directional bias $1 - 2\beta$, which depends on the probability that flow occurs in the positive y-direction (β).

We can recover model B as a special case of model C if we fix $\alpha = \beta = \frac{1}{2}$, assuming isotropy in the two directions, so that $\Delta x = \Delta y$. We can also recover model A if we additionally fix $h = \frac{1}{2}$.

Supplementary figures

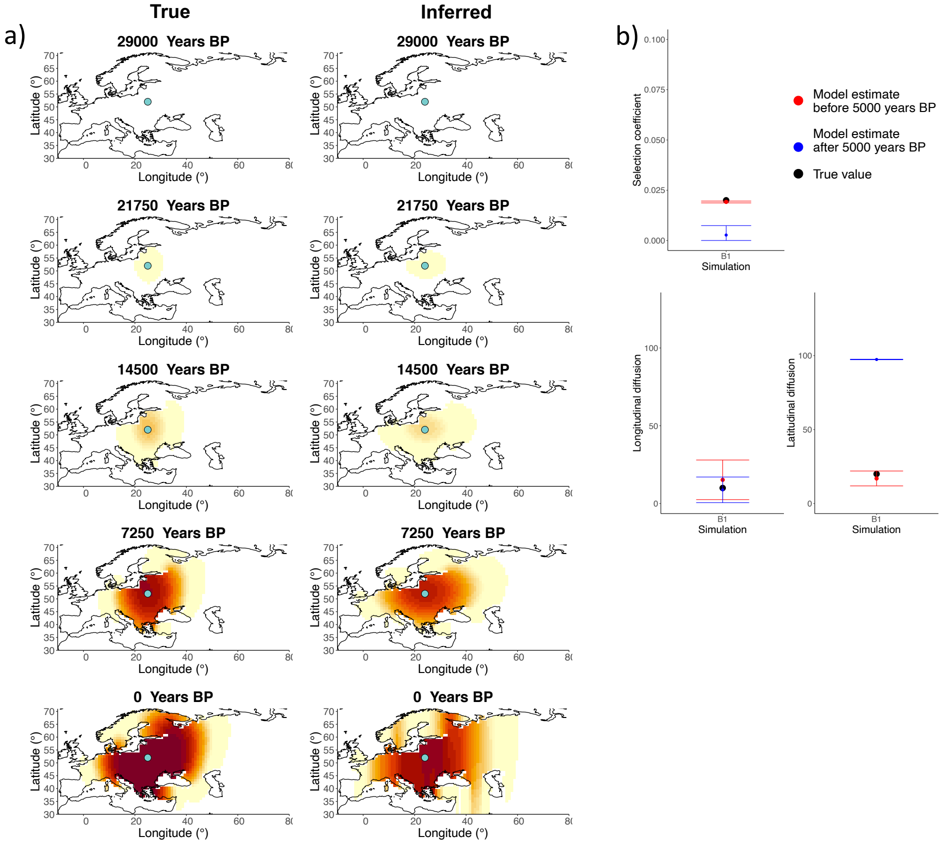


Figure S1: a) Comparison of true and inferred allele frequency dynamics for simulation B1. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S1. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

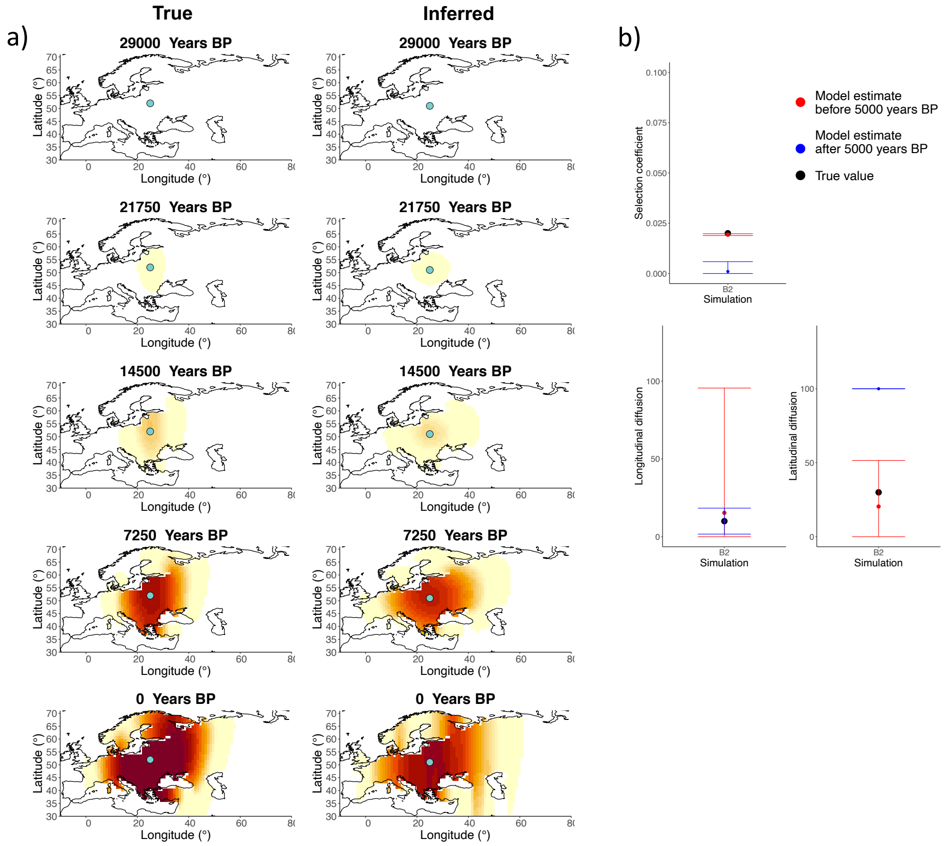


Figure S2: a) Comparison of true and inferred allele frequency dynamics for simulation B2. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S1. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

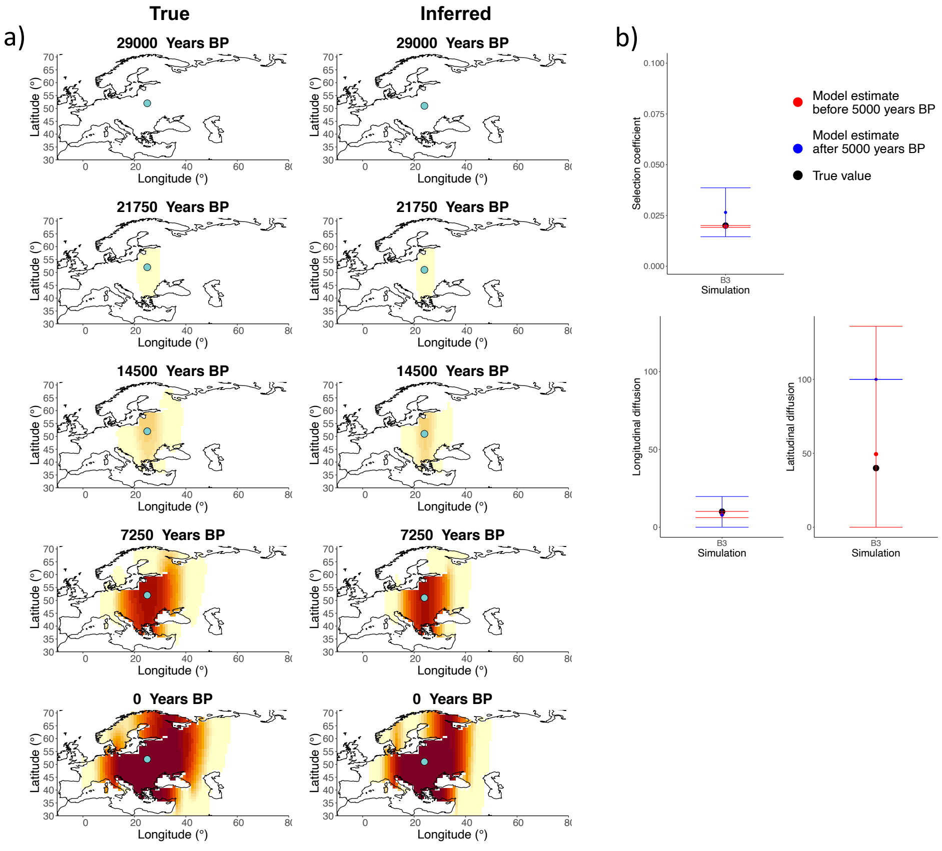


Figure S3: a) Comparison of true and inferred allele frequency dynamics for simulation B3. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S1. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

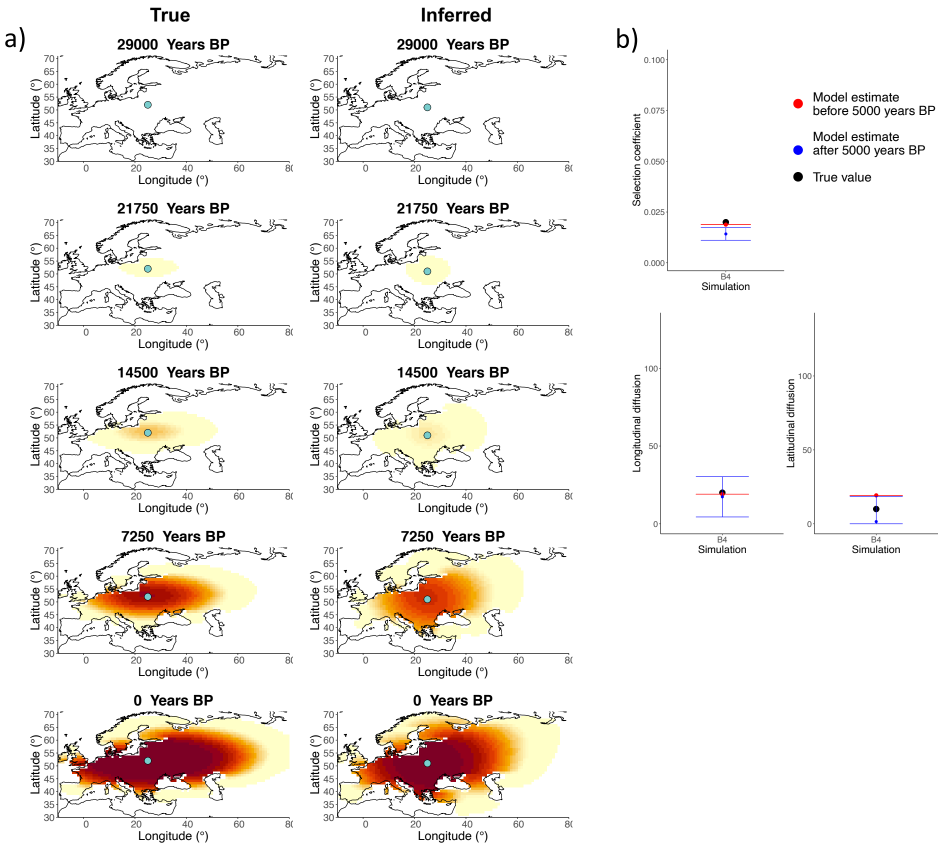


Figure S4: a) Comparison of true and inferred allele frequency dynamics for simulation B4. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S1. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

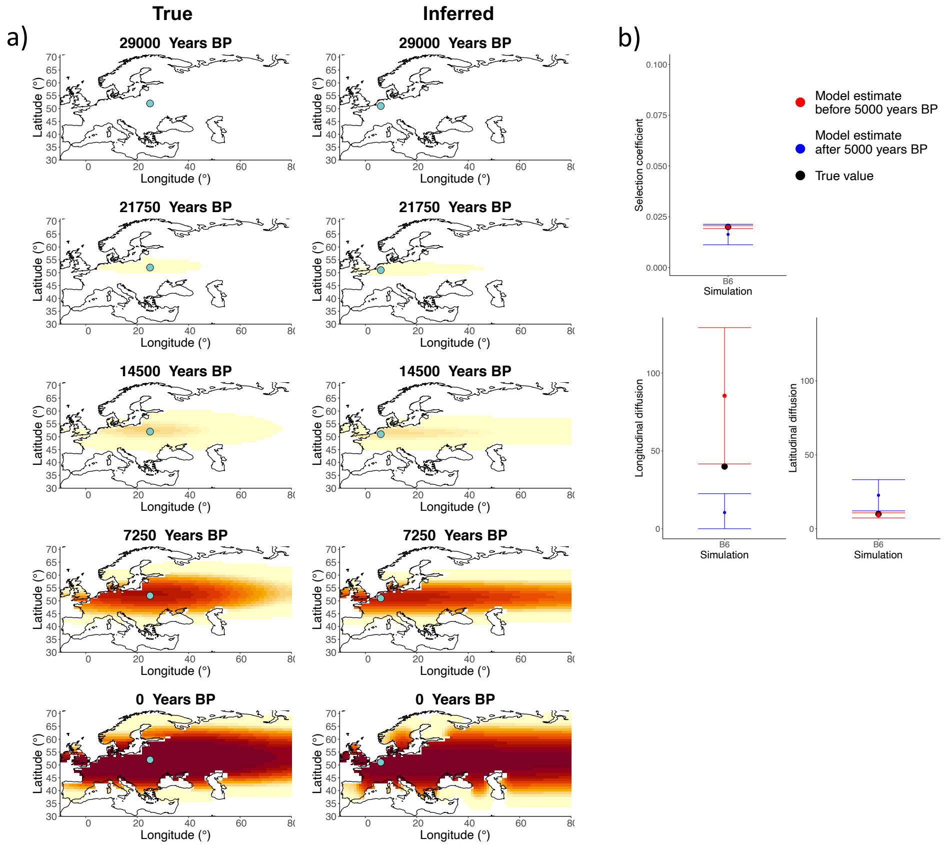


Figure S5: a) Comparison of true and inferred allele frequency dynamics for simulation B6. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S1. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

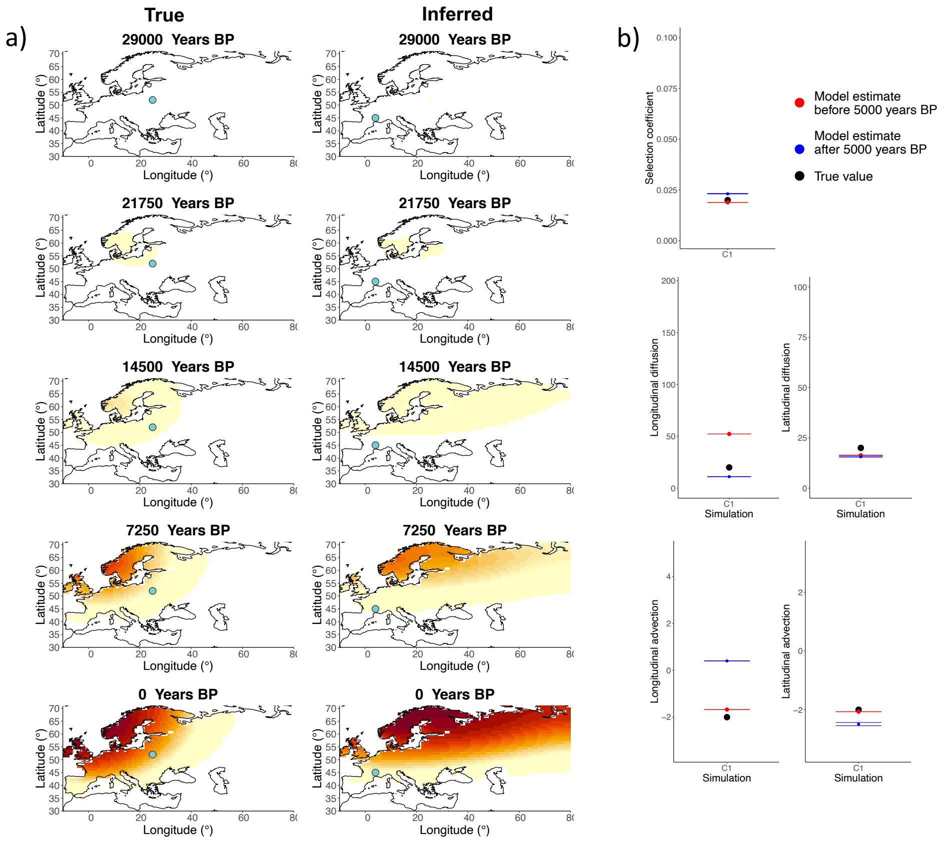


Figure S6: a) Comparison of true and inferred allele frequency dynamics for one of the simulations including advection (C1). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S2. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

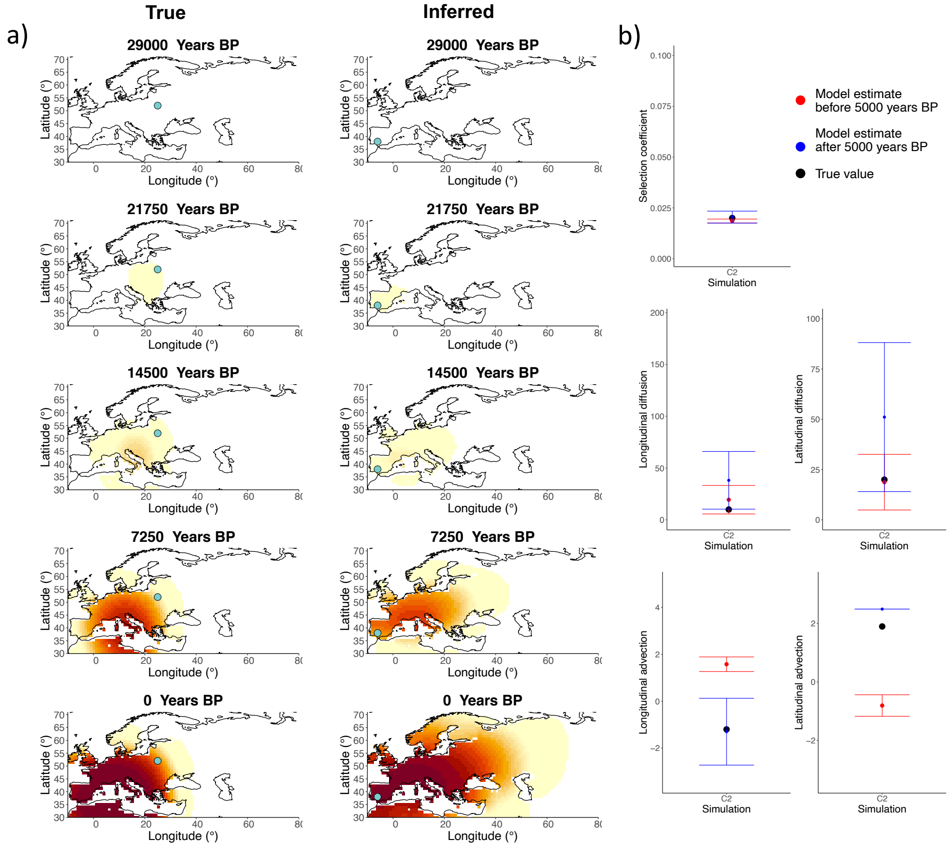


Figure S7: a) Comparison of true and inferred allele frequency dynamics for one of the simulations including advection (C2). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S2. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

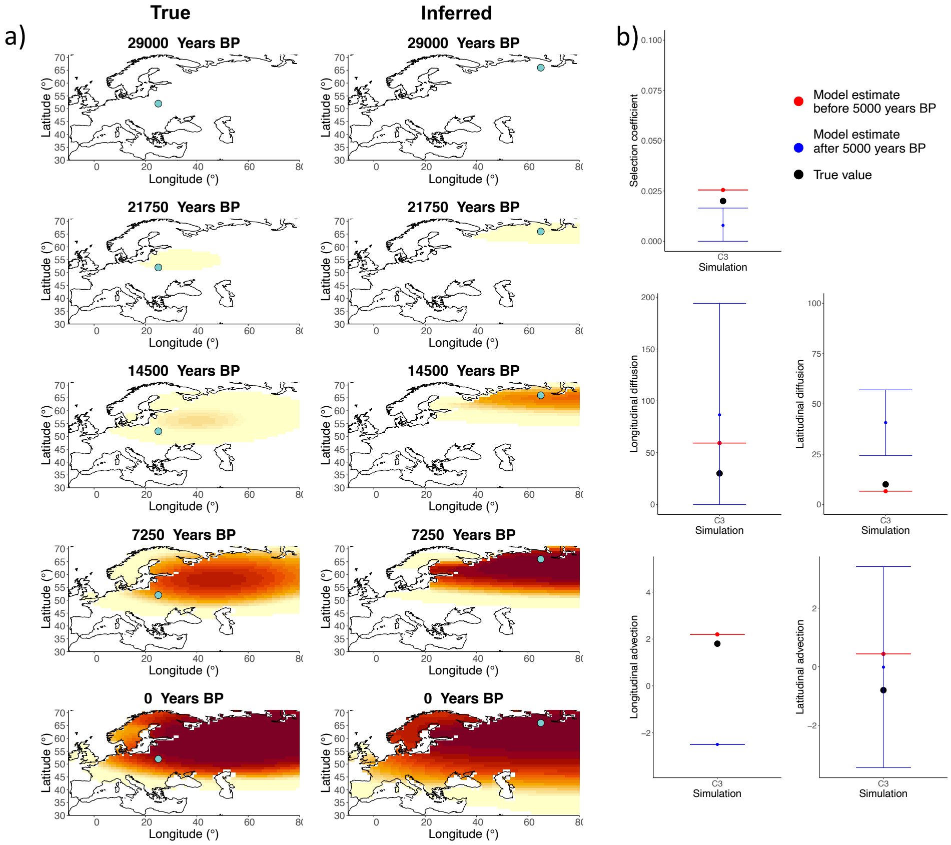


Figure S8: a) Comparison of true and inferred allele frequency dynamics for one of the simulations including advection (C3). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in Table S2. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

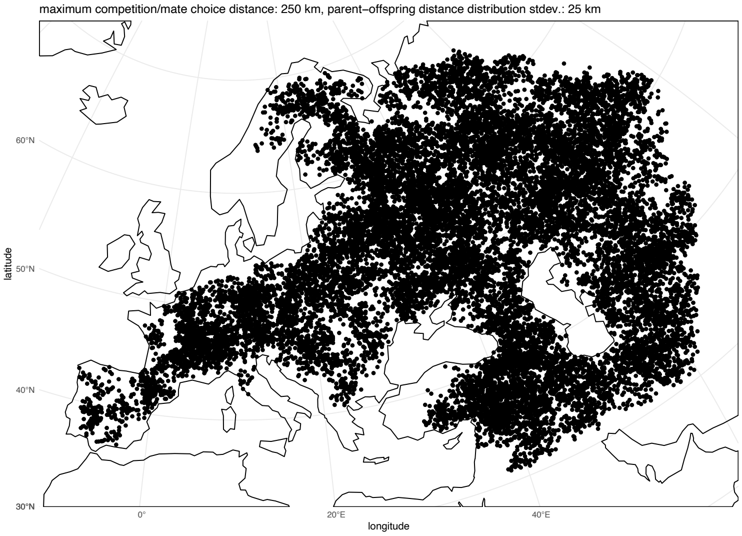

maximum competition/mate choice distance: 250 km, parent–offspring distance distribution stdev.: 25 km
60°N
50°N
40°N
30°N
latitude
0°
20°E
40°E
longitude

Figure S9: Distribution of individuals across the map under neutrality, showing the tendency of individuals to cluster together.

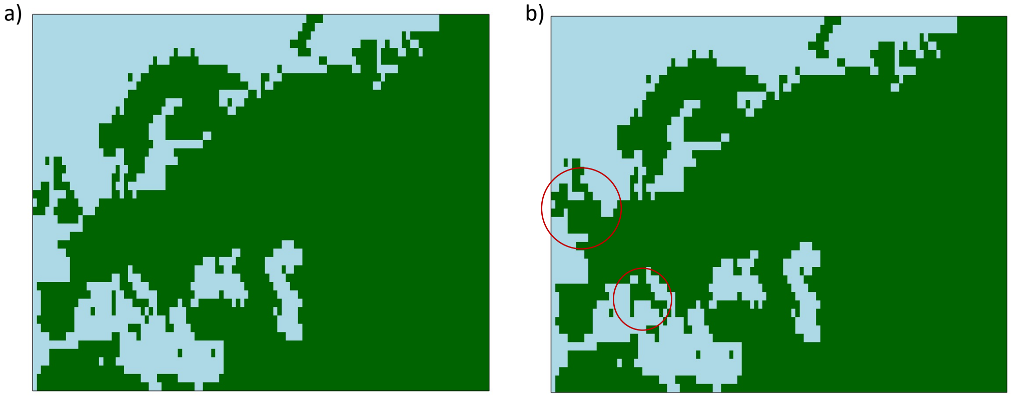


Figure S10: Maps showing areas where diffusion in the model is allowed (green) and where it is forbidden (blue). Figure a) map without land bridges. Figure b) map containing land bridges indicated with red circles.

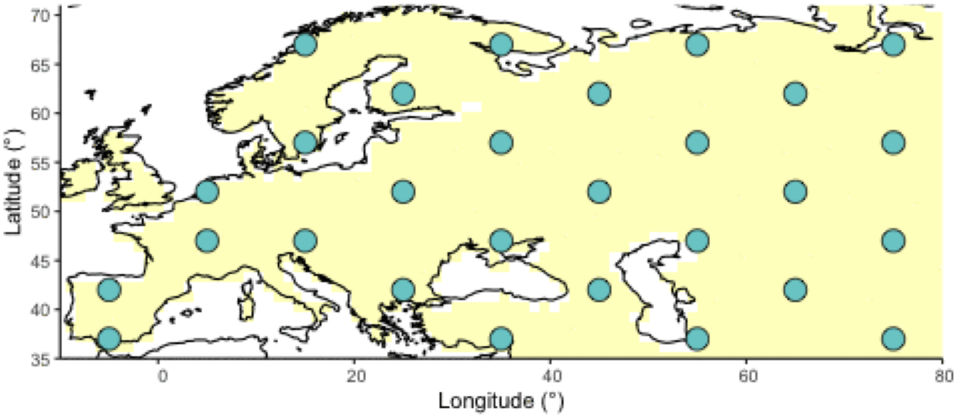


Figure S11: Geographic locations for points used as potential origins of the allele at the initialization of the simulated annealing optimization algorithm. Note that, after initialization, the algorithm can continuously explore any points on the map grid that are not necessarily included in this point set

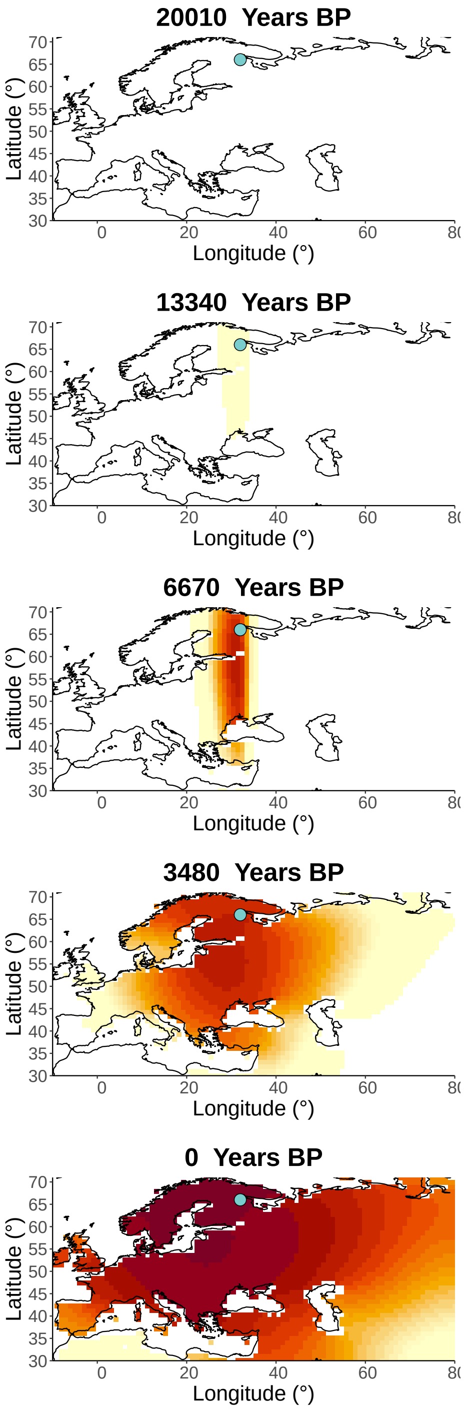


Figure S12: Inferred frequency dynamics of rs4988235(T) using the allele age that was inferred in Albers & McVean (2020).

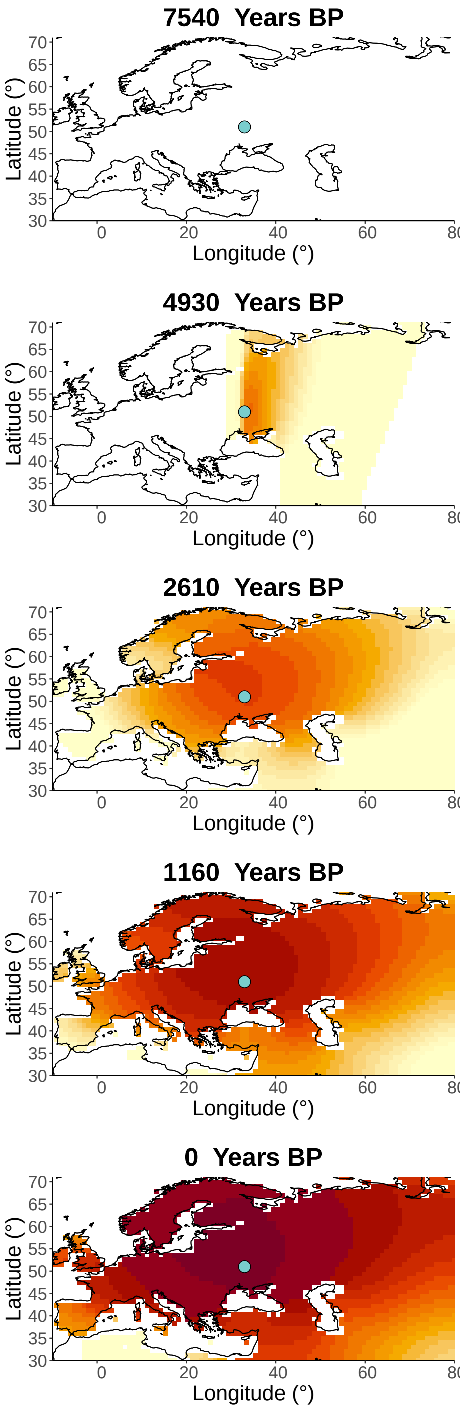
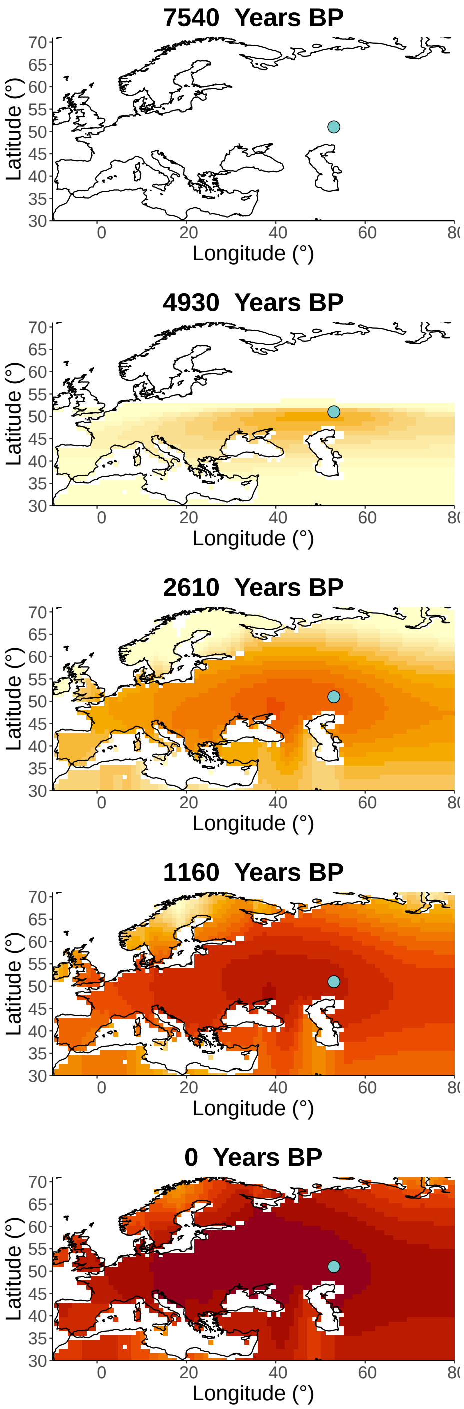


Figure S13: Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees west from the original estimate.



7540 Years BP

4930 Years BP

2610 Years BP

1160 Years BP

0 Years BP

Figure S14: Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees east from the original estimate.

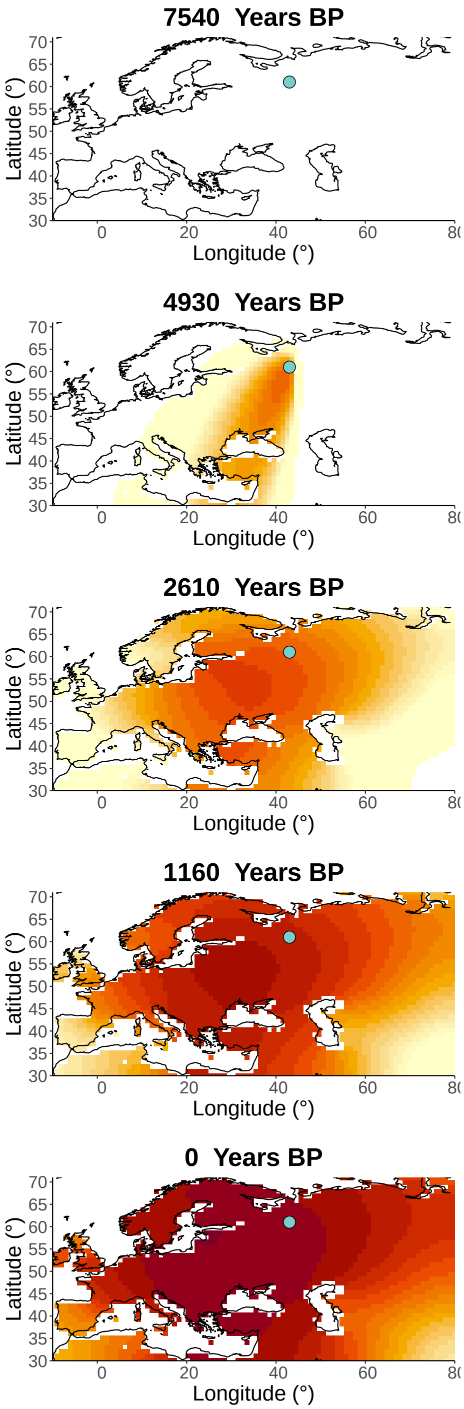
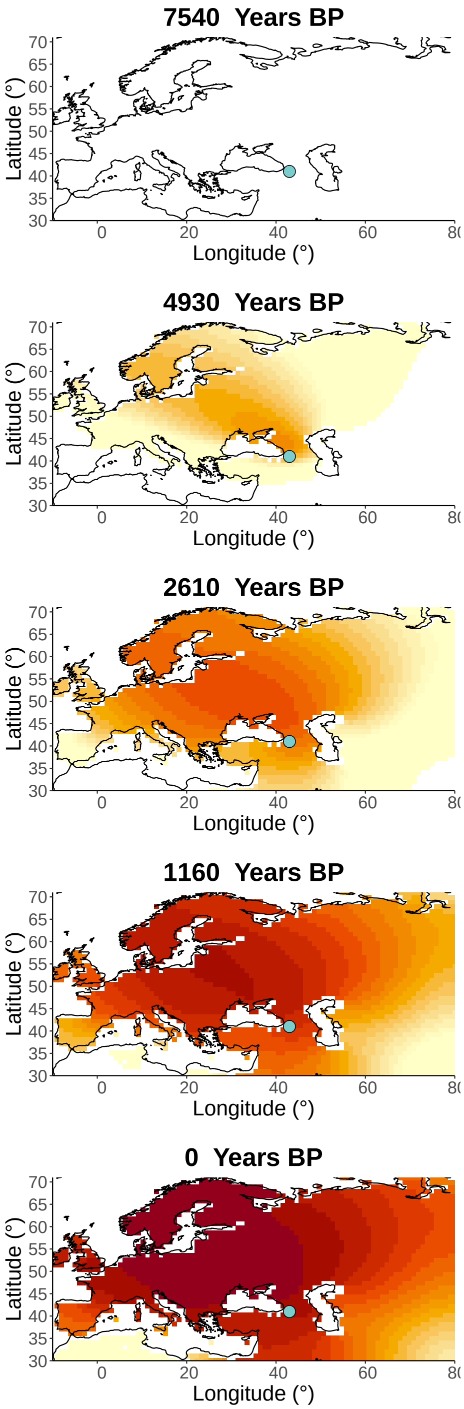


Figure S15: Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees north from the original estimate.


7540 Years BP

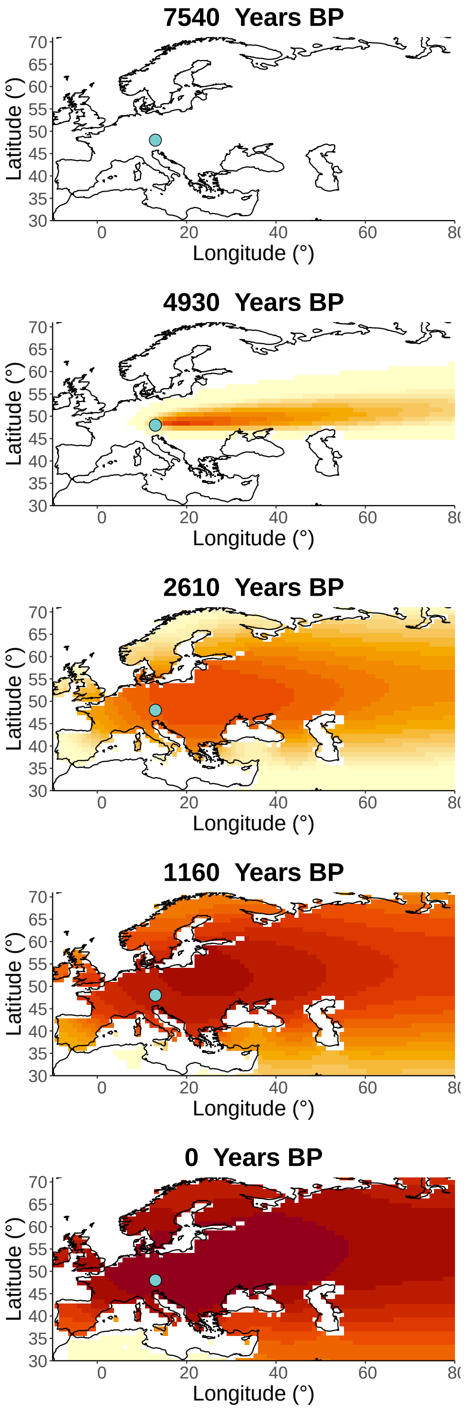
4930 Years BP

2610 Years BP

1160 Years BP

0 Years BP

Figure S16: Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees south from the original estimate.



7540 Years BP

4930 Years BP

2610 Years BP

1160 Years BP

0 Years BP

Figure S17: Inferred frequency dynamics of rs4988235(T) forcing the geographic origin of the allele to be at the location inferred in Itan *et al.* (2009).

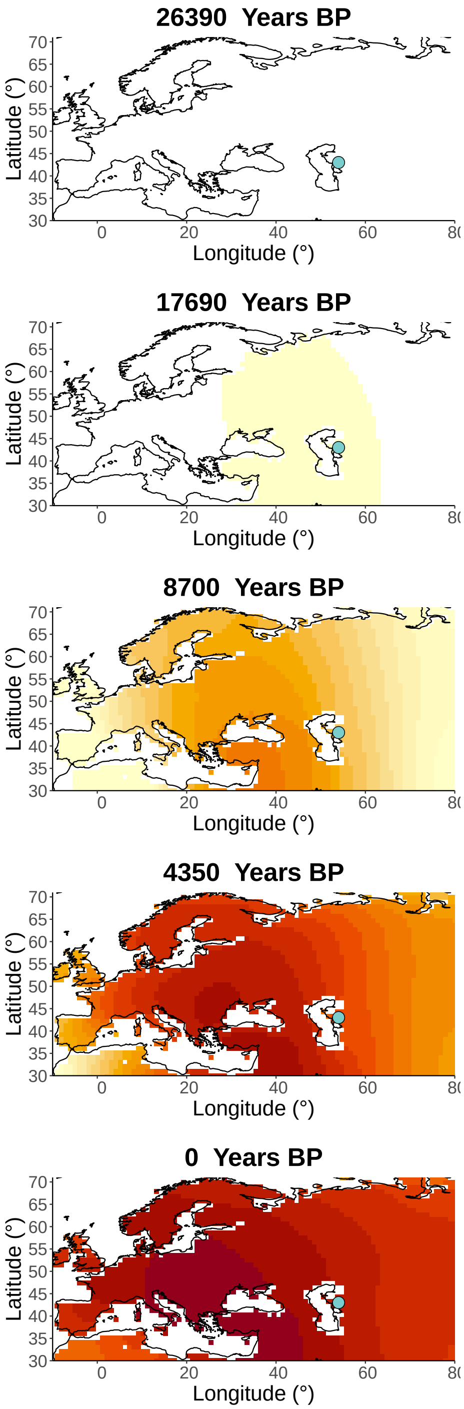


Figure S18: Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees east from the original estimate.

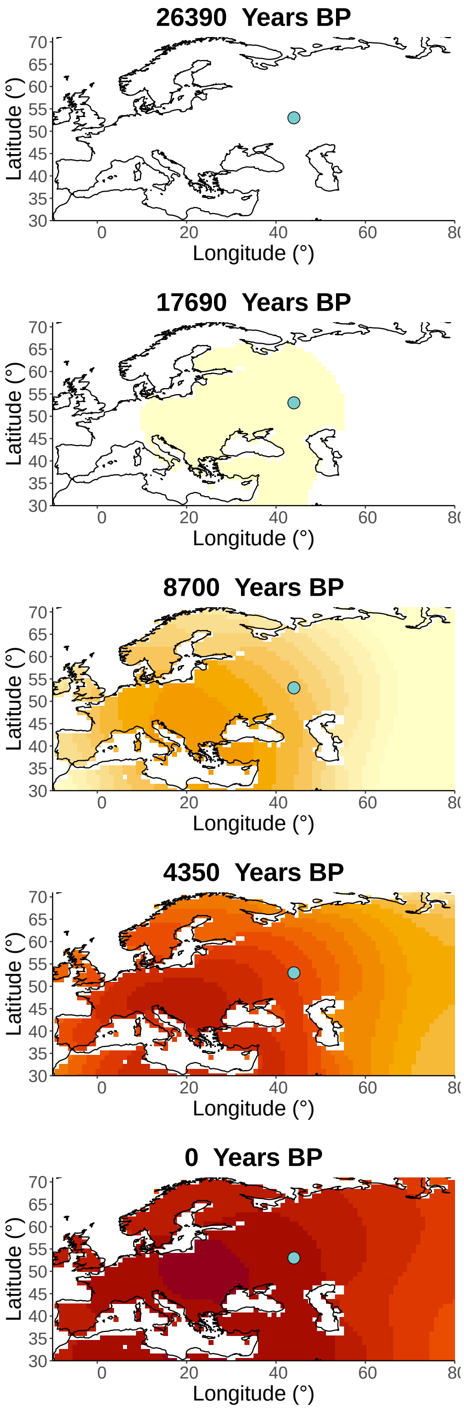


Figure S19: Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees north from the original estimate.

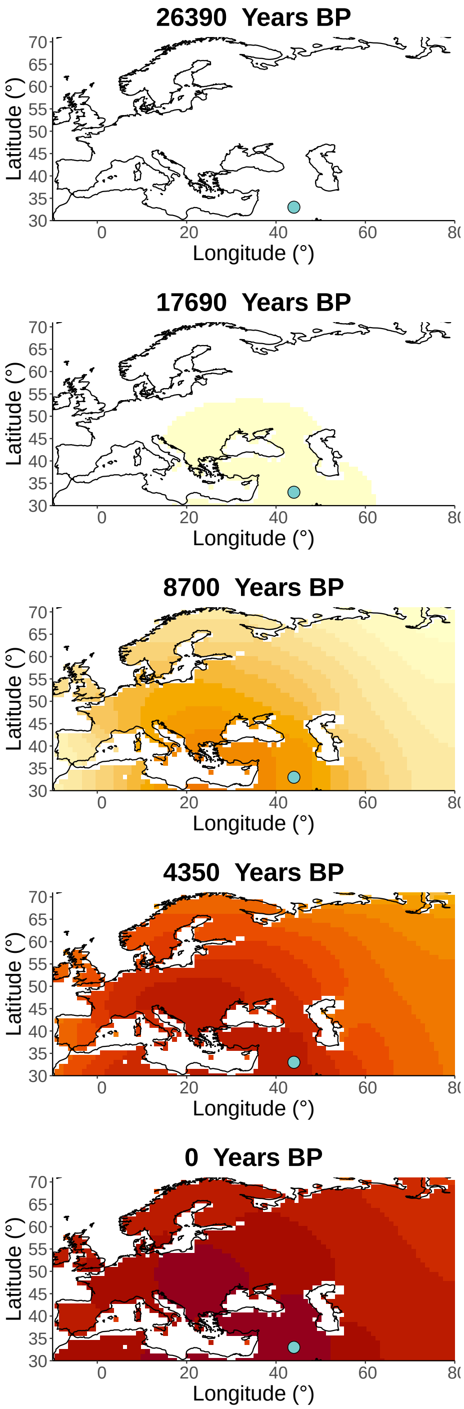


Figure S20: Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees south from the original estimate.

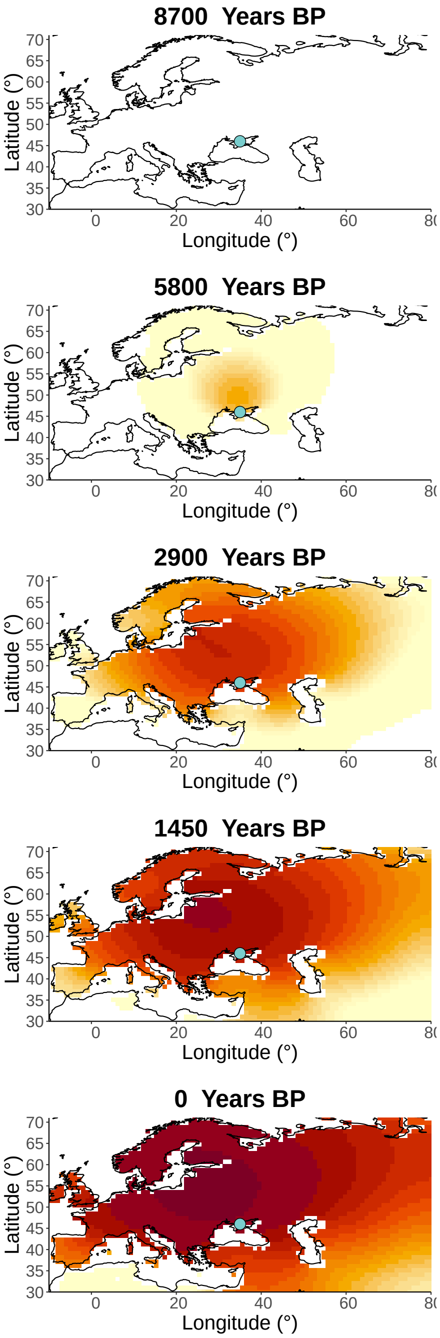


Figure S21: Inferred frequency dynamics of rs4988235(T) assuming the allele age to be the lower end of the 95% credible interval for the allele age inferred in Itan *et al.* (2009).

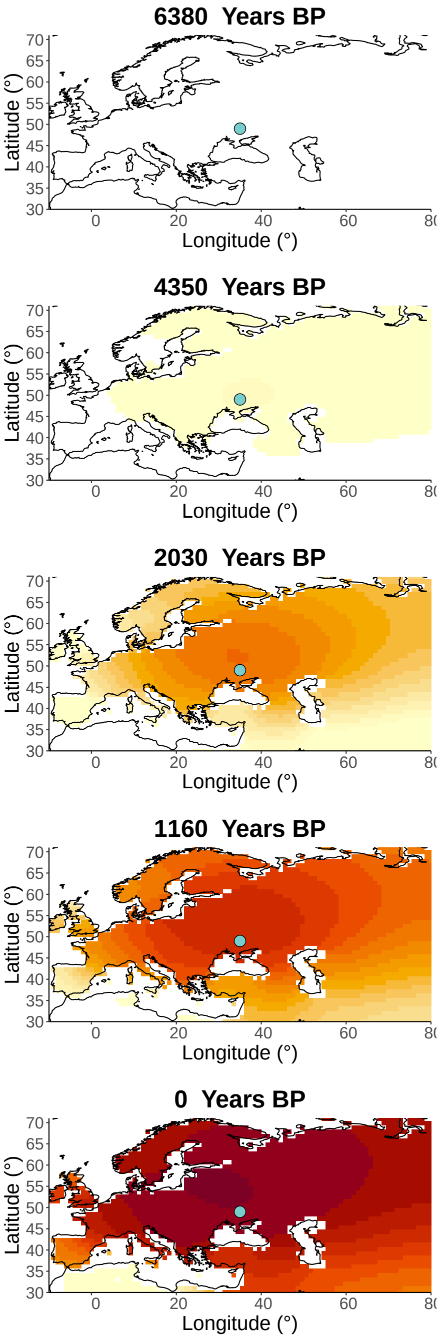


Figure S22: Inferred frequency dynamics of rs4988235(T) assuming the allele age to be the higher end of the 95% credible interval for the allele age inferred in Itan *et al.* (2009).

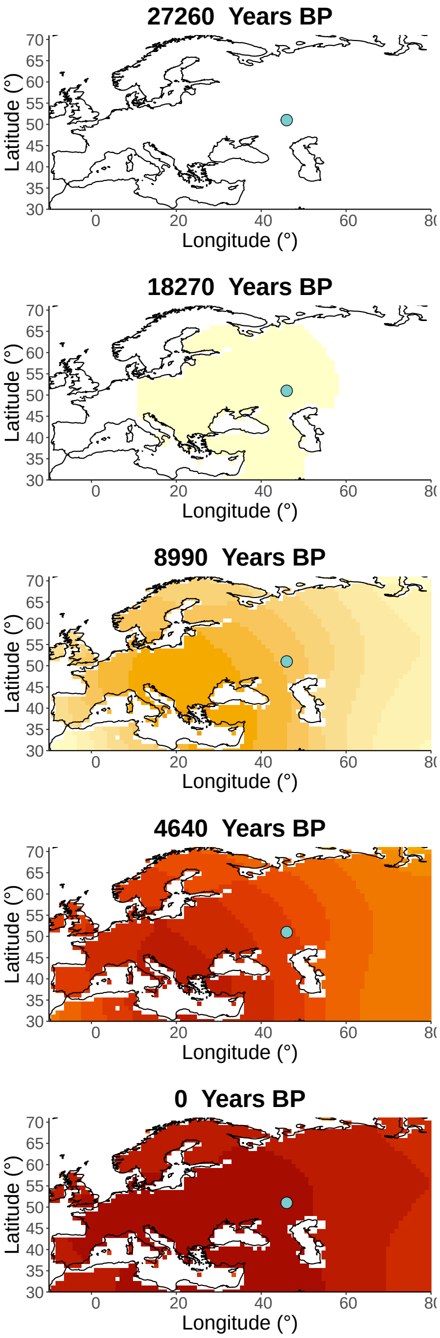


Figure S23: Inferred frequency dynamics of rs1042602(A) assuming the allele age to be the lower end of the 95% confidence interval for the allele age inferred in Albers & McVean (2020).

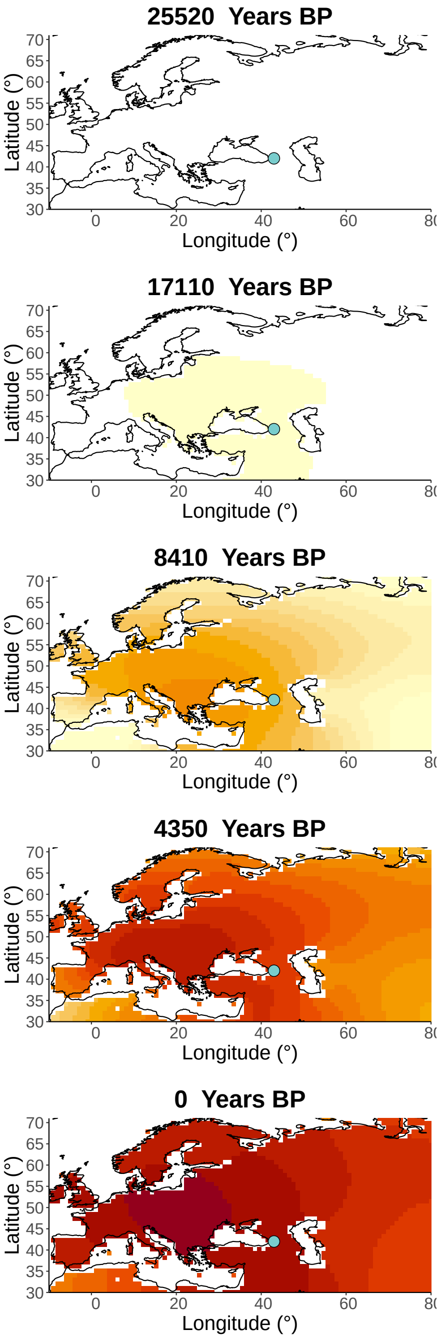


Figure S24: Frequency dynamics of rs1042602(A) assuming the allele age to be the higher end of the 95% confidence interval for the allele age inferred in Albers & McVean (2020).

Supplementary tables

Simulation		s	σ_x (km/gen)	σ_y (km/gen)	Long	Lat
		True/Pred (95% CI)	True/Pred (95% CI)	True/Pred (95% CI)	True/Pred	True/Pred
B1	Sample age >5000	0.02/0.0192 (0.0187–0.0196)	10/15.244 (2.5042–27.9828)	20/16.963 (11.9263–21.9993)	25/24	52/52
	Sample age <5000	0.02/0.0027 (0–0.0074)	10/8.805 (0.5631–17.0468)	20/97.432 (97.2566–97.6081)	–	–
B2	Sample age >5000	0.02/0.0193 (0.0189–0.0198)	10/15.348 (0–95.5192)	30/20.427 (0–51.514)	25/25	52/51
	Sample age <5000	0.02/0.001 (0–0.0059)	10/10.015 (1.6837–18.3472)	30/100 (99.9933–100.0067)	–	–
B3	Sample age >5000	0.02/0.0196 (0.0191–0.02)	10/8.149 (6.1551–10.143)	40/49.432 (0–135.9428)	25/24	52/51
	Sample age <5000	0.02/0.0265 (0.0145–0.0386)	10/7.855 (0–19.751)	40/100 (99.9933–100.0067)	–	–
B4	Sample age >5000	0.02/0.0188 (0.0188–0.0188)	20/19.037 (19.0311–19.0435)	10/19.254 (19.2439–19.2638)	25/25	52/51
	Sample age <5000	0.02/0.0142 (0.0111–0.0173)	20/17.354 (4.4083–30.2991)	10/1.489 (0–18.5993)	–	–
B5	Sample age >5000	0.02/0.0196 (0.019–0.0202)	30/26.409 (11.1997–41.6184)	10/11.429 (7.1825–15.6759)	25/27	52/51
	Sample age <5000	0.02/0.0215 (0.0183–0.0246)	30/14.3 (0–68.176)	10/1.554 (0–16.5985)	–	–
B6	Sample age >5000	0.02/0.0199 (0.0192–0.0206)	40/85.415 (41.6058–129.2248)	10/9.02 (7.2853–10.7538)	25/6	52/51
	Sample age <5000	0.02/0.0163 (0.0112–0.0213)	40/10.403 (0–22.533)	10/22.623 (12.0841–33.1614)	–	–

Table S1: Parameter values used to generate simulations using numerical solutions to equation 3, compared to parameter estimates assuming model B. The age of the allele was set to 29,000 years in all simulations. Columns named ”Long” and ”Lat” indicate the longitude and latitude of the geographic origin of the allele, respectively.

Simulation		s	σ_x (km/gen)	σ_y (km/gen)	v_x (km/gen)	v_y (km/gen)	Long	Lat
		True/Pred (95% CI)	True/Pred (95% CI)	True/Pred (95% CI)	True/Pred (95% CI)	True/Pred (95% CI)	True/Pred	True/Pred
C1	Sample age >5000	0.02/0.0189 (0.0188–0.0189)	20/52.246 (52.2051–52.2872)	20/16.373 (16.332–16.4139)	-2/-1.675 (-1.6771–-1.6722)	-2/-2.067 (-2.0702–-2.0639)	25/4	52/45
	Sample age <5000	0.02/0.0231 (0.023–0.0233)	20/11.086 (10.9286–11.2441)	20/15.606 (15.3659–15.8467)	-2/0.399 (0.3946–0.4037)	-2/-2.491 (-2.5458–-2.436)	–	–
C2	Sample age >5000	0.02/0.0185 (0.0176–0.0195)	10/19.434 (5.6736–33.1952)	20/18.727 (4.8938–32.5605)	-1.2/1.579 (1.2671–1.8905)	1.9/-0.801 (-1.1684–-0.4331)	25/-6	52/38
	Sample age <5000	0.02/0.0205 (0.0175–0.0234)	10/38.144 (10.3123–65.9749)	20/51.094 (14.0489–88.1388)	-1.2/-1.299 (-2.7247–0.1266)	1.9/2.493 (2.4929–2.4933)	–	–
C3	Sample age >5000	0.02/0.0255 (0.0254–0.0256)	30/59.237 (59.1269–59.347)	10/6.604 (6.5991–6.6087)	1.8/2.195 (2.1918–2.1985)	-0.8/0.438 (0.4381–0.4387)	25/65	52/66
	Sample age <5000	0.02/0.0079 (0–0.0165)	30/86.511 (0–194.0772)	10/40.693 (24.3946–56.9905)	1.8/-2.498 (-2.4983–-2.498)	-0.8/-0.014 (-3.4481–3.4204)	–	–
C4	Sample age >5000	0.02/0.0197 (0.0191–0.0204)	10/19.647 (14.975–24.3197)	10/13.585 (0–27.2936)	1.2/-0.054 (-0.0968–-0.0111)	1/0.72 (0.4278–1.0124)	25/44	52/50
	Sample age <5000	0.02/0.0137 (0.0046–0.0229)	10/14.151 (0–32.1031)	10/4.093 (0–49.4651)	1.2/0.8 (-3.4903–5.0895)	1/2.434 (2.4299–2.4387)	–	–

Table S2: Parameter values used to generate simulations using numerical solutions to equation 4, compared to parameter estimates assuming model C. The age of the allele was set to 29,000 years in all simulations. Columns named ”Long” and ”Lat” indicate the longitude and latitude of the geographic origin of the allele, respectively.

s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat	Allele age (years)
0.0366 (0.0357–0.0375)	58.583 (49.1983–67.9669)	63.733 (3.6601–123.8056)	-0.436 (-0.8077–-0.0649)	-1.564 (-3.0915–-0.0355)	15	47	15000

Table S3: Parameter values estimated using model C for the forward simulation created using SLiM. Columns named ”Long” and ”Lat” indicate the longitude and latitude of the geographic origin of the allele, respectively.

	s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat	Allele age (years)
Sample age >5000	0.0993 (0.0993–0.0993)	20.293 (15.5643–25.0226)	15.642 (9.9963–21.2871)	-0.575 (-0.8055–-0.3446)	0.435 (0.319–0.5512)	43	51	7441
Sample age <5000	0.0328 (0.0327–0.0329)	94.901 (94.2585–95.5435)	85.612 (84.6975–86.526)	-1.211(-1.2197–-1.2019)	-2.5 (-2.5136–-2.4855)			
Sample age >5000	0.0867 (0.0866–0.0867)	24.27 (24.2658–24.2734)	28.328 (28.3234–28.3326)	-0.398 (-0.3985–-0.3984)	-2.055 (-2.0562–-2.0547)	35	46	8683
Sample age <5000	0.0321 (0.0319–0.0323)	97.325 (97.1434–97.5061)	87.416 (85.6745–89.1578)	-2.5 (-2.5–-2.4997)	-2.389 (-2.3935–-2.3845)			
Sample age >5000	0.0994 (0.0994–0.0994)	22.92 (15.0004–30.8397)	17.884 (13.8709–21.8967)	0.327 (0.1726–0.4818)	-0.295 (-0.3678–-0.2229)	35	49	6256
Sample age <5000	0.0572 (0.057–0.0574)	95.014 (93.6242–96.4032)	85.249 (82.9662–87.5322)	-2.499 (-2.4992–-2.4989)	-1.679 (-1.7919–-1.5658)			

Table S4: Summary of parameter estimates for rs4988235(T). The upper two rows correspond to results obtained assuming the allele age to be the point estimate from Itan *et al.* (2009): 7,441 years ago. The middle two rows and the bottom two rows show results assuming the age to be either the lower or the higher ends of the allele age's 95% confidence interval from Itan *et al.* (2009). Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

	s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat	Allele age (years)
Sample age >5000	0.0285 (0.0285–0.0285)	1.25 (1.2492–1.25)	44.619 (44.5944–44.6445)	-0.177 (-0.1773–-0.1771)	1.925 (1.9247–1.9262)	32	66	20106
Sample age <5000	0.0255 (0.0252–0.0258)	92.545 (91.6963–93.3941)	87.545 (85.3525–89.7369)	-2.499 (-2.4992–-2.4989)	-2.271 (-2.4127–-2.1297)			

Table S5: Parameter estimates for rs4988235(T) using the allele age inferred in Albers & McVean (2020). Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

	s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat	Allele age (years)
Sample age >5000	0.0221 (0.0216–0.0227)	71.668 (24.7274–118.6092)	50.434 (25.6535–75.2136)	-2.268 (-3.006–-1.5304)	-0.486 (-0.8661–-0.1053)	44	43	26367
Sample age <5000	0.0102 (0.0083–0.012)	69.25 (14.0247–124.4756)	95.281 (95.1087–95.453)	0.849 (-0.0783–1.7769)	-0.503 (-0.929–-0.076)			
Sample age >5000	0.0214 (0.0205–0.0223)	57.914 (0–131.3177)	83.846 (0–246.6688)	-2.111 (-2.8784–-1.3429)	1.305 (-0.8411–3.4519)	46	51	27315
Sample age <5000	0.01 (0.0078–0.0121)	88.218 (0–190.105)	96.216 (96.0422–96.3898)	1.19 (-0.7489–3.1293)	-0.88 (-2.0897–0.3299)			
Sample age >5000	0.023 (0.023–0.0231)	75.857 (75.8065–75.9071)	48.992 (48.9166–49.0674)	-2.362 (-2.3655–-2.3593)	-0.837 (-0.8371–-0.8362)	43	42	25424
Sample age <5000	0.0099 (0.0085–0.0112)	72.847 (67.7991–77.8949)	92.867 (75.4925–110.2412)	0.497 (0.2717–0.7214)	-0.685 (-0.8076–-0.5628)			

Table S6: Summary of parameter estimates for rs1042602(A). Upper two rows corresponds to model fit when allele age is set to be the point estimate Albers & McVean (2020): 26,367 years ago. The middle two rows and the bottom two rows show results assuming the age to be either the lower or the higher ends of the allele age's 95% confidence interval from Albers & McVean (2020). Columns named ”Long” and ”Lat” indicate the longitude and latitude of the geographic origin of the allele, respectively.

	s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat
Sample age >5000	0.0993 (0.0993–0.0993)	20.293 (15.5643–25.0226)	15.642 (9.9963–21.2871)	-0.575 (-0.8055–-0.3446)	0.435 (0.319–0.5512)	43	51
Sample age <5000	0.0328 (0.0327–0.0329)	94.901 (94.2585–95.5435)	85.612 (84.6975–86.526)	-1.211(-1.2197–-1.2019)	-2.5 (-2.5136–-2.4855)		
Sample age >5000	0.0985 (0.0985–0.0985)	3.103 (3.1027–3.1031)	44.876 (44.8747–44.8768)	0.354 (0.3537–0.3537)	-0.663 (-0.6634–-0.6633)	33	51
Sample age <5000	0.0413 (0.0411–0.0415)	96.029 (95.8493–96.2087)	85.711 (83.6634–87.7594)	-2.5 (-2.5002–-2.4998)	-1.318 (-1.46–-1.1764)		
Sample age >5000	0.0979 (0.0978–0.0979)	70.388 (70.3697–70.4065)	2.628 (2.6271–2.6286)	-2.328 (-2.3286–-2.3276)	1.216 (1.2159–1.2164)	53	51
Sample age <5000	0.0376 (0.0374–0.0377)	3.705 (1.9497–5.4607)	77.019 (74.9065–79.1311)	-2.413 (-2.4174–-2.4084)	-2.5 (-2.4999–-2.4995)		
Sample age >5000	0.0991 (0.0991–0.0992)	1.218 (1.218–1.2183)	15.127 (15.1256–15.1287)	-0.781 (-0.781–-0.7807)	2.452 (2.452–2.4526)	43	61
Sample age <5000	0.0359 (0.0357–0.0361)	96.836 (96.6538–97.0183)	86.616 (83.9434–89.2891)	-2.499 (-2.4994–-2.499)	-2.219 (-2.3368–-2.1009)		
Sample age >5000	0.0999 (0.0999–0.0999)	27.442 (27.4385–27.4464)	11.879 (11.8781–11.8801)	-1.582 (-1.5824–-1.582)	-1.638 (-1.6382–-1.638)	43	41
Sample age <5000	0.0355 (0.0353–0.0357)	97.044 (96.8637–97.2236)	86.223 (83.4533–88.992)	-2.499 (-2.4996–-2.4992)	-2.148 (-2.2811–-2.0141)		

Table S7: Summary of parameter estimates for rs4988235(T) when the origin of the allele is forced to be at different points in the map (top panel corresponds to the original fit for the geographic position). In all cases, the estimated age of allele that was inputted into the model is 7,441 years ago. Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

	s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat	Allele age (years)
Sample age >5000	0.0989 (0.0989–0.0989)	9.341 (9.3402–9.341)	3.264 (3.2635–3.2643)	2.338 (2.3379–2.3381)	-0.21 (-0.2098–-0.2098)	13	48	7441
Sample age <5000	0.0358 (0.0357–0.036)	97.086 (96.9059–97.2657)	87.043 (85.1968–88.8895)	-2.434 (-2.4385–-2.4294)	-2.499 (-2.4994–-2.499)			

Table S8: Parameter estimates for rs4988235(T) using the geographic origin of the allele inferred in Itan *et al.* (2009). Columns named ”Long” and ”Lat” indicate the longitude and latitude of the geographic origin of the allele, respectively.

	s (95% CI)	σ_x (km/gen) (95% CI)	σ_y (km/gen) (95% CI)	v_x (km/gen) (95% CI)	v_y (km/gen) (95% CI)	Long	Lat
Sample age >5000	0.0221 (0.0216–0.0227)	71.668 (24.7274–118.6092)	50.434 (25.6535–75.2136)	-2.268 (-3.006–-1.5304)	-0.486 (-0.8661–-0.1053)	44	43
Sample age <5000	0.0102 (0.0083–0.012)	69.25 (14.0247–124.4756)	95.281 (95.1087–95.453)	0.849 (-0.0783–1.7769)	-0.503 (-0.929–-0.076)		

Sample age >5000	0.0227 (0.0223–0.0231)	42.745 (33.6354–51.8541)	96.993 (96.8183–97.1683)	-2.437 (-2.4412–-2.4324)	-0.266 (-0.4848–-0.0468)	54	43
Sample age <5000	0.0095 (0.007–0.0119)	93.477 (7.6582–179.2965)	99.634 (0–205.4586)	-2.499 (-3.2101–-1.7873)	2.057 (-0.7888–4.903)		

Sample age >5000	0.0221 (0.0221–0.0221)	47.691 (47.6686–47.7127)	71.367 (71.336–71.3986)	-2.164 (-2.1652–-2.1637)	1.839 (1.8387–1.8392)	44	53
Sample age <5000	0.0112 (0.0093–0.0131)	87.959 (0–215.8939)	88.951 (25.5422–152.3589)	2.108 (-0.2061–4.4227)	-2.237 (-5.7828–1.3083)		

Sample age >5000	0.0219 (0.0209–0.0229)	73.106 (38.1699–108.043)	76.835 (24.0025–129.6684)	-2.429 (-2.4335–-2.4248)	-1.474 (-2.8769–-0.0706)	44	33
Sample age <5000	0.0102 (0.0083–0.0121)	88.216 (0–192.1057)	95.401 (95.2283–95.573)	0.871 (-0.2474–1.9893)	-1.026 (-2.6161–0.564)		

Table S9: Summary of parameter estimates for rs1042602(A) when the origin of the allele is forced to be at different points in the map (top panel corresponds to the original fit for the geographic position). In all cases, the estimated age of allele that was inputted into the model is 26,367 years ago. Columns named ”Long” and ”Lat” indicate the longitude and latitude of the geographic origin of the allele, respectively.