



**HAL**  
open science

# Modelling the spatiotemporal spread of beneficial alleles using ancient genomes

Rasa Muktupavela, Martin Petr, Laure L. Segurel, Thorfinn Korneliussen,  
John Novembre, Fernando Racimo

► **To cite this version:**

Rasa Muktupavela, Martin Petr, Laure L. Segurel, Thorfinn Korneliussen, John Novembre, et al..  
Modelling the spatiotemporal spread of beneficial alleles using ancient genomes. 2022. hal-03796564

**HAL Id: hal-03796564**

**<https://hal.science/hal-03796564>**

Preprint submitted on 4 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 Modelling the spatiotemporal spread 2 of beneficial alleles using ancient 3 genomes

4 **Rasa Muktupavela<sup>1\*</sup>, Martin Petr<sup>1</sup>, Laure Ségurel<sup>2</sup>, Thorfinn Korneliussen<sup>1</sup>, John  
5 Novembre<sup>3</sup>, Fernando Racimo<sup>1</sup>**

\*For correspondence:

[rasa.muktupavela@gmail.com](mailto:rasa.muktupavela@gmail.com) (

)

6 <sup>1</sup>Lundbeck GeoGenetics Centre, GLOBE Institute, Faculty of Health; <sup>2</sup>Laboratoire de  
7 Biométrie et Biologie Evolutive UMR5558, CNRS - Université Lyon 1, Université de Lyon,  
8 Villeurbanne, France; <sup>3</sup>University of Chicago, Department of Human Genetics, Chicago,  
9 IL, USA

---

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

---

## 28 Introduction

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

34 Patterns of genetic variation can be used to estimate how strongly natural selection has af-  
35 fected the trajectory of an allele and to fit the parameters of the selection process. The problem of  
36 estimating the age of a beneficial allele, for example, has yielded a rich methodological literature  
37 (*Slatkin and Rannala, 2000*), and recent methods have exploited fine-scale haplotype information  
38 to produce highly accurate age estimates (*Mathieson and McVean, 2014; Platt et al., 2019; Albers  
39 and McVean, 2020*). In contrast, efforts to infer the geographic origins of beneficial mutations are  
40 scarcer. These include *Novembre et al. (2005)*, who developed a maximum likelihood method to



41 model the origin and spread of a beneficial mutation and applied it to the *CCR5-Δ32* allele, which  
42 was, at the time, considered to have been under positive selection (*Stephens et al., 1998; Sabeti*  
43 *et al., 2005; Novembre and Han, 2012*). Similarly, *Itan et al. (2009)* developed an approximate  
44 Bayesian computation (ABC) approach using demic simulations, in order to find the geographic  
45 and temporal origins of a beneficial allele, based on present-day allele frequency patterns.

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

56 Dynamic models can also contribute to ongoing debates about the past trajectories of pheno-  
57 typically important loci. For example, the geographic origin of the rs4988235(T) allele—upstream  
58 of the *LCT* gene and associated with adult lactase persistence in most of Western Eurasia (*Enattah*  
59 *et al., 2002*)—remains elusive, as is the way in which it spread (an extensive review can be found in  
60 *Ségurel and Bon, 2017*). The allele has been found in different populations, with frequencies rang-  
61 ing from 5% up to almost 100%, and its selection coefficient has been estimated to be among the  
62 highest in human populations (*Bersaglieri et al., 2004; Enattah et al., 2008; Tishkoff et al., 2007*).  
63 However, the exact causes for its adaptive advantage are contested (*Szpak et al., 2019*), and it has  
64 been suggested that the selection pressures acting on the allele may have been different in differ-  
65 ent parts of the continent (*Gerbault et al., 2009*). Ancient DNA evidence shows that the allele was  
66 rare in Europe during the Neolithic (*Burger et al., 2007; Gamba et al., 2014; Allentoft et al., 2015;*  
67 *Mathieson et al., 2015*) and only became common in Northern Europe after the Iron Age, suggest-  
68 ing a rise in frequency during this period, perhaps mediated by gene flow from regions east of the  
69 Baltic where this allele was more common during the onset of the Bronze Age (*Krüttli et al., 2014;*  
70 *Margaryan et al., 2020*). *Itan et al. (2009)* deployed their ABC approach to model the spatial spread  
71 of the rs4988235(T) allele and estimated that it was first under selection among farmers around  
72 7,500 years ago possibly between the central Balkans and central Europe. Others have postulated  
73 a steppe origin for the allele (*Allentoft et al., 2015*), given that the rise in frequency appears to have  
74 occurred during and after the Bronze Age migration of steppe peoples into Western Eurasia (*Haak*  
75 *et al., 2015; Allentoft et al., 2015*). However, the allele is at low frequency in genomes of Bronze  
76 Age individuals associated with Corded Ware and Bell Beaker assemblages in Central Europe who  
77 have high steppe ancestry (*Mathieson et al., 2015; Margaryan et al., 2020*), complicating the story  
78 further (*Ségurel and Bon, 2017*).

79 The origins and spread dynamics of large-effect pigmentation-associated SNPs in ancient Eurasians  
80 have also been intensely studied (*Ju and Mathieson, 2020*). Major loci of large effect on skin, eye  
81 and hair pigmentation have been documented as having been under recent positive selection in  
82 Western Eurasian history (*Voight et al., 2006; Sabeti et al., 2007; Pickrell et al., 2009; Lao et al.,*  
83 *2007; Mathieson et al., 2015; Alonso et al., 2008; Hudjashov et al., 2013*). These include genes  
84 *SLC45A2, OCA2, HERC2, SLC24A5* and *TYR*. While there is extensive evidence supporting the adap-  
85 tive significance of these alleles, debates around their exact origins and spread are largely driven  
86 by comparisons of allele frequency estimates in population groups which are almost always dis-  
87 cretized in time and/or space. Among these, selection at the *TYR* locus is thought to have occurred  
88 particularly recently, over the last 5,000 years (*Stern et al., 2019*), driven by a recent mutation (*Al-*  
89 *bers and McVean, 2020*) that may have spread rapidly in Western Eurasia.

90 Here, we develop a method to model the spread of a recently selected allele across both space  
91 and time, avoiding artificial discretization schemes to more rigorously assess the evidence for or

92 against a particular dispersal process. We begin with the model proposed by *Novembre et al.*  
93 (2005), and adapt it in order to handle ancient low-coverage genomic data, and explore more com-  
94 plex models that allow for both diffusion and advection (i.e. directional transport) in the distribu-  
95 tion of allele frequencies over space, as well as for a change in these parameters at different periods  
96 of time. We apply the method to alleles in two of the aforementioned loci in the human genome,  
97 which have been reported to have strong evidence for recent positive selection: *LCT/MCM6* and  
98 *TYR*. We focus on Western Eurasia during the Holocene, where ancient genomes are most densely  
99 sampled, and infer parameters relevant to the spread of these alleles, including selection, diffusion  
100 and advection coefficients.

## 101 Results

### 102 Summary of model

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

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

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

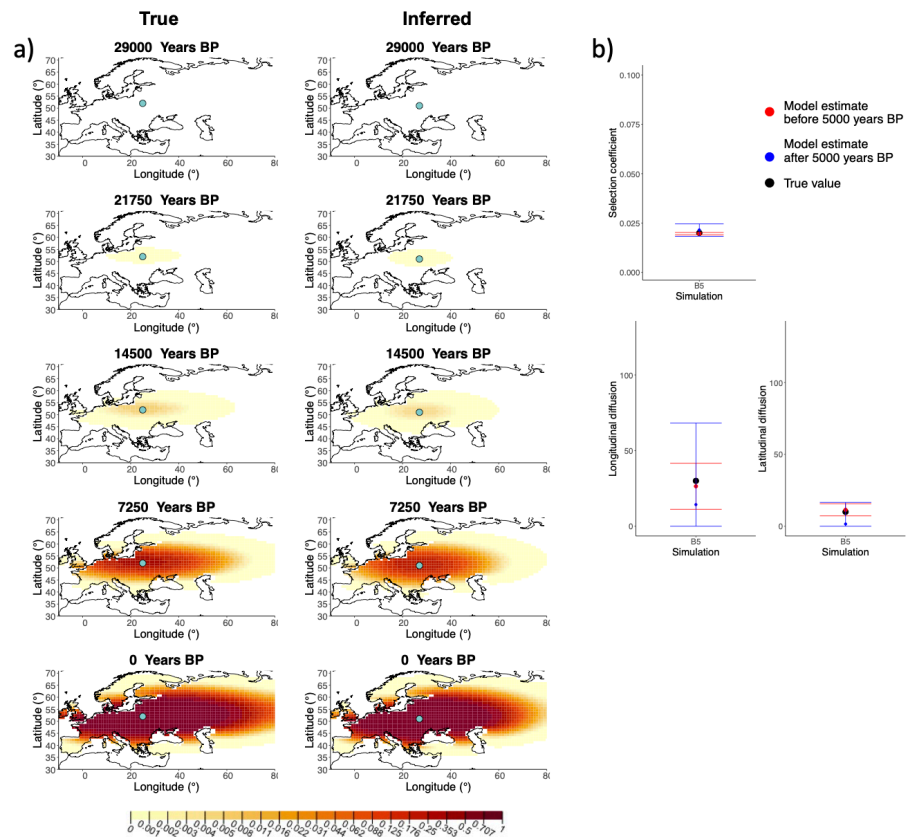
124 In order to establish a starting time point for our diffusion process, we used previously pub-  
125 lished allele age estimates obtained from a non-parametric approach leveraging the patterns of  
126 haplotype concordance and discordance around the mutation of interest (*Albers and McVean,*  
127 *2020*). In the case of the allele in the *LCT/MCM6* region, we also used age estimates based on an  
128 approximate Bayesian computation approach (*Itan et al., 2009*).

### 129 Performance on deterministic simulations

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

135 We generated six different simulations with different diffusion coefficients and afterwards ran  
136 our method assuming model B. The results (simulations B1-B6) are summarised in *Figure 1, Fig-*  
137 *ure 1–Figure Supplement 1, Figure 1–Figure Supplement 2, Figure 1–Figure Supplement 3, Figure 1–*  
138 *Figure Supplement 4, Figure 1–Figure Supplement 5* and *Table A1*. Overall, the model is more ac-

139 curate at correctly inferring the parameters for the time period before 5,000 years BP (**Figure 1b**),  
 140 with decreased performance when longitudinal diffusion is high (**Figure 1–Figure Supplement 5**).



**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 A1**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

**Figure 1–Figure supplement 1.** Comparison of true and inferred allele frequency dynamics for simulation B1.

**Figure 1–Figure supplement 2.** Comparison of true and inferred allele frequency dynamics for simulation B2.

**Figure 1–Figure supplement 3.** Comparison of true and inferred allele frequency dynamics for simulation B3.

**Figure 1–Figure supplement 4.** Comparison of true and inferred allele frequency dynamics for simulation B4.

**Figure 1–Figure supplement 5.** Comparison of true and inferred allele frequency dynamics for simulation B6.

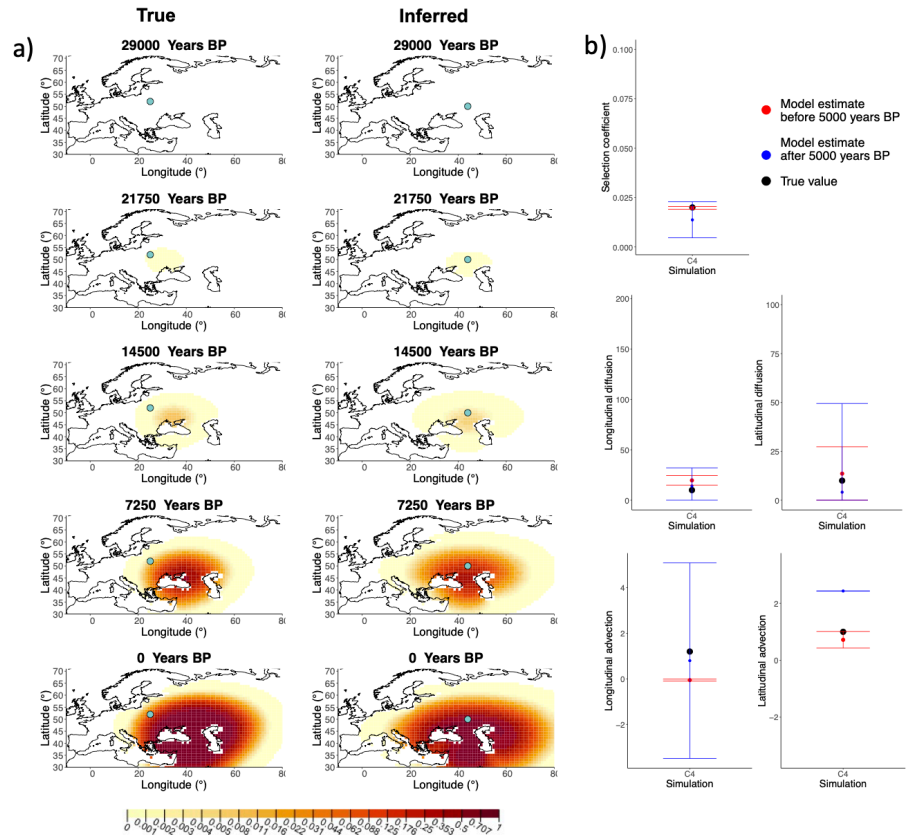
**Figure 1–Figure supplement 6.** Comparison of true allele frequency dynamics for simulation B1 and those inferred by the model C.

**Figure 1–Figure supplement 7.** Comparison of true allele frequency dynamics for simulation B4 and those inferred by the model C.

141 Next, we investigated the performance of model C, which includes advection coefficients. We  
 142 generated four different simulations including advection (simulations C1–C4: **Figure 2**, **Figure 2–**  
 143 **Figure Supplement 1**, **Figure 2–Figure Supplement 2**, **Figure 2–Figure Supplement 3** and **Table A2**).  
 144 We found that our method is generally able to estimate the selection coefficient accurately. How-  
 145 ever, in some of the simulations, we found discrepancies between the estimated and true diffusion  
 146 and advection coefficients, often occurring because of a misestimated origin forcing the other pa-  
 147 rameters to adjust in order to better fit the allele frequency distribution in later stages of the allele’s  
 148 spread (**Figure 2**). Despite the disparities between the true and inferred parameter values, the re-  
 149 sulting surface plots become very similar as we approach the present, suggesting that different



150 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 A2](#). b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

**Figure 2–Figure supplement 1.** Comparison of true and inferred allele frequency dynamics for simulation C1.

**Figure 2–Figure supplement 2.** Comparison of true and inferred allele frequency dynamics for simulation C2.

**Figure 2–Figure supplement 3.** Comparison of true and inferred allele frequency dynamics for simulation C3.

### 151 Advection model application to non-advection simulations

152 We assessed the model performance when we apply the model C, which includes advection coef-  
 153 ficient estimates, to simulations generated without advection (see [Figure 1–Figure Supplement 6](#)  
 154 and [Figure 1–Figure Supplement 7](#)). We can observe that the advection coefficients are inferred to  
 155 be non-zero ([Figure 1–Figure Supplement 6b](#) and [Figure 1–Figure Supplement 7b](#)), however the in-  
 156 ferred allele frequency dynamic plots closely resemble the ones obtained with with true parameter  
 157 values ([Figure 1–Figure Supplement 6a](#) and [Figure 1–Figure Supplement 7a](#)). This shows that complex  
 158 interactions between the diffusion and advection coefficients can result in similar outcomes  
 159 even when only diffusion is considered in the model.

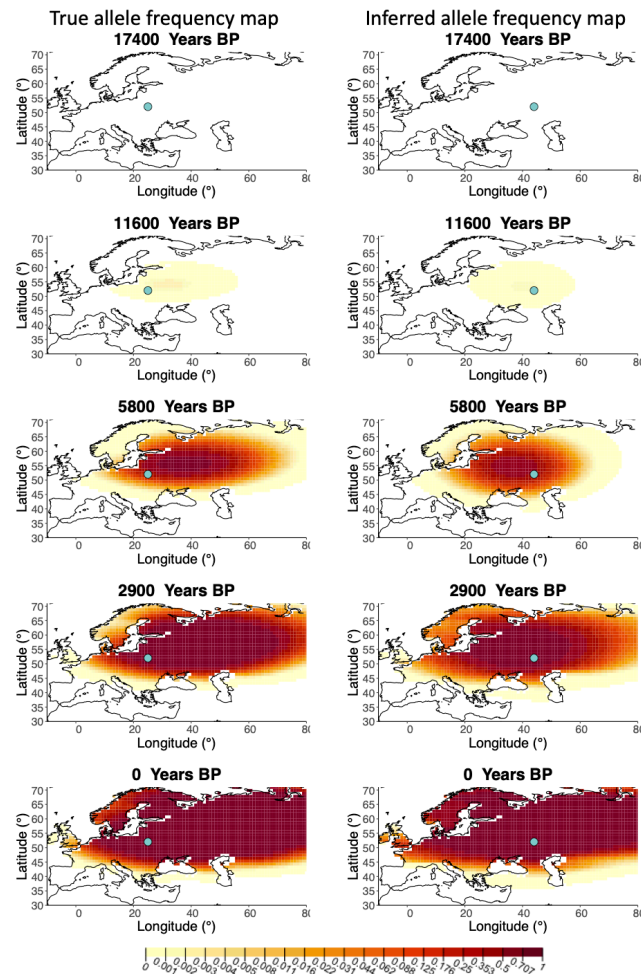
160 The inference of the origin of the allele also differs when we compare the results for using  
 161 model B and model C. In order to understand better how the model estimates the allele origin, we  
 162 highlighted the first individual in simulations B1 and B4 that contains the derived allele. We can  
 163 see that in case of simulation B1 the inferred origin of the allele is close to the first observance  
 164 of the derived allele in the model which includes advection. In contrast when the advection is not  
 165 included, the origin of the allele is inferred to be closer to where it is initially rising in frequency  
 166 ([Figure 1–Figure Supplement 1a](#) and [Figure 1–Figure Supplement 4a](#)). However, this is not always

167 the case. For instance, if we look at the results from the advection model on simulation B4, we can  
168 see that the origin of the allele is inferred relatively far from the sample known to have carried the  
169 first instance of the derived allele. Therefore, if there is a relatively large interval between the time  
170 when the allele originated and when the first ancient genomes are available, the beneficial allele  
171 can spread widely, but as this spread is not captured by any of the data points, inference of the  
172 precise origin of the selected allele is nearly impossible.

### 173 **Impact of sample clustering on parameter estimates**

174 We evaluated the impact of different sampling and clustering schemes on our inferences that could  
175 potentially arise by aggregating aDNA data from studies with different sampling schemes. We used  
176 a deterministic simulation to create three different degrees of clustering which we will refer to as  
177 “homogeneous”, “intermediate” or “extreme” by varying the area from which we sample individuals  
178 to be used in our inferences (**Figure 3–Figure Supplement 1**). Additionally, we also tested the impact  
179 of biased temporal sampling in the periods before and after 5000 year BP by oversampling in the  
180 ancient period (75%/25%), equal sampling in the two periods (50%/50%), and oversampling in the  
181 recent period (25%/75%). Because we evaluated this temporal bias for each of the three spatial  
182 clustering sampling scenarios, this resulted in a total of 9 different sampling scenarios. We note  
183 that the third “extreme” spatial clustering scenarios is completely unrealistic and one would not  
184 expect inferences of any degree of accuracy from it, but we believe it gives a good idea of the  
185 behaviour of our method in the limit case of extremely restricted spatial sampling.

186 A comparison of allele frequency maps generated using true parameter values and using pa-  
187 rameter estimates from the different sampling schemes are shown in **Figure 3–Figure Supple-**  
188 **ment 2, Figure 3–Figure Supplement 3, Figure 3–Figure Supplement 4, Figure 3–Figure Supple-**  
189 **ment 5, Figure 3–Figure Supplement 6, Figure 3–Figure Supplement 7, Figure 3–Figure Supple-**  
190 **ment 8, Figure 3–Figure Supplement 9**. In **Figure 3** we show the allele frequency map generated  
191 using the “intermediate 75%/25%” clustering scheme. Parameter estimates used to generate all  
192 these figures are summarised in **Table A3**. Overall we can see that the allele frequency maps in-  
193 ferred from these scenarios closely resemble the maps generated using the true parameter values,  
194 despite the challenges in finding accurate values for the individual point estimates of some of the  
195 parameters, highlighting that various combinations of diffusion and advection coefficients can pro-  
196 duce similar underlying frequency maps (as discussed in the manuscript section “Performance on  
197 deterministic simulations”). This suggests that the joint spatiotemporal information encoded in the  
198 inferred maps (not just the individual parameters estimates) should be used in interpreting model  
199 outputs, particularly when it comes to the advection and diffusion parameters. The selection coef-  
200 ficient estimates are inferred highly accurately, regardless of the sampling scheme chosen, and lie  
201 close to the true value, with only a slight underestimation in the time period after 5000 years BP  
202 (with the exception of “extreme 25%/75%”).



**Figure 3.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “intermediate 75%/25%” clustering scheme. Parameter values used to generate the maps are summarised in [Table A3](#).

**Figure 3–Figure supplement 1.** Examples of spatial sampling scenarios for each of the three clustering schemes.

**Figure 3–Figure supplement 2.** Allele frequency map generated using true parameter values and using parameter estimates for “homogeneous 75%/25%” clustering scheme

**Figure 3–Figure supplement 3.** Allele frequency map generated using true parameter values and using parameter estimates for “homogeneous 50%/50%” clustering scheme

**Figure 3–Figure supplement 4.** Allele frequency map generated using true parameter values and using parameter estimates for “homogeneous 25%/75%” clustering scheme

**Figure 3–Figure supplement 5.** Allele frequency map generated using true parameter values and using parameter estimates for “intermediate 50%/50%” clustering scheme

**Figure 3–Figure supplement 6.** Allele frequency map generated using true parameter values and using parameter estimates for “intermediate 25%/75%” clustering scheme

**Figure 3–Figure supplement 7.** Allele frequency map generated using true parameter values and using parameter estimates for “extreme 75%/25%” clustering scheme

**Figure 3–Figure supplement 8.** Allele frequency map generated using true parameter values and using parameter estimates for “extreme 50%/50%” clustering scheme

**Figure 3–Figure supplement 9.** Allele frequency map generated using true parameter values and using parameter estimates for “extreme 25%/75%” clustering scheme

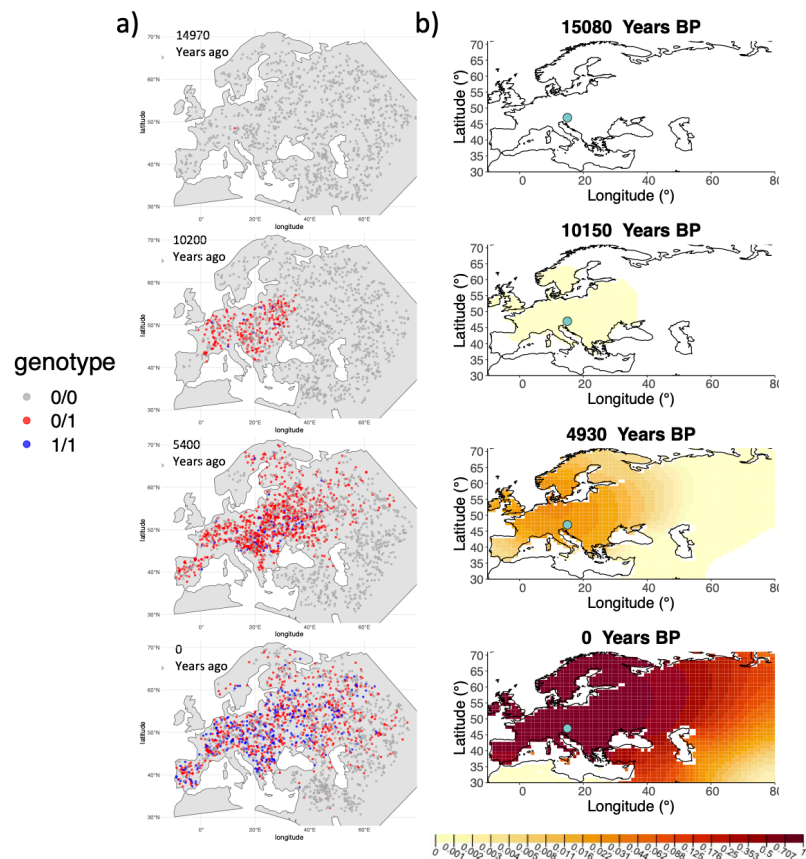


### 203 **Spatially-explicit forward simulations**

204 In addition to drawing simulated samples from a diffusion model, we used SLiM (*Haller and Messer*  
205 *(2019)*) to perform spatially explicit individual-based forward-in-time simulations of selection acting  
206 on a beneficial allele, by leveraging an R interface for spatial population genetics now implemented  
207 in an R package *slendr* (*Petr (2021)*).

208 We introduced a single beneficial additive mutation in a single individual and let it evolve across  
209 the European landscape. Before applying our method on the simulated data, we sampled 1,040  
210 individuals whose ages were log-uniformly distributed, to ensure that there were more samples  
211 closer to the present, as in the real data. We transformed the diploid genotypes to pseudohaploid  
212 genotypes by assigning a heterozygous individual an equal probability of carrying the ancestral or  
213 the derived genotype. The parameter values estimated by our model to the simulations described  
214 in this section are summarised in *Table A4*.

215 We can see that the origin of the allele inferred by the model closely corresponds to the first  
216 observation of the derived allele in the simulation (*Figure 4*). The inferred selection coefficient is  
217 only slightly higher than the true value from the simulation (0.0366 vs 0.030). In general, the model  
218 accurately captures the spread of the allele centered in central Europe, though we observe some  
219 discrepancies due to differences between the model assumed in the simulation (which, for exam-  
220 ple, accounts for local clustering of individuals, *Figure 4–Figure Supplement 1*), and that assumed  
221 by our diffusion-based inference.



**Figure 4.** 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 ages of sampled individuals were log-uniformly distributed. The estimated parameter values of the fitted model are shown in [Table A4](#).

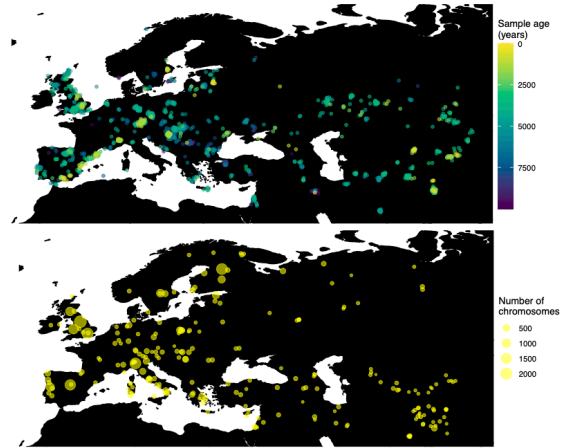
**Figure 4–Figure supplement 1.** Distribution of individuals across the map under neutrality, showing the tendency of individuals to cluster together.

## 222 Dynamics of the rs4988235(T) allele

223 Having tested the performance of our method on simulated data, we set out to infer the allele  
 224 frequency dynamics of the rs4988235(T) allele (associated with adult lactase persistence) in ancient  
 225 Western Eurasia. For our analysis, we used a genotype dataset compiled by [Segurel et al. \(2020\)](#),  
 226 which amounts to 1,434 genotypes from ancient Eurasian genomes individuals, and a set of 36,659  
 227 genotypes from present-day Western and Central Eurasian genomes ([Ségurel and Bon, 2017](#); [Heyer  
 228 et al., 2011](#); [Marchi et al., 2018](#); [Liebert et al., 2017](#); [Gallego Romero et al., 2012](#); [Itan et al., 2010](#);  
 229 [Charati et al., 2019](#)). After filtering out individuals falling outside of the range of the geographic  
 230 boundaries considered in this study, we retained 1,332 ancient individuals. The locations of ancient  
 231 and present-day individuals used in the analysis to trace the spread of rs4988235(T) are shown in  
 232 [Figure 5](#).

233 We used a two-period scheme by allowing the model to have two sets of estimates for the  
 234 selection coefficient and the diffusion and advection coefficients in two different periods of time:  
 235 before and after 5,000 years ago, reflecting the change in population dynamics and mobility before  
 236 and after the Bronze Age transition ([Loog et al., 2017](#); [Racimo et al., 2020a](#)). We used two allele

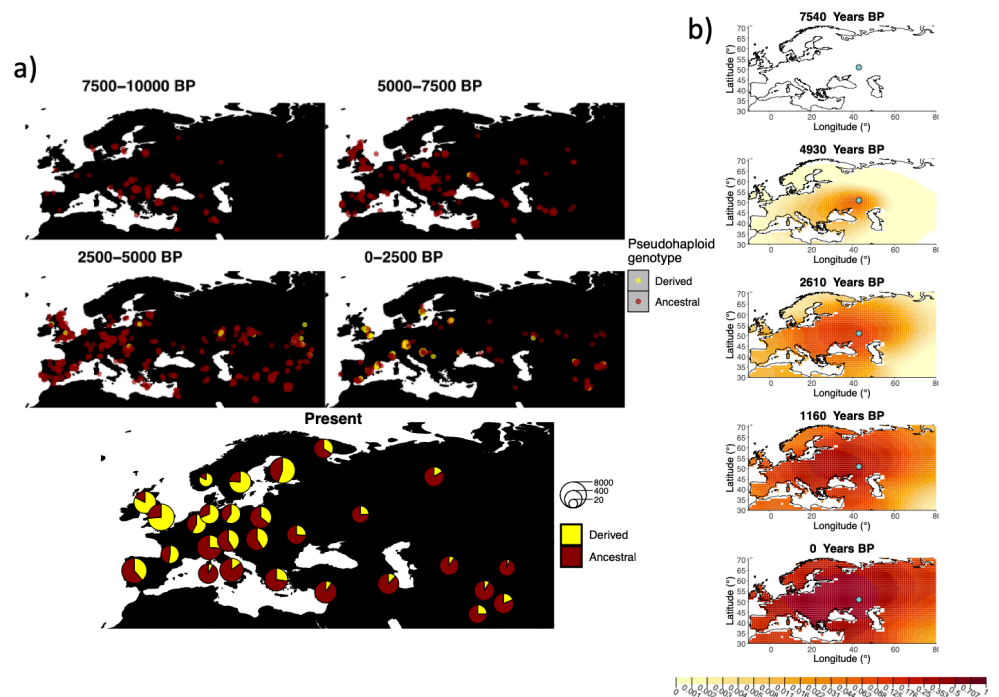
237 age estimates as input: a relatively young one (7,441 years ago) obtained from *Itan et al. (2009)*,  
238 and a relatively old one (20,106 years ago) obtained from *Albers and McVean (2020)*. The results  
239 obtained for fitting the model on rs4988235(T) are summarised in *Table A5* and *Table A6*, and in  
240 *Figure 6b* (younger age) and *Figure 6–Figure Supplement 1* (older age).



**Figure 5.** 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.

241 Assuming the age estimate from (*Itan et al., 2009*), the origin of the allele is estimated to be  
242 north of the Caucasus, around what is now southwestern Russia and eastern Ukraine (*Figure 6b*).  
243 Given that this age is relatively young, our method fits a very strong selection coefficient ( $\approx 0.1$ )  
244 during the first period in order to accommodate the early presence of the allele in various points  
245 throughout Eastern Europe, and a weaker (but still strong) selection coefficient ( $\approx 0.03$ ) in the sec-  
246 ond period. We also estimate stronger diffusion in the second period than in the first, to accommo-  
247 date the rapid expansion of the allele throughout Western Europe, and a net westward advection  
248 parameter, indicating movement of the allele frequency's center of mass to the west as we ap-  
249 proach the present.

250 Assuming the older age estimate from *Albers and McVean (2020)*, the origin of the allele is  
251 estimated to be in the Northeast of Europe (*Figure 6–Figure Supplement 1*), which is at a much  
252 higher latitude than the first occurrence of the allele, in Ukraine. Due to the deterministic nature  
253 of the model, the frequency is implicitly imposed to expand in a region where there are no actual  
254 observed instances of the allele. The model compensates for this by placing the origin in an area  
255 with a lower density of available aDNA data and thus avoiding an overlap of the increasing allele  
256 frequencies with individuals who do not carry the derived rs4988235(T) allele (see *Figure 6a*). As  
257 the model expands rapidly in the southern direction (*Table A6*) it eventually reaches the sample  
258 carrying the derived variant in Ukraine.



**Figure 6.** 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.

**Figure 6–Figure supplement 1.** Inferred frequency dynamics of rs4988235(T) using the allele age that was inferred in *Albers and McVean (2020)*.

**Figure 6–Figure supplement 2.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees west from the original estimate.

**Figure 6–Figure supplement 3.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees east from the original estimate.

**Figure 6–Figure supplement 4.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees north from the original estimate.

**Figure 6–Figure supplement 5.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees south from the original estimate.

**Figure 6–Figure supplement 6.** 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 6–Figure supplement 7.** 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)*.

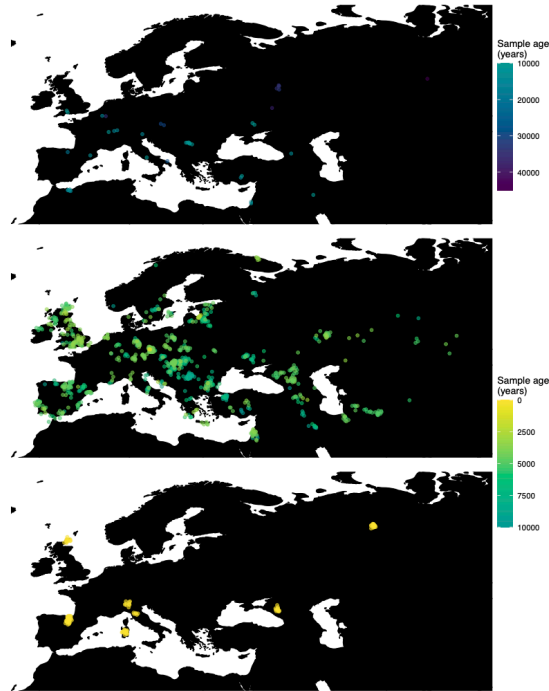
**Figure 6–Figure supplement 8.** Inferred frequency dynamics of rs4988235(T) assuming the allele age to be the higher end of the 95% credible interval for the allele age inferred in *Itan et al. (2009)*.

**Figure 6–Figure supplement 9.** Log-likelihood values for model runs using different ages of the rs4988235(T) allele as input.

## 259 Dynamics of the rs1042602(A) allele

260 Next, we investigated the spatiotemporal dynamics of the spread of an allele at a pigmentation-  
 261 associated SNP in the *TYR* locus (rs1042602(A)), which has been reported to be under recent selec-  
 262 tion in Western Eurasian history (*Stern et al., 2019*). For this purpose, we applied our method to  
 263 the Allen Ancient DNA Resource data (*Reich and Mallick, 2019*), which contains randomly sampled  
 264 pseudohaploid genotypes from 1,513 published ancient Eurasian genomes (listed in Supplemen-

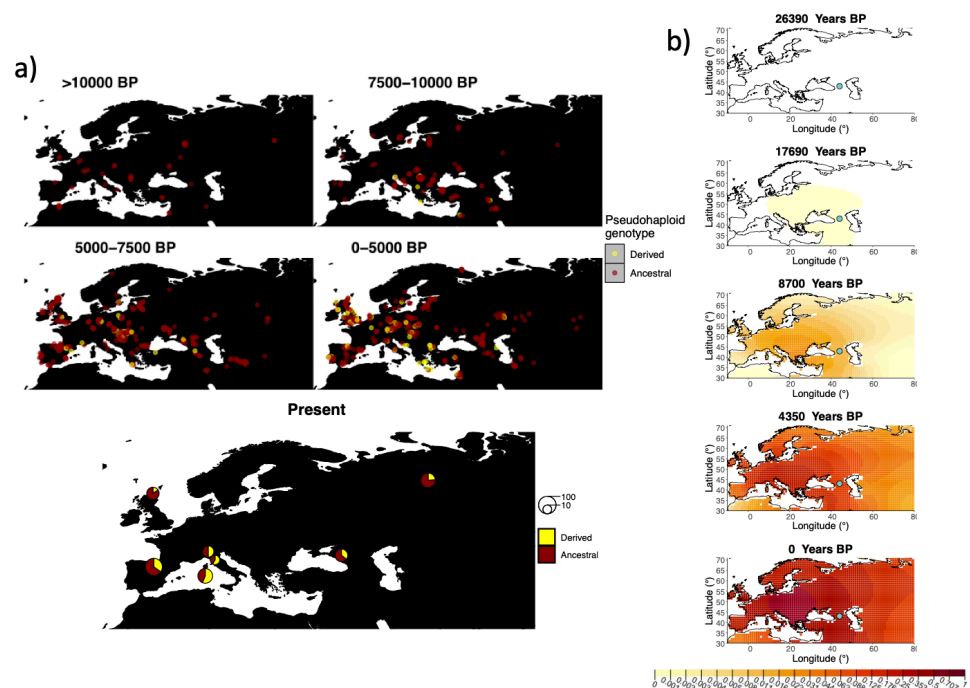
265 tary File 1), from which we extracted those genomes that had genotype information at this locus in  
266 Western Eurasia. We merged this dataset with diploid genotype information from high-coverage  
267 present-day West Eurasian genomes from the Human Genome Diversity Panel (HGDP) (*Bergström*  
268 *et al., 2020*), which resulted in a total of 1,040 individuals with genotype information at rs1042602,  
269 which were as input to our analysis. Geographic locations of individuals in the final dataset are  
270 shown in *Figure 7*.



**Figure 7.** 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.

271 Similarly to our analysis of the spread of the allele in rs4988235(T), we inferred the dynamics of  
272 the rs1042602(A) allele separately for the time periods before and after 5,000 years BP and assum-  
273 ing the age of the allele to be 26,361 years (*Albers and McVean, 2020*). The inferred parameters  
274 for both time periods are summarised in *Table A7* and the allele frequency surface maps gener-  
275 ated using these parameters are shown in *Figure 8b*. The origin of the rs1042602(A) corresponds  
276 closely to the region where the allele initially starts to segregate in the time period between 7,500  
277 and 10,000 years BP as seen in *Figure 8a*. Estimates of the selection coefficient for both time pe-  
278 riods (0.0221 and 0.0102 for the period before and after 5000 years BP, respectively) suggest that  
279 selection acting on the allele has decreased after 5000 years BP.





**Figure 8.** 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.

**Figure 8–Figure supplement 1.** Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees east from the original estimate.

**Figure 8–Figure supplement 2.** Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees north from the original estimate.

**Figure 8–Figure supplement 3.** Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees south from the original estimate.

**Figure 8–Figure supplement 4.** 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 and McVean (2020)*.

**Figure 8–Figure supplement 5.** 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 and McVean (2020)*.

**Figure 8–Figure supplement 6.** Log-likelihood values for model runs using different ages of the rs1042602(A) allele as input.

## 280 **Robustness of parameters to the inferred geographic origin of allele**

281 We carried out an analysis to characterize how sensitive the selection, diffusion and advection  
 282 parameters are to changes in the assumed geographic origin of the allele. For the rs4988235(T)  
 283 allele, we forced the origin of the allele to be 10 degrees away from our inferred origin in each car-  
 284 dinal direction, while assuming the allele age from *Itan et al. (2009) (Table A8)*. In *Figure 6–Figure*  
 285 *Supplement 2, Figure 6–Figure Supplement 3, Figure 6–Figure Supplement 4* and *Figure 6–Figure*  
 286 *Supplement 5*, we can see the allele frequency dynamics of these four scenarios, respectively. We  
 287 also forced the allele origin to be at the geographic origin estimated in *Itan et al. (2009) (Table A9,*  
 288 *Figure 6–Figure Supplement 6)*, which is westward of our estimate. In all five cases during the pe-  
 289 riod prior 5,000 years BP, the allele is inferred to expand in the direction of the first sample that  
 290 is observed to carry the rs4988235(T) allele and is located in Ukraine. During the time period after  
 291 5000 years BP, the patterns produced by the model are rather similar, although the parameters  
 292 associated with diffusion and advection differ, in order to account for the different starting condi-  
 293 tions.



294 We also investigated how the results are affected when the estimated geographic origin of the  
295 rs1042602(A) allele is moved with respect to the initial estimate. We set the allele to be 10 degrees  
296 east, 10 degrees north and 10 degrees south of the original estimate as shown in **Figure 8–Figure**  
297 **Supplement 1, Figure 8–Figure Supplement 2 and Figure 8–Figure Supplement 3**, respectively. We  
298 did not look at a scenario in which the origin of the allele is moved to the west, since it would either  
299 end up in the Black sea or more westwards than 10 degrees. The selection coefficient remains  
300 similar to the original estimate throughout all three scenarios. The way the allele spreads across  
301 the landscape is also similar in all cases and, as in the case of rs4988235(T), the model accounts for  
302 the different origins of the allele by adjusting the diffusion and advection coefficients in the time  
303 period after 5000 years BP.

### 304 **Robustness of parameters to the assumed age of the allele**

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

312 When re-fitting the model for the rs4988235(T) allele, we found that the inferred selection co-  
313 efficient is slightly lower when the allele age is assumed to be at the lower bound of the 95% credi-  
314 ble interval (0.0867 vs 0.0993 before 5000 years BP and 0.0321 vs 0.0328 after 5000 years BP) and  
315 slightly higher when assumed to be at the upper bound (0.0994 vs 0.0993 before 5000 years BP and  
316 0.0572 vs 0.0328 after 5000 years BP) (**Table A5 and Figure 6–Figure Supplement 7 and Figure 6–**  
317 **Figure Supplement 8**). This occurs because the selection intensity must be higher or lower when  
318 there is more or less time, respectively, for the allele to reach the allele frequencies observed in the  
319 data. In the case of the rs1042602(A) allele, this only affects the earlier time period (**Table A7**). The  
320 rs4988235(T) allele’s geographic distribution in the more recent time periods is also less extended  
321 geographically when the age is assumed to be young. The inferred geographic origin of both alleles  
322 slightly differs under different assumed ages (**Figure 8–Figure Supplement 4 and Figure 8–Figure**  
323 **Supplement 5**).

324 In addition, we assessed the likelihood of the best fitted models with varying the ages of the  
325 rs4988235(T) and rs1042602(A) alleles (**Figure 6–Figure Supplement 9 and Figure 8–Figure Supple-**  
326 **ment 6**, respectively). We can see that in the case of rs4988235(T) allele the allele age used in  
327 this study (7,441 years) gives the most likely solution among the explored ages. In case of the  
328 rs1042602(A) allele, we found that there are multiple nearly equally likely ages when looking at  
329 ages at least as old as 15,000 years.

### 330 **Discussion**

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

341 The inferred geographic origin of the rs4988235(T) allele reflects the best guess of our frame-  
342 work given the constraints provided by its input, namely the previously inferred age of the allele

343 and the observed instances of this allele throughout Western Eurasia. We are also assuming that  
344 the allele must have arisen somewhere within the bounding box of our studied map. When assum-  
345 ing a relatively young allele age (7,441 years ago, *Itan et al. (2009)*), the origin of the allele is placed  
346 north of the Caucasus, perhaps among steppe populations that inhabited the area at this time  
347 (*Haak et al., 2015; Allentoft et al., 2015*). This origin is further east than the geographic origin es-  
348 timate from *Itan et al. (2009)*, likely reflecting additional ancient DNA information that is available  
349 to us, and indicates an early presence of the allele in eastern Europe. When assuming a relatively  
350 old allele age (20,106 years ago, *Albers and McVean (2020)*), the age is placed in northeast Europe,  
351 perhaps among Eastern hunter-gatherer groups that inhabited the region in the early Holocene.  
352 We note that the number of available genomes for eastern and northeastern Europe during the  
353 early Holocene is scarce, so the uncertainty of the exact location of this origin is relatively high. Re-  
354 gardless of the assumed age, we estimate a net westward displacement of the allele frequency's  
355 center of mass, and a rapid diffusion, particularly in the period after 5,000 years ago.

356 Various studies have estimated the selection coefficient for the rs4988235(T) allele, and these  
357 range from as low as 0.014 to as high as 0.19 (*Enattah et al., 2008; Mathieson and Mathieson,*  
358 *2018; Mathieson, 2020; Stern et al., 2019; Burger et al., 2020; Peter et al., 2012; Gerbault et al.,*  
359 *2009; Itan et al., 2009; Bersaglieri et al., 2004*). Recent papers incorporating ancient DNA estimate  
360 the selection coefficient to be as low as 0 (in certain regions of Southern Europe) and as high as  
361 0.06 (*Mathieson and Mathieson, 2018; Mathieson, 2020; Burger et al., 2020*). It is also likely that the  
362 selection coefficient was different for different regions of Europe, perhaps due to varying cultural  
363 practices (*Mathieson, 2020*). In our case, the estimated selection coefficient during the first period  
364 - before 5,000 years ago - depends strongly on the assumed allele age ( $s = 0.0993$  vs.  $s = 0.0285$ ). As  
365 in the case of the geographic origin, these estimates should be taken with caution as the number of  
366 available allele observations in the early Holocene is fairly low. The estimates for the second period  
367 - after 5,000 years ago - are more robust to the assumed age:  $s = 0.0328$  (95% CI: 0.0327–0.0329)  
368 if we assume the younger allele age (7,441 years ago) and  $s = 0.0255$  (95% CI: 0.0252–0.0258) if  
369 we assume the older allele age (20,106 years ago). These estimates are also within the range of  
370 previous estimates.

371 In the case of the rs1042602(A) allele, our estimated selection coefficients of 0.0221 (95% CI:  
372 0.0216–0.0227) and 0.0102 (95% CI: 0.0083–0.0120) for the time periods before and after 5000 years  
373 BP, respectively, are generally in agreement with previous results. *Wilde et al. (2014)* used a for-  
374 ward simulation approach to infer a point estimate of 0.026. Another study using an approximate  
375 Bayesian computation framework (*Nakagome et al., 2019*) estimated the strength of selection act-  
376 ing on rs1042602 to be 0.013 (0.002–0.029). Although both studies utilized ancient DNA data, the  
377 estimates were obtained without explicitly modelling the spatial dimension of the selection pro-  
378 cess.

379 Our estimates of the longitudinal advection parameter are negative for both the SNPs in the *TYR*  
380 and *LCT* loci: the mutation origins are always to the east of the center of mass of the allele frequency  
381 distribution seen in present-day data. This perhaps reflects common migratory processes, like the  
382 large-scale Neolithic and Bronze Age population movements from east to west, affecting the allele  
383 frequencies at these loci across the Eurasian landscape (*Allentoft et al., 2015; Haak et al., 2015*). As  
384 a form of regularization, we kept the range of explored values for the advection parameters to be  
385 small (-2.5 to 2.5 km per generation), while allowing the diffusion parameters to be explored over  
386 a much wider range of values. In certain cases, like the second period of the rs4988235(T) spread  
387 when the allele age is assumed to be young (*Table A5*), we find that the advection parameters are  
388 fitted at the boundary of the explored range, because the allele needs to spread very fast across  
389 the landscape to fit the data.

390 A future improvement to our method could include other forms of regularization that better ac-  
391 count for the joint behavior of the advection and diffusion processes, or the use of priors for these  
392 parameters under a Bayesian setting, which could be informed by realistic assumptions about the  
393 movement of individuals on a landscape. Bayesian parameter fitting would likely provide a more

394 robust understanding of the uncertainty of the estimates as well as an opportunity to formally  
395 compare different models using Bayes factors, although at the cost of an increase of computa-  
396 tional intensity.

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

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

413 Potential future extensions of our method could incorporate geographic features and historical  
414 migration events that create spatially or temporally varying moderators of gene flow. An example  
415 of this type of processes is the retreat of glaciers after the last Glacial maximum, which allowed  
416 migration of humans into Scandinavia (*Günther et al., 2018*). These changing geographic features  
417 could lead to changes in the rate of advection or diffusion across time or space. They could also  
418 serve to put more environmentally-aware constraints on the geographic origin of the allele, given  
419 that it cannot have existed in regions uninhabitable by humans, and to extend our analyses beyond  
420 the narrow confines of the Western Eurasian map chosen for this study. One could also envision  
421 incorporating variation in population densities over time, or known migration processes in the  
422 time frames and regions of interest. These might have facilitated rapid, long-range dispersal of  
423 beneficial alleles (*Bradburd et al., 2016; Hallatschek and Fisher, 2014*) or caused allelic surfing on  
424 the wave of range expansions (*Klopfstein et al., 2006*). Additional information like this could come,  
425 for example, from previously inferred spatiotemporal demographic processes (e.g. *Racimo et al.*  
426 *(2020b)*).

427 As described above, our model only accounts for diffusion in two directions. Further extension  
428 of our model could therefore incorporate anisotropic diffusion (*Othmer et al., 1988; Painter and*  
429 *Hillen, 2018*). Another possibility could be the introduction of stochastic process components, in or-  
430 der to convert the partial differential equations into stochastic differential equations (*Brown et al.,*  
431 *2000*). Stochastic components could serve to induce spatial autocorrelation and capture local pat-  
432 terns of allele frequency covariance in space that might not be well modeled by the deterministic  
433 PDEs (*Cressie and Wikle, 2015*). They could also serve to induce stochasticity in allele frequency  
434 changes over time as a consequence of genetic drift (*Crow et al., 1970*), allowing one to model  
435 the dynamics of more weakly selected variants, where drift plays an important role. Eventually,  
436 one could perhaps combine information across loci to jointly model the spatiotemporal frequency  
437 surfaces at multiple loci associated with the same trait. This could help clarify the dynamics of poly-  
438 genic adaptation and negative selection on complex traits (*Irving-Pease et al., 2021*), and perhaps  
439 hindcast the genetic value of traits across a landscape.

440 The availability of hundreds of ancient genomes (*Marciniak and Perry, 2017*) and the increas-  
441 ing interest in spatiotemporal method development (*Bradburd and Ralph, 2019*), such as the one  
442 described in this manuscript, will likely lead researchers to posit new questions and hypotheses  
443 about the behavior of natural selection. In the case of a beneficial allele spreading on a landscape,  
444 new ontologies and vocabulary for describing positive selection in time and space will be needed.



445 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) 446 (Hermisson and Pennings, 2005; Pritchard and Di Rienzo, 2010; Hermisson and Pennings, 2017). In 447 contrast, there is a lack of vocabulary for distinguishing between a scenario of strong selection that 448 is locally constrained in space from a scenario of widespread selection extended over a landscape, 449 or a model of neutral diffusion in space followed by parallel non-neutral increases in frequency 450 at multiple locations. For example, Ralph and Coop (2010) showed how multiple localized hard 451 sweeps may be seen as a soft sweep at a larger population-wide scale. Existing vocabulary for spa- 452 tiotemporal genetic processes is clearly not enough, limiting the types of questions or hypotheses 453 we can pose about them. 454

455 Population genetic models that explicitly account for space and time are an important area of 456 future methodological development (Bradburd and Ralph, 2019). We believe that methods such 457 as the one described in this study show great promise at broadening the horizon of our under- 458 standing of natural selection across space and time in humans and other species. As in the case 459 of demographic reconstruction (Ray and Excoffier, 2009), spatiotemporal information can greatly 460 help improve our knowledge of how natural selection operated in the past.

## 461 Methods

### 462 The model

463 To describe the allele frequency dynamics in time and space, we first begin by using a deterministic 464 model based on a two-dimensional partial differential equation (PDE) (Fisher, 1937; Kolmogorov 465 et al., 1937; Novembre et al., 2005). This PDE represents the distribution  $p(x, y, t)$  of the allele fre- 466 quency across a two dimensional  $(x, y)$  landscape at time  $t$ :

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

467 where

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

468 Here,  $\sigma$  is the diffusion coefficient,  $s$  is the selection coefficient, and  $d$  is the dominance coeffi- 469 cient (Novembre et al., 2005). We assumed an additive model and fixed  $d = 2s$  in all analyses below. 470 We call this "model A", but we also evaluated the fit of our data under more complex models which 471 are more flexible, and are described below.

472 Model B is a more general diffusion-reaction model, which incorporates distinct diffusion terms 473 in the longitudinal and latitudinal directions ( $\sigma_x$  and  $\sigma_y$ , respectively):

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

474 Model C is a generalization of model B that incorporates advection terms in the longitudinal and 475 latitudinal directions (see e.g. Cantrell and Cosner (2004) for a motivation of this type of model in 476 the context of spatial ecology):

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

477 Here,  $v_x$  and  $v_y$  represent the coefficients for advective velocity along the longitude and latitude 478 respectively.

479 In the Appendix, we motivate the construction of these equations using model C as an example, 480 and show that equation 4 can be obtained by taking an infinitesimal limit of a random walk on a 481 two-dimensional lattice, after including a reaction term due to selection. Models A and B are then 482 shown to be special cases of model C.

483 For evaluating the likelihood of the observed data, we use a binomial genotype sampling model.  
484 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  
485 carrying ancestral alleles, let  $d_i$  be the number of reads carry derived reads. Let  $(x_i, y_i)$  be the  
486 coordinates of the location from which individual  $i$  was sampled, and  $t_i$  its estimated age (e.g. from  
487 radiocarbon dating). Then, the likelihood for individual  $i$  can be computed as follows:

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

488 Here,  $p(x_i, y_i, t_i)$  is the solution to one of the partial differential equations described above (equa-  
489 tions (1), (2) or (4), depending on the process model chosen), evaluated at location  $(x_i, y_i)$  and time  
490  $t_i$ . 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  
491 binomial distribution, where  $n$  represents the ploidy level, which in this case is 2:

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

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

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

493 where  $M$  is the total number of individuals for which we have data,  $\mathbf{d}$  is the vector containing the  
494 derived read count for each individual and  $\mathbf{a}$  is the vector containing the ancestral read count  
495 for each individual. We computed genotype likelihoods directly on the BAM file read data, using  
496 the SAMtools genotype model (Li, 2011) implemented in the software ANGSD (Korneliussen et al.,  
497 2014).

498 When only randomly sampled pseudohaploid allele counts are available, we used a Bernoulli  
499 sampling likelihood (conditional on the genotype  $g_i$ ) on the right-hand side of equation 6 instead.  
500 Briefly, assuming that the probability of an individual having genotype  $g$  at a particular locus given  
501 the underlying allele frequency  $p$  follows a binomial distribution and that the probability of sam-  
502 pling a read given the genotype of an individual follows a Bernoulli distribution with probability of  
503 success  $\frac{1}{2}g$ , then the probability of sampling a read given the genotype follows a Bernoulli distribu-  
504 tion with probability of success  $p$ .

## 505 Map

506 We restricted the geographic area explored by our model fit to be between 30°N to 75°N, and be-  
507 tween 10°W and 80°E. For numerical calculations, we used a grid constructed using a resolution  
508 of approximately 1 grid cell per latitude and longitude. We used Haversine functions in order to  
509 transform the distance from degrees to kilometers between two geographic points. The diffusion  
510 of the allele frequency was disallowed in the map regions where the topology is negative (i.e. re-  
511 gions under water), based on ETOPO5 data (NOAA (1988)). For this reason we added land bridges  
512 between the European mainland and Sardinia, and between the mainland and Great Britain, in  
513 order to allow the allele to diffuse in these regions (see Figure A1).

## 514 Parameter search

515 Parameter optimization was done via maximum likelihood estimation with a two-layer optimiza-  
516 tion set-up. The first layer consists of a simulated annealing approach (Bélisle (1992)) starting from  
517 50 random points in the parameter space. The initial 50 points are sampled using latin hypercube  
518 sampling to ensure an even spread across the parameter space. The output of this fit was then  
519 fed to the L-BFGS-B algorithm to refine the parameter estimates around the obtained maximum  
520 and obtain confidence intervals for the selection, diffusion and advection parameters (Byrd et al.  
521 (1995)).

522 The parameters optimised were:

- 523 • the selection coefficient ( $s$ ), restricted to the range 0.001-0.1
- 524 • two dispersal parameters  $\sigma_x$  and  $\sigma_y$ , in the longitudinal and latitudinal directions respectively,
- 525 restricted to the range of 1-100 square-kilometers per generation
- 526 • the longitudinal and latitudinal advection coefficients  $v_x$  and  $v_y$ , respectively. As a form of
- 527 regularization, we set the range of explored values to be narrowly centered around zero:
- 528 -2.5 to 2.5 kilometers per generation
- 529 • the geographic origin of the allele, which is randomly initialized to be any of the 28 spatial
- 530 points shown in **Figure A2** at the start of the optimization process

531 We chose to construct our method in a way that uses the age of the allele as an input parameter  
532 rather than estimating it. We do this since there are multiple equally possible solutions with various  
533 combinations of allele age and selection coefficient values as shown in **Figure A3**. The latitude and  
534 longitude are discretized in our model in order to solve the differential equations numerically, thus  
535 the origin of a mutation is measured in terms of discrete units. For this reason, when using the  
536 L-BFGS-B algorithm, we fixed the previously estimated origin of the allele, and did not explore it  
537 during this second optimization layer. For numerical calculations we used the Livermore Solver for  
538 Ordinary Differential Equations (*Hindmarsh, 1983*) implemented in R package “deSolve” (*Soetaert*  
539 *et al., 2010a*), which is a general purpose solver that can handle both stiff and nonstiff systems. In  
540 case of stiff problems the solver uses a Jacobian matrix. Absorbing boundary conditions were used  
541 at the boundaries of the map. For visualisation purposes we masked the allele frequencies from  
542 areas with negative topology (i.e. areas covered by large bodies of water). Time was measured in  
543 generations, assuming 29 years per generation. During the optimization we scaled the time and  
544 the parameters by a factor of 10, which allowed us to decrease the execution time of the model.

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

Asymptotic 95% confidence intervals for a given parameter  $\theta_j$  were calculated using equation

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

551 where  $F(\theta)$  is an estimate of the observed Fisher information matrix (*Fisher, 1922; Efron and Hastie,*  
552 *2016; Casella and Berger, 2021*).

## 553 Implementation

554 The above described model was implemented in R version 3.6. To numerically solve the differential  
555 equations and obtain maximum likelihood estimates, we used the libraries *deSolve* (*Soetaert et al.,*  
556 *2010b*), *ReacTran* (*Soetaert and Meysman, 2012*) and *bbmle* (*Bolker and R Development Core Team,*  
557 *2020*). Scripts containing the code used in this paper are available on github:  
558 <https://github.com/RasaMukti/stepadna>

## 559 Individual-based simulations

560 For the individual-based spatiotemporal forward simulations, we first defined a spatial boundary  
561 for a population spread across a broad geographic region of Europe. In order to ensure a rea-  
562 sonably uniform distribution of individuals across this spatial range throughout the course of the  
563 simulation, we set the maximum distance for spatial competition and mating choice between indi-  
564 viduals to 250 km (translated, on a SLiM level, to the interaction parameter *maxDistance*), and the  
565 standard deviation of the normal distribution governing the spread of offspring from their parents  
566 at 25 km (leveraged in SLiM's *modifyChild()* callback function) (*Haller and Messer, 2019*). We note



567 that we have chosen the values of these parameters merely to ensure a uniform spread of individ-  
568 uals across a simulated landscape. They are not intended to represent realistic estimates for these  
569 parameters at any time in human history.

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

## 579 Acknowledgments

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

## 586 Competing interests

587 The authors declare that they have no conflict of interest.

## 588 References

- 589 **Albers PK**, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS*  
590 *biology*. 2020; 18(1):e3000586.
- 591 **Allentoft ME**, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H,  
592 Ahlström T, Vinner L, et al. Population genomics of bronze age Eurasia. *Nature*. 2015; 522(7555):167–172.
- 593 **Alonso S**, Izagirre N, Smith-Zubiaga I, Gardeazabal J, Díaz-Ramón JL, Díaz-Pérez JL, Zelenika D, Boyano MD, Smit  
594 N, De la Rúa C. Complex signatures of selection for the melanogenic loci TYR, TYRP1 and DCT in humans.  
595 *BMC evolutionary biology*. 2008; 8(1):1–14.
- 596 **Bergström A**, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecsek P, Chen Y, Felkel S, Hallast P, Kamm J, et al.  
597 Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020;  
598 367(6484).
- 599 **Bersaglieri T**, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN.  
600 Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human*  
601 *Genetics*. 2004; 74(6):1111–1120.
- 602 **Bolker B**, R Development Core Team. *bbmle: Tools for General Maximum Likelihood Estimation*; 2020, <https://CRAN.R-project.org/package=bbmle>, r package version 1.0.23.1.
- 604 **Bradburd GS**, Ralph PL. Spatial population genetics: it's about time. *Annual Review of Ecology, Evolution, and*  
605 *Systematics*. 2019; .
- 606 **Bradburd GS**, Ralph PL, Coop GM. A spatial framework for understanding population structure and admixture.  
607 *PLoS genetics*. 2016; 12(1):e1005703.
- 608 **Brown PE**, Roberts GO, Kåresen KF, Tonellato S. Blur-generated non-separable space-time models. *Journal of*  
609 *the Royal Statistical Society: Series B (Statistical Methodology)*. 2000; 62(4):847–860.
- 610 **Burger J**, Kirchner M, Bramanti B, Haak W, Thomas MG. Absence of the lactase-persistence-associated allele  
611 in early Neolithic Europeans. *Proceedings of the National Academy of Sciences*. 2007; 104(10):3736–3741.

- 612 **Burger J**, Link V, Blöcher J, Schulz A, Sell C, Pochon Z, Diekmann Y, Žegarac A, Hofmanová Z, Winkelbach  
613 L, Reyna-Blanco CS, Bieker V, Orschiedt J, Brinker U, Scheu A, Leuenberger C, Bertino TS, Bollongino R,  
614 Lidke G, Stefanović S, et al. Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongo-  
615 ing Strong Selection over the Last 3,000 Years. *Current Biology*. 2020 Nov; 30(21):4307–4315.e13. <https://www.sciencedirect.com/science/article/pii/S0960982220311878>, doi: 10.1016/j.cub.2020.08.033.
- 617 **Byrd RH**, Lu P, Nocedal J, Zhu C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM*  
618 *Journal on Scientific Computing*. 1995 Sep; 16(5):1190–1208. <https://epubs.siam.org/doi/10.1137/0916069>,  
619 doi: 10.1137/0916069, publisher: Society for Industrial and Applied Mathematics.
- 620 **Bélisle CJP**. Convergence Theorems for a Class of Simulated Annealing Algorithms on Rd. *Journal of Applied*  
621 *Probability*. 1992; 29(4):885–895. <https://www.jstor.org/stable/3214721>, doi: 10.2307/3214721, publisher:  
622 Applied Probability Trust.
- 623 **Cantrell RS**, Cosner C. *Spatial ecology via reaction-diffusion equations*. John Wiley & Sons; 2004.
- 624 **Casella G**, Berger RL. *Statistical inference*. Cengage Learning; 2021.
- 625 **Charati H**, Peng MS, Chen W, Yang XY, Jabbari Ori R, Aghajanzadeh M, Esmailzadeh A, Zhang YP. The evolu-  
626 tionary genetics of lactase persistence in seven ethnic groups across the Iranian plateau. *Human Genomics*.  
627 2019 Feb; 13(1):7. <https://doi.org/10.1186/s40246-019-0195-5>, doi: 10.1186/s40246-019-0195-5.
- 628 **Colin McEvedy RJ**. *Atlas of World Population History*. Great Britain: Penguin Books Ltd. and Allen Lane; 1978.
- 629 **Cressie N**, Wikle CK. *Statistics for spatio-temporal data*. John Wiley & Sons; 2015.
- 630 **Crow JF**, Kimura M, et al. *An introduction to population genetics theory*. An introduction to population genetics  
631 theory. 1970; .
- 632 **Dehasque M**, Ávila-Arcos MC, Díez-del Molino D, Fumagalli M, Guschanski K, Lorenzen ED, Malaspina AS,  
633 Marques-Bonet T, Martin MD, Murray GG, et al. Inference of natural selection from ancient DNA. *Evolution*  
634 *Letters*. 2020; 4(2):94–108.
- 635 **Efron B**, Hastie T. *Computer age statistical inference, vol. 5*. Cambridge University Press; 2016.
- 636 **Enattah NS**, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M,  
637 Khalil IF, et al. Independent introduction of two lactase-persistence alleles into human populations reflects  
638 different history of adaptation to milk culture. *The American Journal of Human Genetics*. 2008; 82(1):57–72.
- 639 **Enattah NS**, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with  
640 adult-type hypolactasia. *Nature genetics*. 2002; 30(2):233–237.
- 641 **Ewens WJ**. *Mathematical population genetics 1: theoretical introduction, vol. 27*. Springer Science & Business  
642 Media; 2012.
- 643 **Fisher RA**. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal So-*  
644 *ciety of London Series A, Containing Papers of a Mathematical or Physical Character*. 1922; 222(594-604):309–  
645 368.
- 646 **Fisher RA**. The wave of advance of advantageous genes. *Annals of eugenics*. 1937; 7(4):355–369.
- 647 **Gallejo Romero I**, Basu Mallick C, Liebert A, Crivellaro F, Chaubey G, Itan Y, Metspalu M, Easwarkhanth M,  
648 Pitchappan R, Vilems R, Reich D, Singh L, Thangaraj K, Thomas MG, Swallow DM, Mirazón Lahr M, Kivisild T.  
649 Herders of Indian and European Cattle Share Their Predominant Allele for Lactase Persistence. *Molecular*  
650 *Biology and Evolution*. 2012 Jan; 29(1):249–260. [https://academic.oup.com/mbe/article-lookup/doi/10.1093/](https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msr190)  
651 [molbev/msr190](https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msr190), doi: 10.1093/molbev/msr190.
- 652 **Gamba C**, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboróczki L, Kóvári I,  
653 Pap I, Anders A, et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nature*  
654 *communications*. 2014; 5(1):1–9.
- 655 **Gerbault P**, Moret C, Currat M, Sanchez-Mazas A. Impact of selection and demography on the diffusion of  
656 lactase persistence. *PLoS One*. 2009; 4(7):e6369.
- 657 **Günther T**, Malmström H, Svensson EM, Omrak A, Sánchez-Quinto F, Kilinç GM, Krzewińska M, Eriksson G,  
658 Fraser M, Edlund H, et al. Population genomics of Mesolithic Scandinavia: Investigating early postglacial  
659 migration routes and high-latitude adaptation. *PLoS biology*. 2018; 16(1):e2003703.

- 660 **Haak W**, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson  
661 K, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*.  
662 2015; 522(7555):207–211.
- 663 **Hallatschek O**, Fisher DS. Acceleration of evolutionary spread by long-range dispersal. *Proceedings of the*  
664 *National Academy of Sciences*. 2014; 111(46):E4911–E4919.
- 665 **Haller BC**, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular*  
666 *Biology and Evolution*. 2019 Mar; 36(3):632–637. <https://doi.org/10.1093/molbev/msy228>, doi: 10.1093/mol-  
667 bev/msy228.
- 668 **Hermisson J**, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic  
669 variation. *Genetics*. 2005; 169(4):2335–2352.
- 670 **Hermisson J**, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection  
671 footprints under rapid adaptation. *Methods in Ecology and Evolution*. 2017; 8(6):700–716.
- 672 **Heyer E**, Brazier L, Ségurel L, Hegay T, Austerlitz F, Quintana-Murci L, Georges M, Pasquet P, Veuille M. Lactase  
673 persistence in central Asia: phenotype, genotype, and evolution. *Human Biology*. 2011 Jun; 83(3):379–392.  
674 doi: 10.3378/027.083.0304.
- 675 **Hindmarsh AC**. ODEPACK, A Systematized Collection of ODE Solvers. *Scientific Computing*. 1983; p. 55–64.
- 676 **Hudjashov G**, Villems R, Kivisild T. Global patterns of diversity and selection in human tyrosinase gene. *PLoS*  
677 *One*. 2013; 8(9):e74307.
- 678 **Irving-Pease EK**, Muktupavela R, Dannemann M, Racimo F. Quantitative Paleogenetics: what can ancient DNA  
679 tell us about complex trait evolution? arXiv preprint arXiv:210502754. 2021; .
- 680 **Itan Y**, Jones BL, Ingram CJ, Swallow DM, Thomas MG. A worldwide correlation of lactase persistence phenotype  
681 and genotypes. *BMC Evolutionary Biology*. 2010 Feb; 10(1):36. <https://doi.org/10.1186/1471-2148-10-36>, doi:  
682 10.1186/1471-2148-10-36.
- 683 **Itan Y**, Powell A, Beaumont MA, Burger J, Thomas MG. The Origins of Lactase Persistence in Europe. *PLoS*  
684 *Computational Biology*. 2009 Aug; 5(8):e1000491. <https://dx.plos.org/10.1371/journal.pcbi.1000491>, doi:  
685 10.1371/journal.pcbi.1000491.
- 686 **Ju D**, Mathieson I. The evolution of skin pigmentation associated variation in West Eurasia. *bioRxiv*. 2020; .
- 687 **Karlin S**, Taylor H. A first course in Stochastic Processes. Academic Press, New York; 1975.
- 688 **Klopfstein S**, Currat M, Excoffier L. The fate of mutations surfing on the wave of a range expansion. *Molecular*  
689 *biology and evolution*. 2006; 23(3):482–490.
- 690 **Kolmogorov A**, Petrovskii I, Piskunov N. A Study of the Diffusion Equation with Increase in the Amount of  
691 Substance, and Its Application to a Biological Problem. *Byull Moskov Univ Ser AMat Mekh*. 1937; 1(6):1–26.
- 692 **Korneliussen TS**, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC bioin-*  
693 *formatics*. 2014; 15(1):356.
- 694 **Krüttli A**, Bouwman A, Akgül G, Della Casa P, Rühli F, Warinner C. Ancient DNA analysis reveals high frequency  
695 of European lactase persistence allele (T-13910) in medieval central europe. *PLoS One*. 2014; 9(1):e86251.
- 696 **Lao O**, De Gruijter J, van Duijn K, Navarro A, Kayser M. Signatures of positive selection in genes associated with  
697 human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of human*  
698 *genetics*. 2007; 71(3):354–369.
- 699 **Li H**. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical  
700 parameter estimation from sequencing data. *Bioinformatics*. 2011; 27(21):2987–2993.
- 701 **Liebert A**, López S, Jones BL, Montalva N, Gerbault P, Lau W, Thomas MG, Bradman N, Maniatis N, Swallow  
702 DM. World-wide distributions of lactase persistence alleles and the complex effects of recombination and  
703 selection. *Human Genetics*. 2017 Nov; 136(11-12):1445–1453. doi: 10.1007/s00439-017-1847-y.
- 704 **Loog L**, Lahr MM, Kovacevic M, Manica A, Eriksson A, Thomas MG. Estimating mobility using sparse data: Appli-  
705 cation to human genetic variation. *Proceedings of the National Academy of Sciences*. 2017; 114(46):12213–  
706 12218.

- 707 **Malaspinas AS**, Malaspinas O, Evans SN, Slatkin M. Estimating allele age and selection coefficient from time-  
708 serial data. *Genetics*. 2012; 192(2):599–607.
- 709 **Marchi N**, Menecier P, Georges M, Lafosse S, Hegay T, Dorzhu C, Chichlo B, Ségurel L, Heyer E. Close inbreeding  
710 and low genetic diversity in Inner Asian human populations despite geographical exogamy. *Scientific Reports*.  
711 2018 Jun; 8(1):9397. doi: 10.1038/s41598-018-27047-3.
- 712 **Marciniak S**, Perry GH. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews*  
713 *Genetics*. 2017; 18(11):659.
- 714 **Margaryan A**, Lawson DJ, Sikora M, Racimo F, Rasmussen S, Moltke I, Cassidy LM, Jørsboe E, Ingason A, Peder-  
715 sen MW, et al. Population genomics of the Viking world. *Nature*. 2020; 585(7825):390–396.
- 716 **Mathieson I**. Estimating time-varying selection coefficients from time series data of allele frequencies. *Genet-*  
717 *ics*; 2020.
- 718 **Mathieson I**, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes  
719 D, Novak M, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 2015; 528(7583):499–  
720 503.
- 721 **Mathieson I**, McVean G. Demography and the age of rare variants. *PLoS Genet*. 2014; 10(8):e1004528.
- 722 **Mathieson S**, Mathieson I. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology*  
723 *and Evolution*. 2018 Dec; 35(12):2957–2970. <https://academic.oup.com/mbe/article/35/12/2957/5112969>, doi:  
724 10.1093/molbev/msy180.
- 725 **Nakagome S**, Hudson RR, Rienzo AD. Inferring the model and onset of natural selection under varying popula-  
726 tion size from the site frequency spectrum and haplotype structure. *The Royal Society*. 2019 Feb; 286(1896):8.
- 727 **NOAA BC National Geophysical Data Center**, Data Announcement 88-MGG-02, , Digital relief of the Surface  
728 of the Earth.; 1988. <https://www.ngdc.noaa.gov/mgg/global/etopo5.HTML>, publisher: U.S. Department of  
729 Commerce.
- 730 **Novembre J**, Galvani AP, Slatkin M. The geographic spread of the CCR5  $\Delta$ 32 HIV-resistance allele. *PLoS Biol*.  
731 2005; 3(11):e339.
- 732 **Novembre J**, Han E. Human population structure and the adaptive response to pathogen-induced selection  
733 pressures. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2012; 367(1590):878–886.
- 734 **Okubo A**, et al. Diffusion and ecological problems: mathematical models. . 1980; .
- 735 **Othmer HG**, Dunbar SR, Alt W. Models of dispersal in biological systems. *Journal of mathematical biology*.  
736 1988; 26(3):263–298.
- 737 **Painter KJ**, Hillen T. From Random Walks to Fully Anisotropic Diffusion Models for Cell and Animal Movement.  
738 In: *Cell Movement* Springer; 2018.p. 103–141.
- 739 **Peter BM**, Huerta-Sanchez E, Nielsen R. Distinguishing between Selective Sweeps from Standing Variation and  
740 from a De Novo Mutation. *PLOS Genetics*. 2012 Oct; 8(10):e1003011. [https://journals.plos.org/plosgenetics/](https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003011)  
741 [article?id=10.1371/journal.pgen.1003011](https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003011), doi: 10.1371/journal.pgen.1003011, publisher: Public Library of Sci-  
742 ence.
- 743 **Petr M**, bodkan/slendr; 2021. <https://github.com/bodkan/slendr>, original-date: 2021-02-18T15:07:15Z.
- 744 **Pickrell JK**, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW,  
745 et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome research*.  
746 2009; 19(5):826–837.
- 747 **Platt A**, Pivrotto A, Knoblach J, Hey J. An estimator of first coalescent time reveals selection on young variants  
748 and large heterogeneity in rare allele ages among human populations. *PLoS genetics*. 2019; 15(8):e1008340.
- 749 **Pritchard JK**, Di Rienzo A. Adaptation—not by sweeps alone. *Nature Reviews Genetics*. 2010; 11(10):665–667.
- 750 **Racimo F**, Sikora M, Vander Linden M, Schroeder H, Lalueza-Fox C. Beyond broad strokes: sociocultural insights  
751 from the study of ancient genomes. *Nature Reviews Genetics*. 2020; 21(6):355–366.



- 752 **Racimo F**, Woodbridge J, Fyfe RM, Sikora M, Sjögren KG, Kristiansen K, Vander Linden M. The spatiotemporal  
753 spread of human migrations during the European Holocene. *Proceedings of the National Academy of*  
754 *Sciences*. 2020; 117(16):8989–9000.
- 755 **Ralph P**, Coop G. Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*.  
756 2010; 186(2):647–668.
- 757 **Ray N**, Excoffier L. Inferring past demography using spatially explicit population genetic models. *Human*  
758 *Biology*. 2009; 81(3):141–157.
- 759 **Reich D**, Mallick S, Allen Ancient DNA Resource (AADR): Downloadable genotypes  
760 of present-day and ancient DNA data; 2019. [https://reich.hms.harvard.edu/  
761 allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data](https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data).
- 762 **Sabeti PC**, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R,  
763 et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*.  
764 2007; 449(7164):913–918.
- 765 **Sabeti PC**, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N,  
766 Cooper R, Reich D, Altshuler D, O'Brien S, Lander ES. The Case for Selection at CCR5-32. *PLoS Biology*. 2005  
767 Nov; 3(11). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1275522/>, doi: 10.1371/journal.pbio.0030378.
- 768 **Ségurel L**, Bon C. On the evolution of lactase persistence in humans. *Annual review of genomics and human*  
769 *genetics*. 2017; 18.
- 770 **Segurel L**, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, Fabre A, Hegay T, Heyer E. Why and when  
771 was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biology*.  
772 2020; 18(6):e3000742.
- 773 **Slatkin M**, Rannala B. Estimating allele age. *Annual review of genomics and human genetics*. 2000; 1(1):225–  
774 249.
- 775 **Soetaert K**, Meysman F. Reactive transport in aquatic ecosystems: Rapid model prototyping in the open source  
776 software R. *Environmental Modelling Software*. 2012; 32:49–60.
- 777 **Soetaert K**, Petzoldt T, Setzer RW. Solving Differential Equations in R: Package deSolve. *Journal of*  
778 *Statistical Software*. 2010; 33(9):1–25. <https://www.jstatsoft.org/index.php/jss/article/view/v033i09>, doi:  
779 10.18637/jss.v033.i09.
- 780 **Soetaert K**, Petzoldt T, Setzer RW. Solving Differential Equations in R: Package deSolve. *Journal of Statistical*  
781 *Software*. 2010; 33(9):1–25. <http://www.jstatsoft.org/v33/i09>, doi: 10.18637/jss.v033.i09.
- 782 **Stephens JC**, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R,  
783 Schriml L, et al. Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes.  
784 *The American Journal of Human Genetics*. 1998; 62(6):1507–1515.
- 785 **Stern AJ**, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele fre-  
786 quency trajectories from DNA sequence data. *PLOS Genetics*. 2019 Sep; 15(9):e1008384. [https://dx.plos.org/  
787 10.1371/journal.pgen.1008384](https://dx.plos.org/10.1371/journal.pgen.1008384), doi: 10.1371/journal.pgen.1008384.
- 788 **Szpak M**, Xue Y, Ayub Q, Tyler-Smith C. How well do we understand the basis of classic selective sweeps in  
789 humans? *FEBS letters*. 2019; 593(13):1431–1448.
- 790 **Tishkoff SA**, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman  
791 M, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*. 2007;  
792 39(1):31–40.
- 793 **Voight BF**, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS*  
794 *Biol*. 2006; 4(3):e72.
- 795 **Wikle CK**, Zammit-Mangion A, Cressie N. *Spatio-temporal Statistics with R*. CRC Press; 2019.
- 796 **Wilde S**, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hoffelder N, Potekhina ID, Schier W,  
797 Thomas MG, Burger J. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans  
798 during the last 5,000 y. *Proceedings of the National Academy of Sciences*. 2014 Apr; 111(13):4832–4837.  
799 <http://www.pnas.org/lookup/doi/10.1073/pnas.1316513111>, doi: 10.1073/pnas.1316513111.

## 800 Appendix

## 801 Appendix 1

802 Here, we motivate the construction of model C as a large scale limit of a random walk model  
 803 on a lattice (*Karlin and Taylor, 1975; Cantrell and Cosner, 2004*). We think of the allele fre-  
 804 quency as a variable  $p$  that can increase in magnitude due to its inherent advantage (se-  
 805 lection), spread across a landscape (diffusion) or move directionally as a consequence of  
 806 migration (advection). We imagine a lattice composed of small square cells of size  $\Delta x \times \Delta y$ ,  
 807 where a certain amount of allele frequency  $p$  can occur at a given time point  $t$ . At each small  
 808 time step (of duration  $\Delta t$ ), inflow and outflow of  $p$  can occur in the  $x$ -direction with proba-  
 809 bility  $h$  or in the  $y$ -direction with probability  $1-h$ , and the magnitude of these flows depend  
 810 on the amount of  $p$  present in neighboring cells. If flow of  $p$  is along the  $x$ -axis, it does so in  
 811 the positive direction with probability  $\alpha$  and in the negative direction with probability  $1 - \alpha$ .  
 812 If flow of  $p$  is along the  $y$ -axis, it does so in the positive direction with probability  $\beta$  and in  
 813 the negative direction with probability  $1 - \beta$ . The allele frequency can also increase in mag-  
 814 nitude locally, via a function  $\gamma()$  that depends on its dominance ( $d$ ), selection coefficient ( $s$ )  
 815 and current magnitude ( $p(x, y, t)$ ). Then, we obtain:

$$816 \begin{aligned} 817 p(x, y, t + \Delta t) = & p(x, y, t) + \gamma(p(x, y, t), s, d)\Delta t + \\ 818 & h\alpha p(x - \Delta x, y, t) + h(1 - \alpha)p(x + \Delta x, y, t) + \\ 819 & (1 - h)\beta p(x, y - \Delta y, t) + (1 - h)(1 - \beta)p(x, y + \Delta y, t) \end{aligned} \quad (8)$$

820 We can also write this as:

$$821 \begin{aligned} 822 p(x, y, t + \Delta t) - p(x, y, t) = & h \left( \frac{1}{2} - \alpha \right) [p(x + \Delta x, y, t) - p(x - \Delta x, y, t)] + \\ 823 & (1 - h) \left( \frac{1}{2} - \beta \right) [p(x, y + \Delta y, t) - p(x, y - \Delta y, t)] + \\ 824 & h \frac{1}{2} [p(x + \Delta x, y, t) - 2p(x, y, t) + p(x - \Delta x, y, t)] + \\ & (1 - h) \frac{1}{2} [p(x, y + \Delta y, t) - 2p(x, y, t) + p(x, y - \Delta y, t)] + \\ & \gamma(p(x, y, t), s, d)\Delta t \end{aligned} \quad (9)$$

825 If we divide both sides by  $\Delta t$  and take the limit of infinitesimally small  $\Delta x$ ,  $\Delta y$  and  $\Delta t$ ,  
 826 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:

$$827 \frac{\partial p}{\partial t} = \frac{1}{2} h \lambda_x \frac{\partial^2 p}{\partial x^2} + \frac{1}{2} (1 - h) \lambda_y \frac{\partial^2 p}{\partial y^2} + h(1 - 2\alpha) u_x \frac{\partial p}{\partial x} + (1 - h)(1 - 2\beta) u_y \frac{\partial p}{\partial y} + \gamma(p(x, y, t), s, d) \quad (10)$$

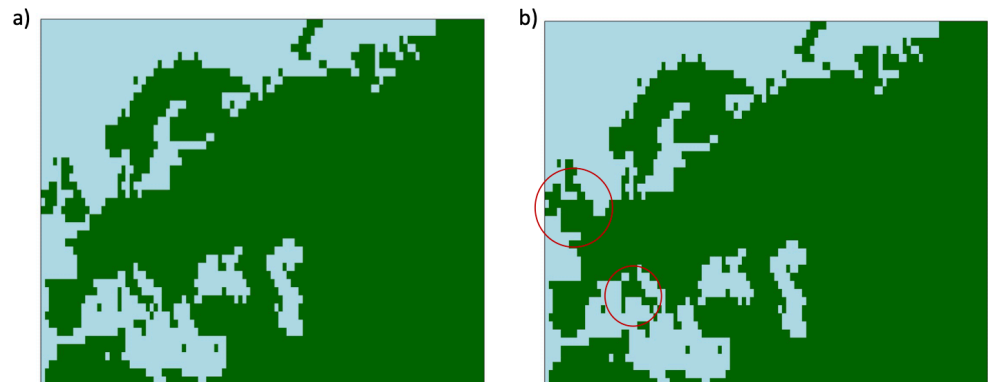
831 where  $u_x = \frac{\Delta x}{\Delta t}$ ,  $u_y = \frac{\Delta y}{\Delta t}$ ,  $\lambda_x = \frac{\Delta x^2}{\Delta t}$ ,  $\lambda_y = \frac{\Delta y^2}{\Delta t}$ .

832 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 equa-  
 833 tion 4. Thus, we can see that the squared diffusion coefficient  $\sigma_x^2$  depends on the square of  
 834 the length of the cells in the  $x$ -axis relative to the duration of a time step ( $\lambda_x$ ), and on the  
 835 probability that flows occurs in the  $x$ -axis at a given time step ( $h$ ). Similarly, the squared  
 836 diffusion coefficient  $\sigma_y^2$  depends on the square of the length of the cells in the  $y$ -axis rela-  
 837 tive to the duration of a time step ( $\lambda_y$ ), and on the probability that flows occurs in the  $y$ -axis  
 838 at a given time step ( $1 - h$ ). The advection coefficient  $v_x$  depends on the advective velocity  
 839 along the  $x$ -axis ( $u_x$ ) as well as on the probability of flow occurring along the  $x$ -axis ( $h$ ) and  
 840 the directional bias  $1 - 2\alpha$ , which depends on the probability that flow occurs in the positive  
 841  $x$ -direction ( $\alpha$ ). Finally, the advection coefficient  $v_y$  depends on the advective velocity along  
 842 the  $y$ -axis ( $u_y$ ) as well as on the probability of flow occurring along the  $y$ -axis ( $1 - h$ ) and the  
 843 directional bias  $1 - 2\beta$ , which depends on the probability that flow occurs in the positive  
 844  $y$ -direction ( $\beta$ ).

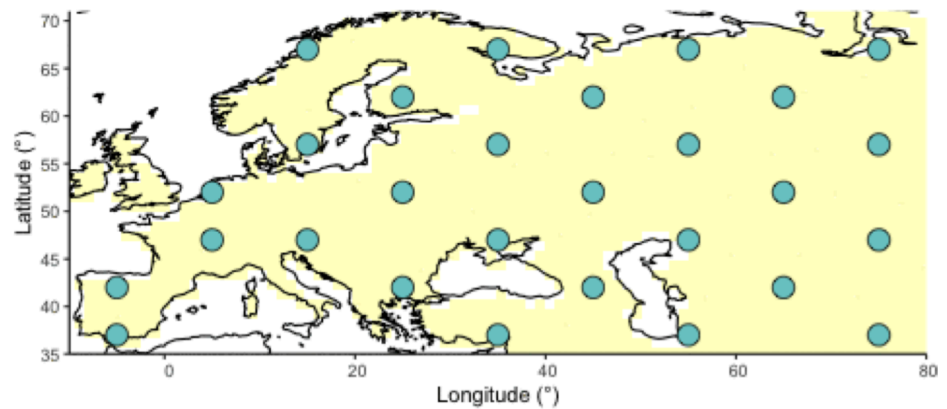


845 We can recover model B as a special case of model C if we fix  $\alpha = \beta = \frac{1}{2}$ , assuming isotropy  
846 in the two directions, so that  $\Delta x = \Delta y$ . We can also recover model A if we additionally fix  
847  $h = \frac{1}{2}$ .

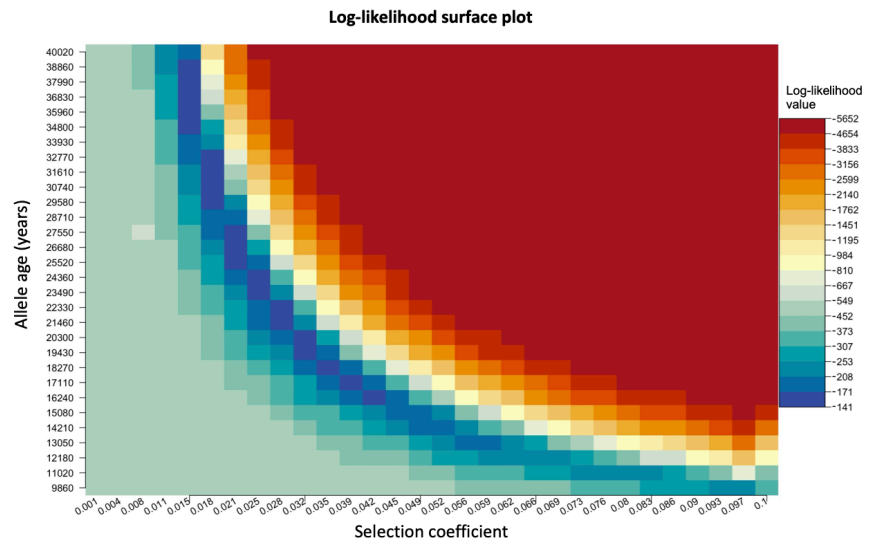
848 **Appendix 2**



**Figure A1.** 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 A2.** 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 A3.** Log-likelihood as a function of selection coefficient and age of the allele. Dark blue regions correspond to optimal solutions.

Simulation		$s$	$\sigma_x$ (km/gen)	$\sigma_y$ (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 A1.** 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		$s$	$\sigma_x$ (km/gen)	$\sigma_y$ (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 A2.** 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.

	Sampling scheme	$s$ (95% CI)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (km/gen) (95% CI)	$v_x$ (km/gen) (95% CI)	$v_y$ (km/gen) (95% CI)	Long	Lat
Sample age >5000	Homogeneous 75%/25%	0.0385 (0.0364–0.0406)	24.592 (14.9174–34.2675)	16.194 (4.8309–27.5568)	-1.209 (-1.6947–-0.7229)	-1.555 (-2.0804–-1.0294)	44	48
Sample age <5000		0.0261 (0.0201–0.0321)	45.725 (19.0071–72.4437)	11.950 (0.0000–27.2152)	2.499 (2.4993–2.4996)	-0.905 (-2.3876–0.5783)		
Sample age >5000	Homogeneous 50%/50%	0.0379 (0.0364–0.0394)	23.071 (0.0000–66.2585)	11.455 (0.0000–26.4585)	-0.827 (-3.2669–1.6124)	-0.934 (-1.5199–-0.3476)	46	51
Sample age <5000		0.0339 (0.0292–0.0385)	33.944 (13.1707–54.7167)	6.183 (0.0000–13.3805)	0.478 (-0.5529–1.5089)	0.315 (-0.1828–0.8127)		
Sample age >5000	Homogeneous 25%/75%	0.0379 (0.0364–0.0394)	22.619 (14.8534–30.3839)	13.588 (6.1189–21.0574)	-1.400 (-1.9213–0.8782)	-1.021 (-1.4258–-0.6161)	46	50
Sample age <5000		0.0322 (0.0257–0.0388)	70.446 (24.6065–116.2854)	3.786 (0.0000–-21.6379)	2.499 (2.4984–2.4987)	-0.990 (-2.0881–0.1079)		
Sample age >5000	Intermediate 75%/25%	0.0378 (0.0378–0.0378)	20.905 (20.904–20.9065)	14.583 (14.5818–14.5844)	-1.069 (-1.0687–-1.0684)	-0.547 (-0.5468–-0.5467)	44	52
Sample age <5000		0.0342 (0.0276–0.0407)	70.405 (29.0665–111.7428)	1.936 (0.0000–18.9234)	2.500 (2.4995–2.4998)	-1.865 (-3.0637–-0.6655)		
Sample age >5000	Intermediate 50%/50%	0.0379 (0.0378–0.0379)	93.136 (93.0316–93.2406)	10.99 (10.9808–10.9994)	1.103 (1.1009–1.1048)	0.695 (0.6939–0.6954)	34	57
Sample age <5000		0.0327 (0.0288–0.0367)	22.409 (0.0000–-69.8122)	18.11 (7.9198–28.2994)	2.496 (2.4962–2.4965)	-2.499 (-2.4992–-2.4989)		
Sample age >5000	Intermediate 25%/75%	0.0386 (0.0371–0.0402)	21.385 (14.0301–28.7407)	12.335 (3.3756–21.2943)	-1.028 (-1.4606–-0.5956)	-1.307 (-1.7696–-0.845)	43	49
Sample age <5000		0.0295 (0.026–0.0329)	21.197 (6.0797–36.3142)	11.318 (2.651–19.9851)	2.500 (2.4997–2.5000)	-0.757 (-1.391–-0.123)		
Sample age >5000	Extreme 75%/25%	0.0362 (0.0336–0.0389)	33.07 (0.0000–78.1418)	26.744 (0.0000–155.2413)	-0.087 (-0.3579–0.1832)	-2.001 (-4.7547–0.7524)	39	46
Sample age <5000		0.0299 (0.0266–0.0332)	16.702 (0.0000–40.5995)	3.048 (0.0000–13.034)	2.197 (0.8463–3.5479)	-2.499 (-2.4995–-2.4992)		
Sample age >5000	Extreme 50%/50%	0.0392 (0.0369–0.0416)	95.472 (95.2997–95.6441)	11.22 (5.3235–17.1167)	1.633 (-2.5434–5.8102)	0.258 (0.0818–0.434)	36	57
Sample age <5000		0.0355 (0.0314–0.0396)	11.756 (10.3763–13.1361)	11.817 (10.1474–13.4863)	2.500 (2.2069–2.7928)	-0.362 (-0.4325–-0.2919)		
Sample age >5000	Extreme 25%/75%	0.047 (0.047–0.047)	7.909 (7.9075–7.9106)	5.941 (5.9403–5.942)	0.454 (0.4537–0.4538)	-2.273 (-2.2732–-2.2729)	34	38
Sample age <5000		0.0434 (0.0385–0.0483)	40.097 (33.7903–46.4034)	12.118 (5.781–18.4555)	2.500 (1.6706–3.3285)	-1.435 (-2.3958–-0.4736)		
	True parameter values	0.04	25	10	1.8	-0.8	25	52

**Table A3.** Parameter value estimates for each of the 9 clustering schemes and true parameter values used to generate the deterministic simulation. The age of the allele was set to 17,400 years. Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

$s$ (95% CI)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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 A4.** 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)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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 A5.** 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)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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 A6.** Parameter estimates for rs4988235(T) using the allele age inferred in *Albers and McVean (2020)*. Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

	$s$ (95% CI)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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 A7.** 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 and 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 and McVean (2020)*. Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.



	$s$ (95% CI)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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)		

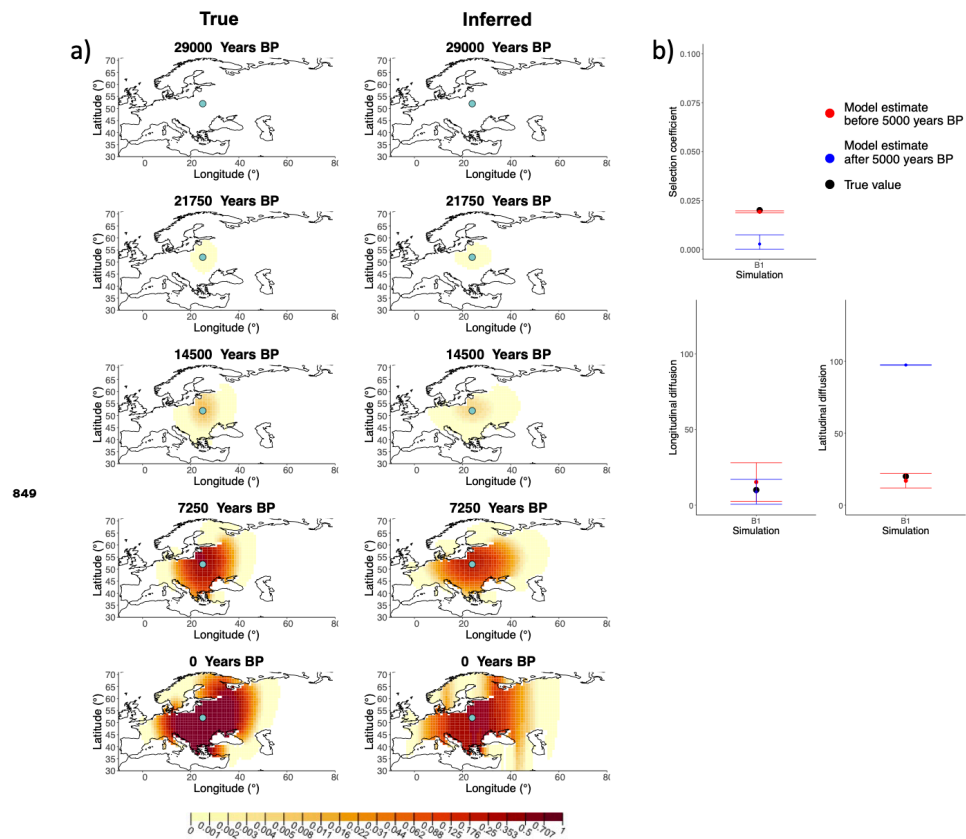
**Table A8.** Summary of parameter estimates for rs4988235(T) when the origin of the allele is forced to be at different points in the map (top panel corresponds to the original fit for the geographic position). In all cases, the estimated age of allele that was inputted into the model is 7,441 years ago. Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

	$s$ (95% CI)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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)			

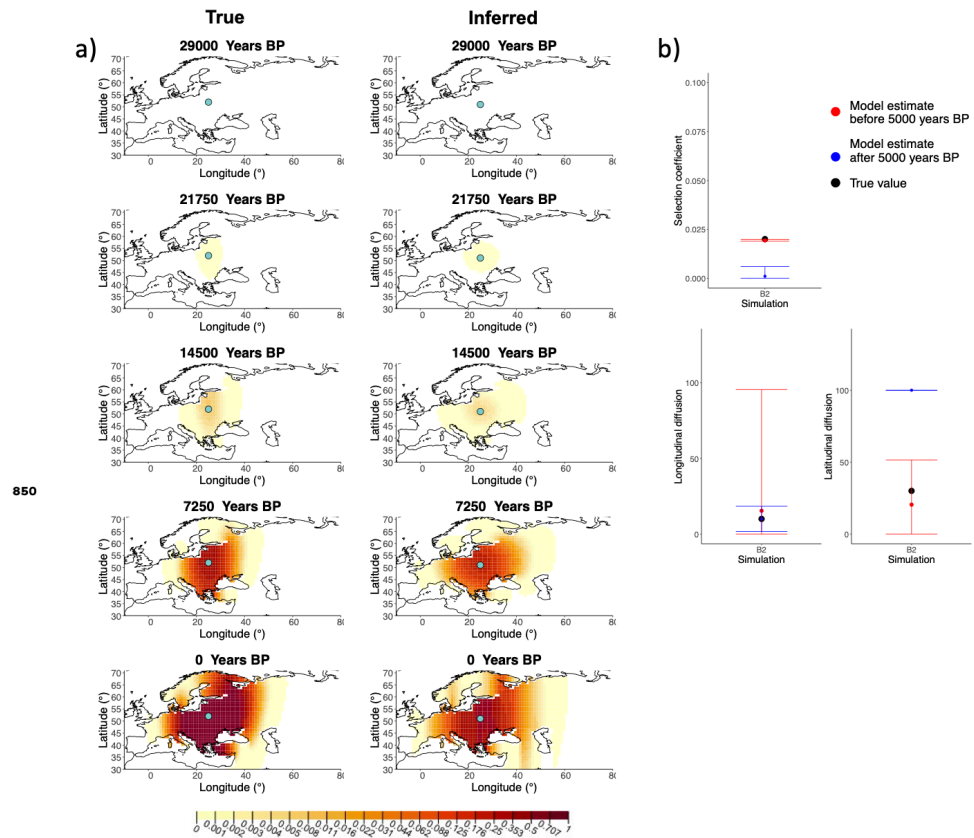
**Table A9.** Parameter estimates for rs4988235(T) using the geographic origin of the allele inferred in *Itan et al. (2009)*. Columns named "Long" and "Lat" indicate the longitude and latitude of the geographic origin of the allele, respectively.

	$s$ (95% CI)	$\sigma_x$ (km/gen) (95% CI)	$\sigma_y$ (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 A10.** 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.

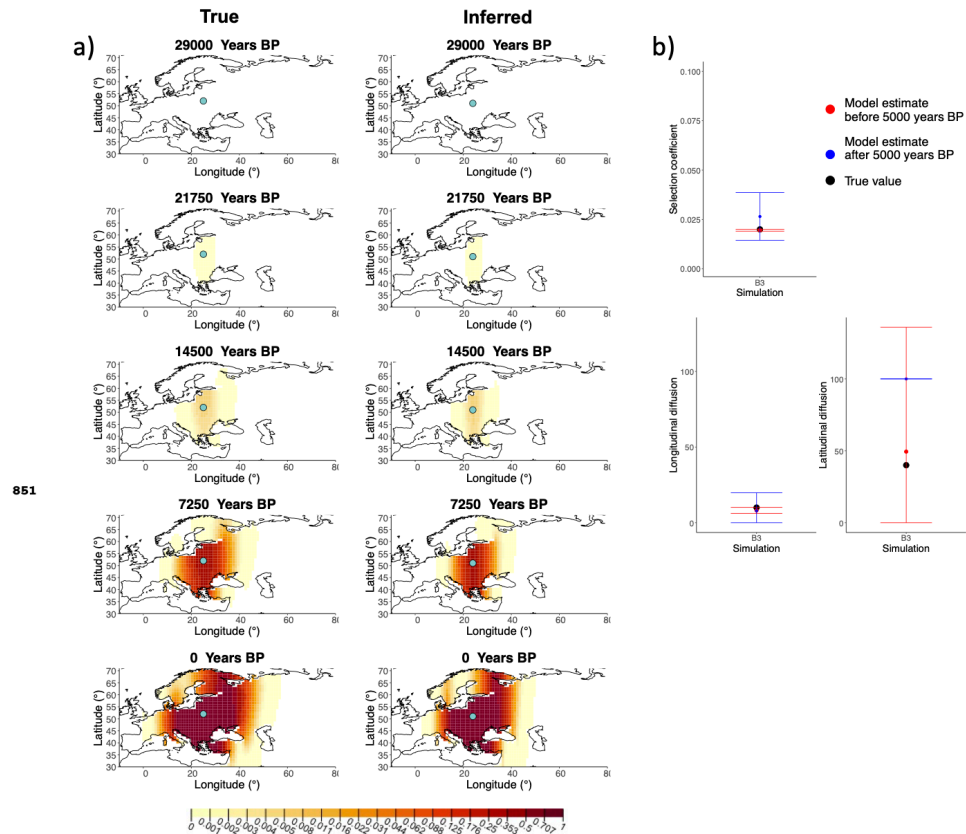


**Figure 1-Figure supplement 1.** a) Comparison of true and inferred allele frequency dynamics for simulation B1. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in **Table A1**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

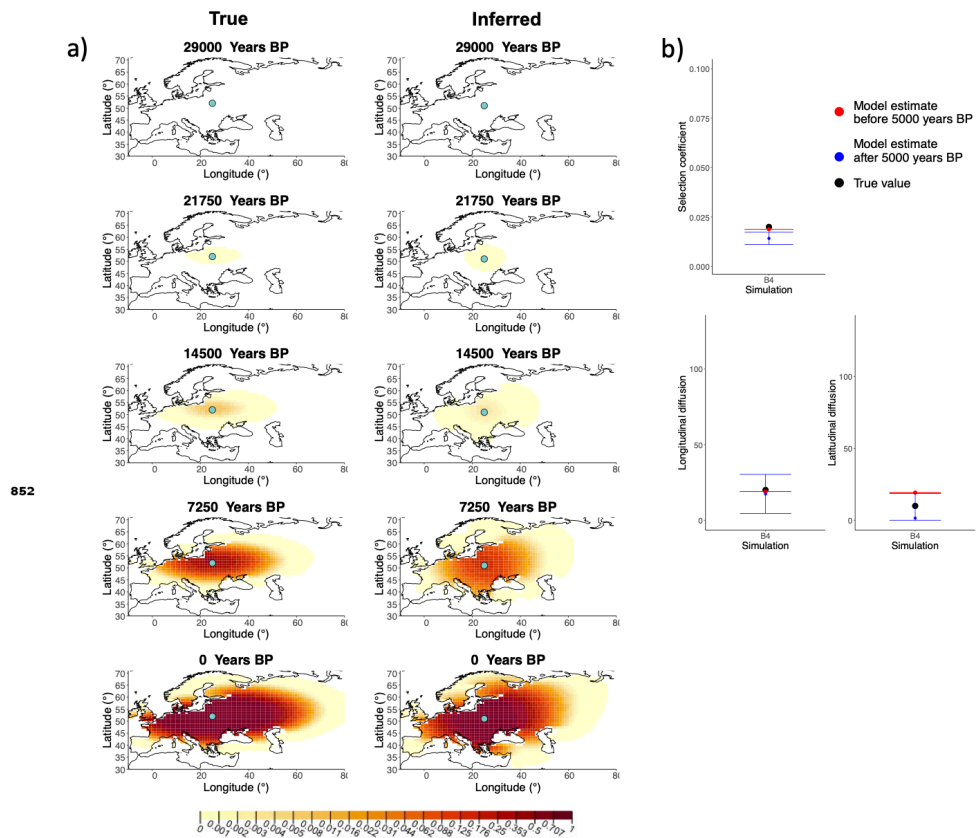


**Figure 1-Figure supplement 2.** a) Comparison of true and inferred allele frequency dynamics for simulation B2. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in **Table A1**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

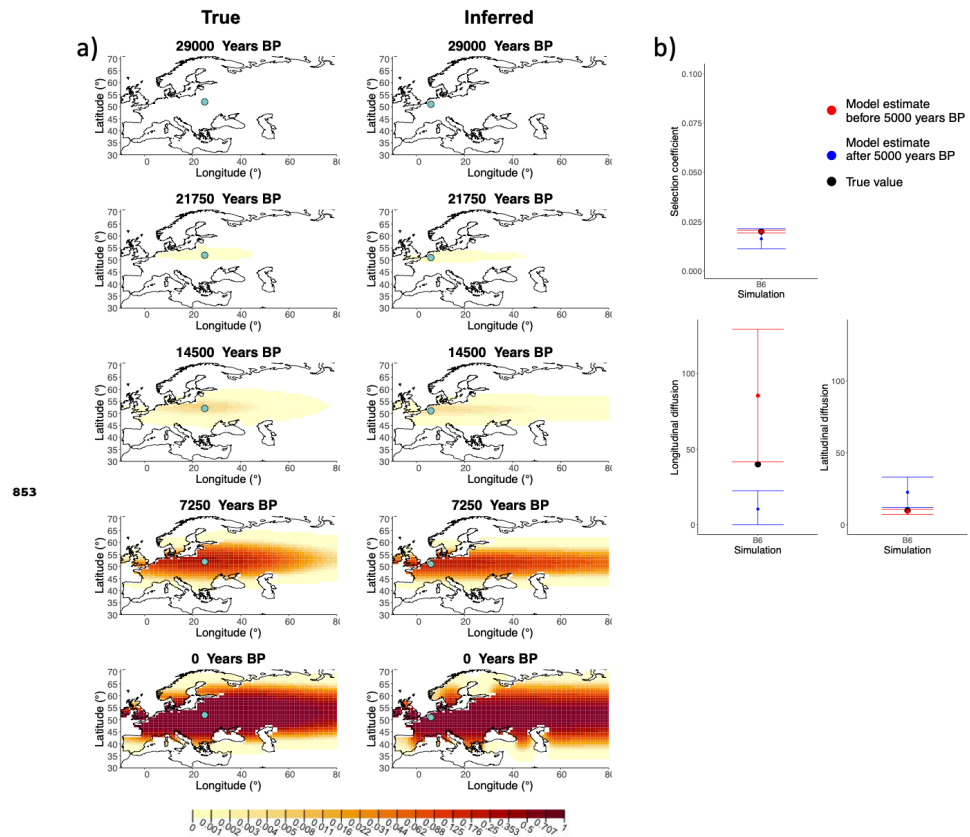




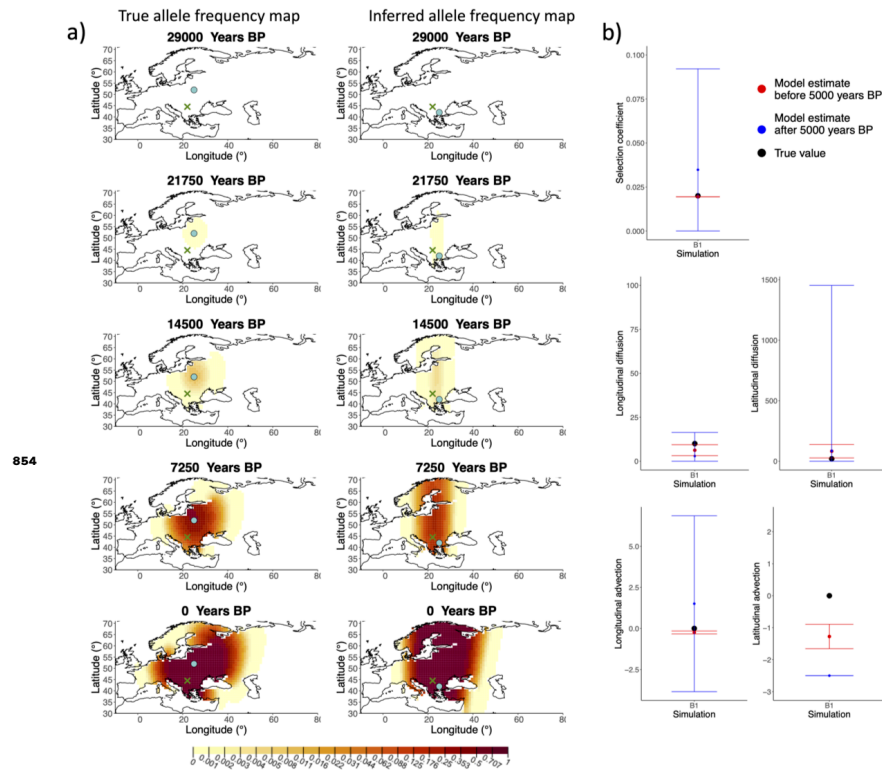
**Figure 1-Figure supplement 3.** a) Comparison of true and inferred allele frequency dynamics for simulation B3. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in **Table A1**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.



**Figure 1-Figure supplement 4.** 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 A1**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

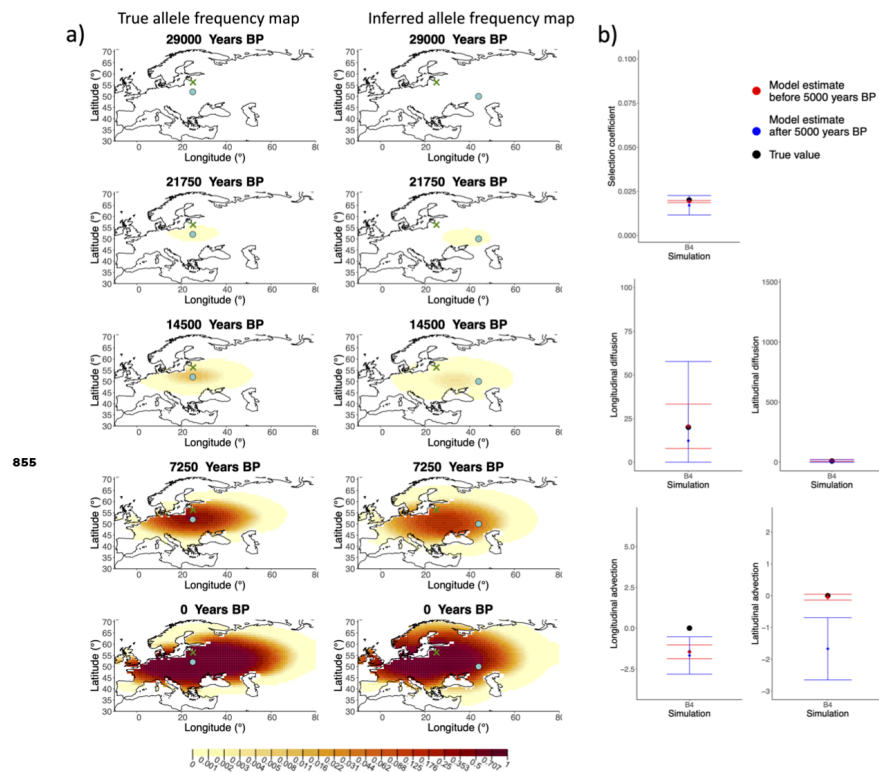


**Figure 1-Figure supplement 5.** a) Comparison of true and inferred allele frequency dynamics for simulation B6. The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in **Table A1**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

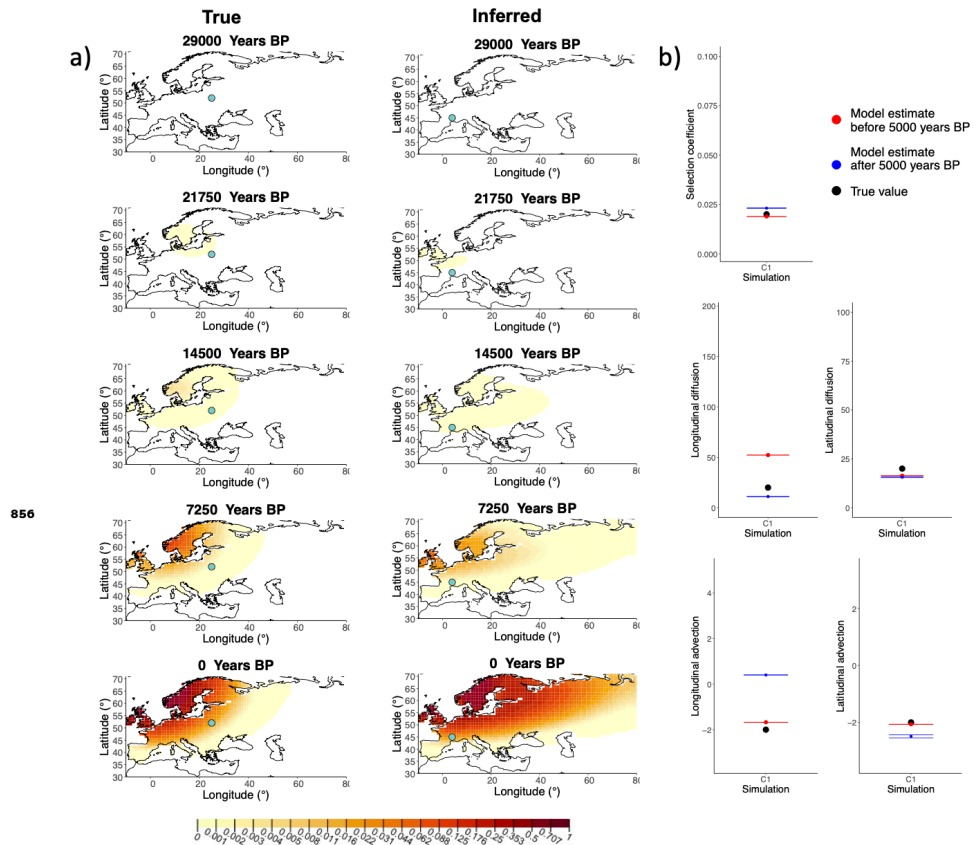


**Figure 1-Figure supplement 6.** a) Comparison of true allele frequency dynamics for simulation B1 and those inferred by the model C. The green dot shows the origin of the derived allele and the cross represents the location of the first individual that carried it. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

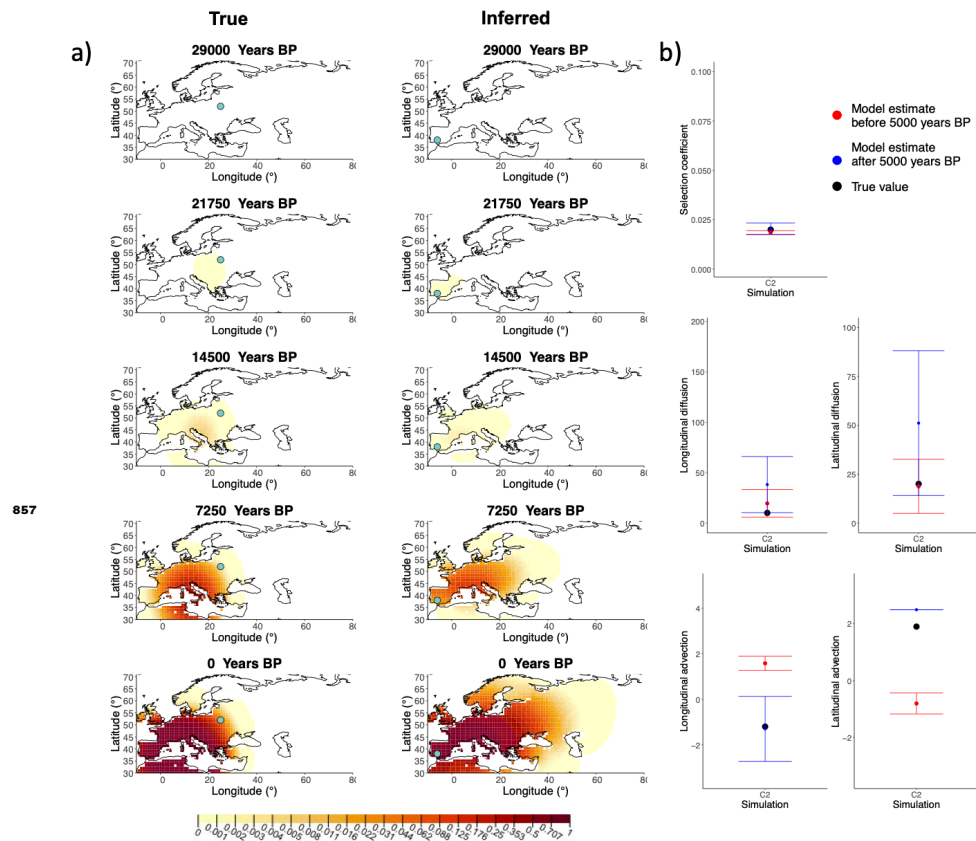




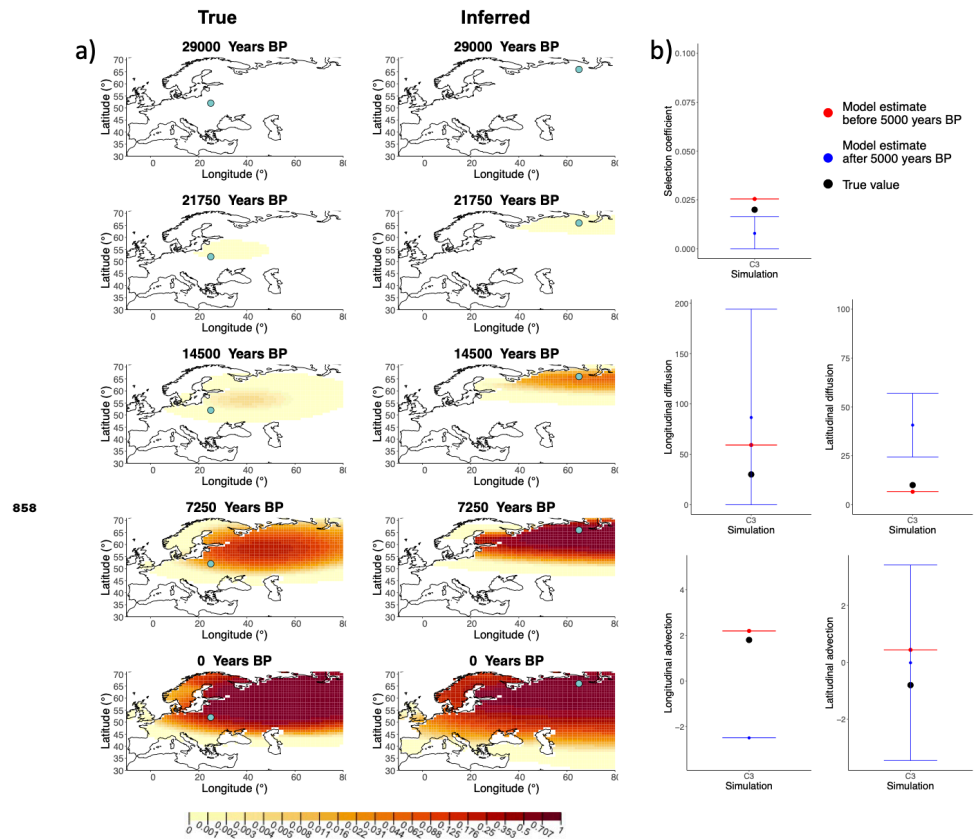
**Figure 1-Figure supplement 7.** a) Comparison of true allele frequency dynamics for simulation B4 and those inferred by the model C. The green dot corresponds to the origin of the allele and the cross represents the first sample having the derived variant. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.



**Figure 2-Figure supplement 1.** 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 A2**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

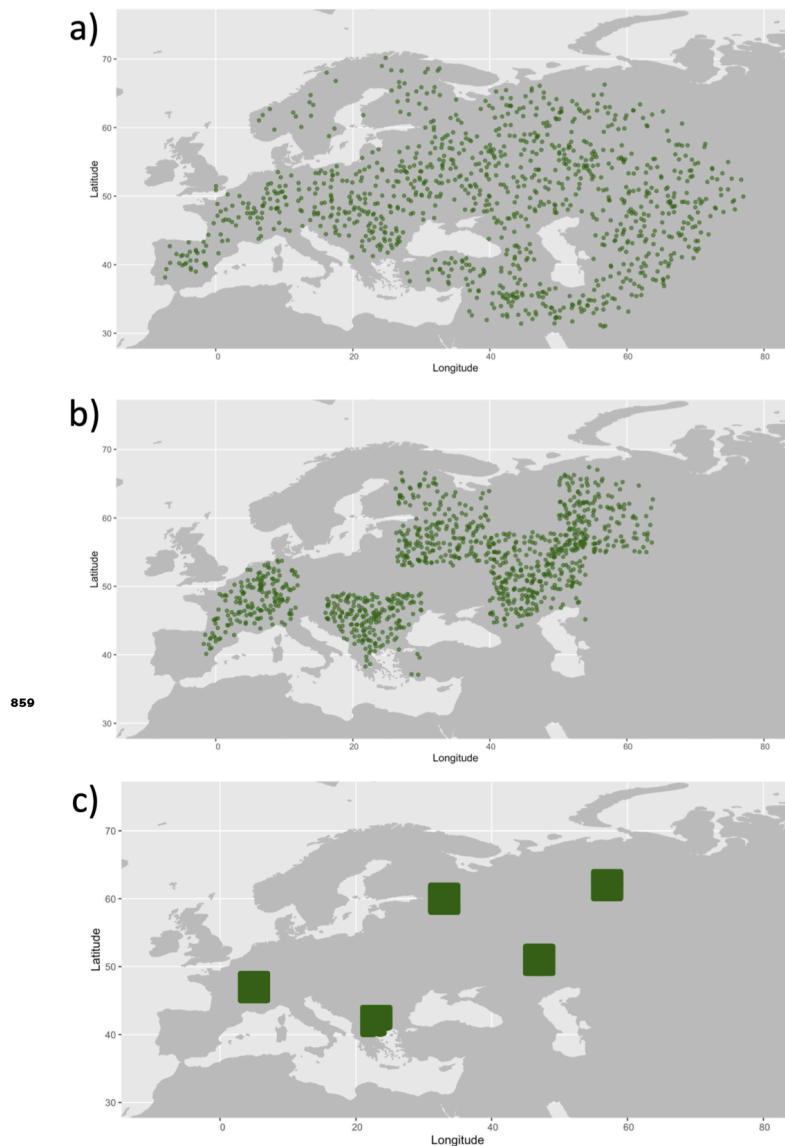


**Figure 2-Figure supplement 2.** a) Comparison of true and inferred allele frequency dynamics for one of the simulations including advection (C2). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in **Table A2**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

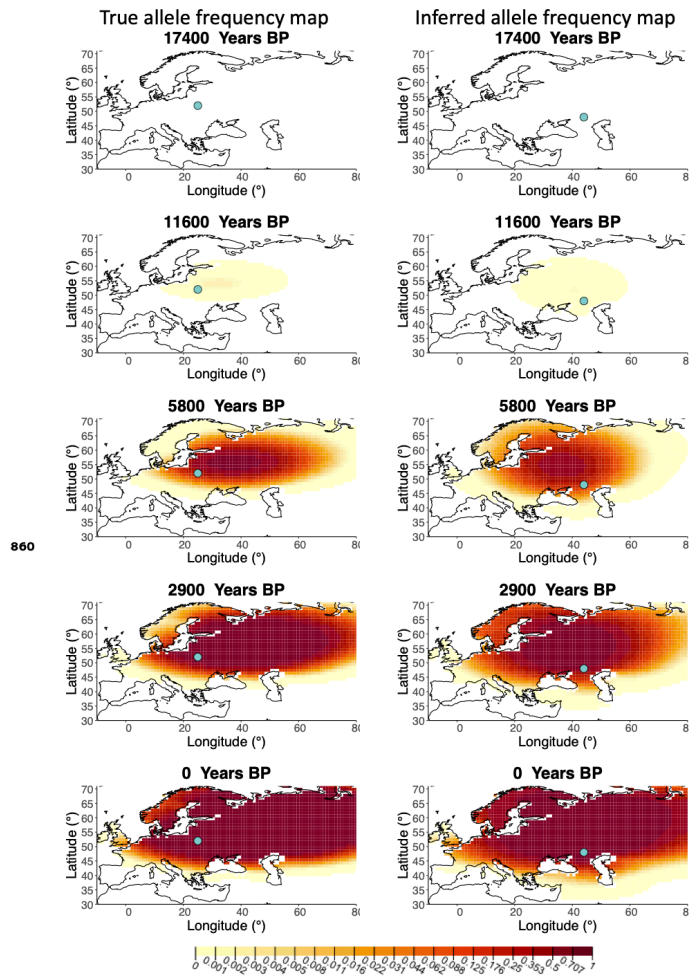


**Figure 2-Figure supplement 3.** a) Comparison of true and inferred allele frequency dynamics for one of the simulations including advection (C3). The green dot corresponds to the origin of the allele. The parameter values used to generate the frequency surface maps are summarised in **Table A2**. b) Comparison of true parameter values and model estimates. Whiskers represent 95% confidence intervals.

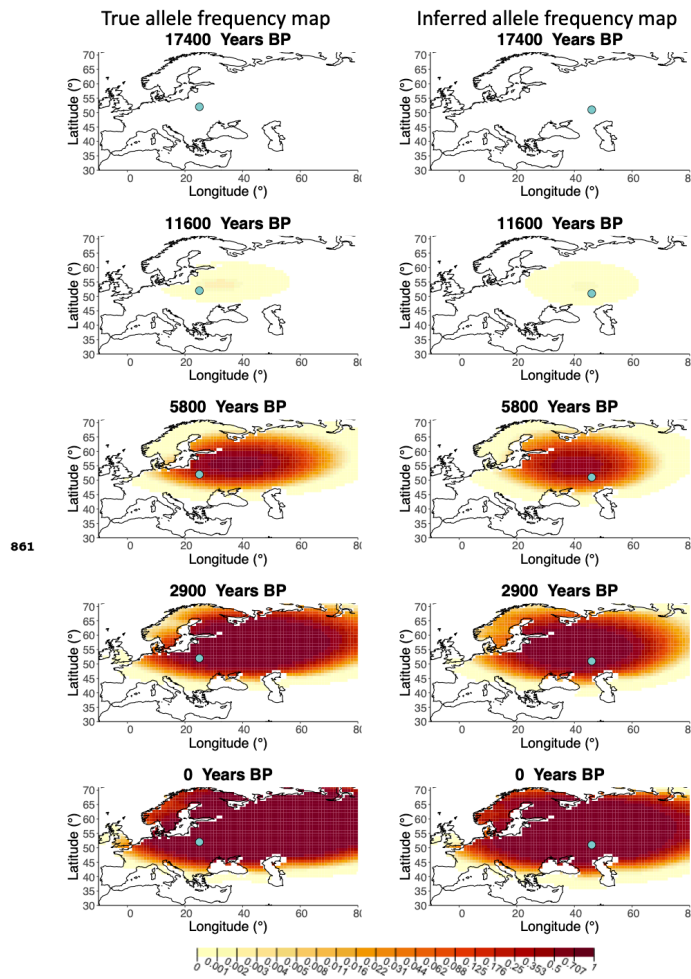




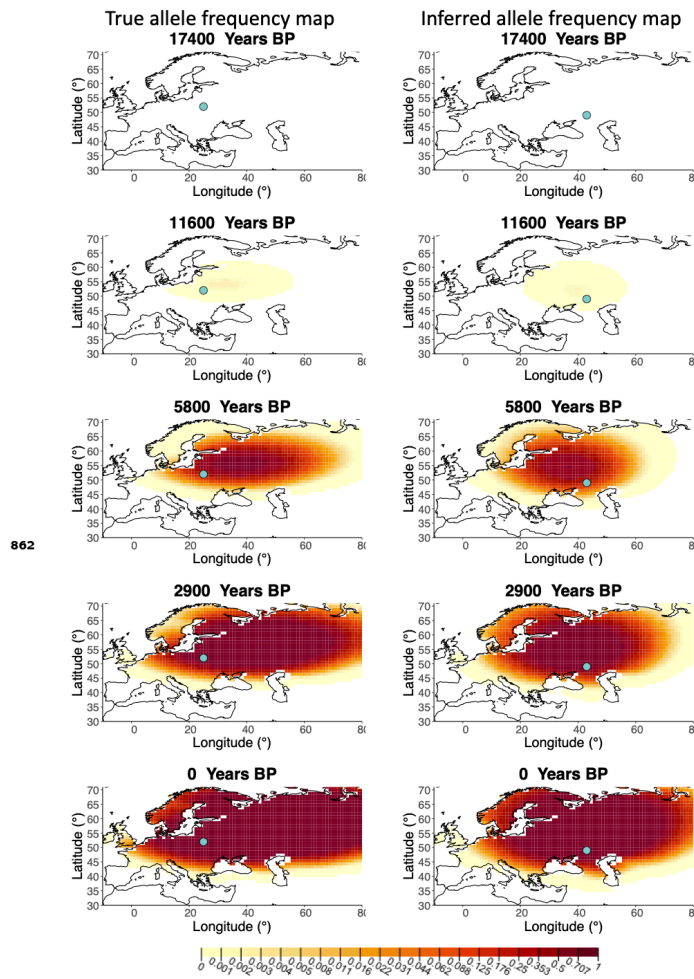
**Figure 3-Figure supplement 1.** Examples of spatial sampling scenarios for each of the three clustering schemes. We chose five locations and increasingly restricted the area where we allowed the individuals to be sampled. a) Map showing homogeneous sampling scheme in which we did not impose any spatial restrictions of individuals sampled. b) Intermediate sampling scheme with the region restricted to 7 degrees in each cardinal direction from each of the chosen locations c) Extreme sampling scheme with the sampling region restricted to 2 degrees in each cardinal direction from the chosen locations.



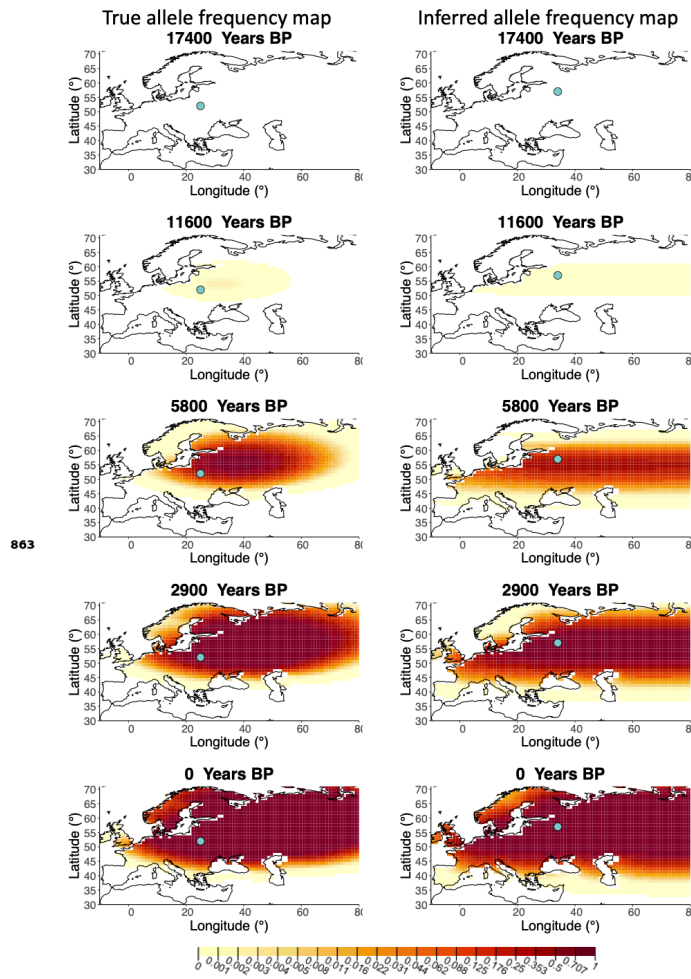
**Figure 3-Figure supplement 2.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “homogeneous 75%/25%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.



**Figure 3-Figure supplement 3.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “homogeneous 50%/50%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.

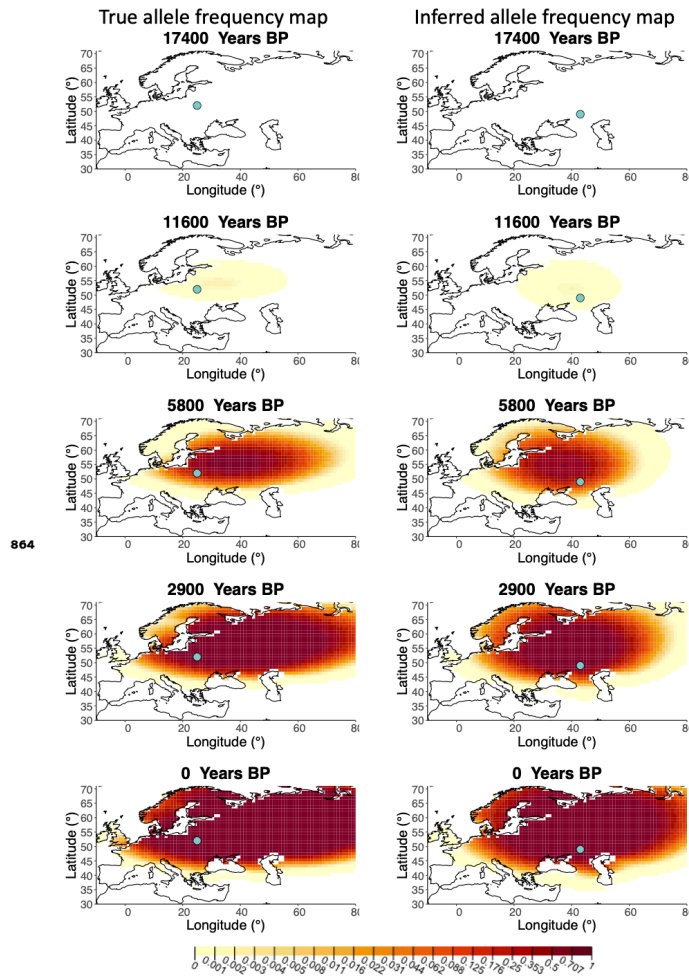


**Figure 3-Figure supplement 4.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “homogeneous 25%/75%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.

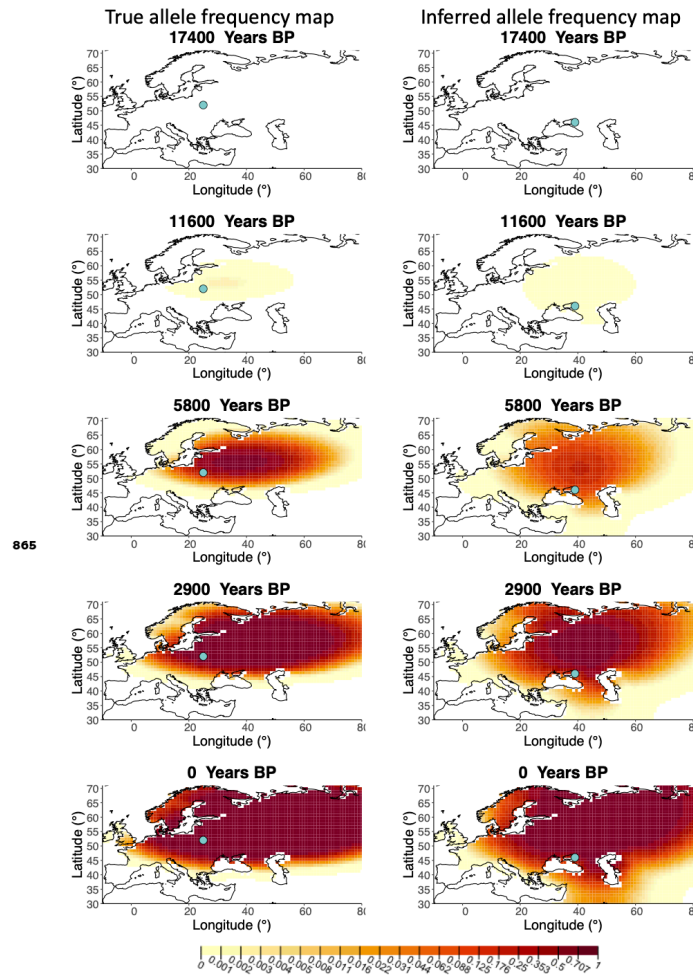


**Figure 3-Figure supplement 5.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “intermediate 50%/50%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.



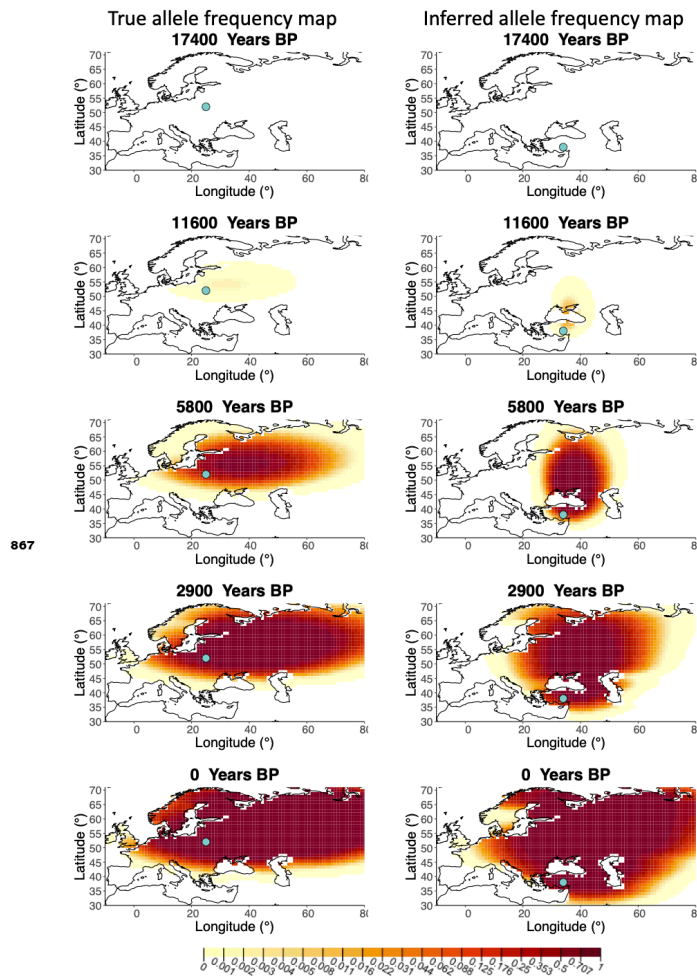


**Figure 3-Figure supplement 6.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “intermediate 25%/75%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.

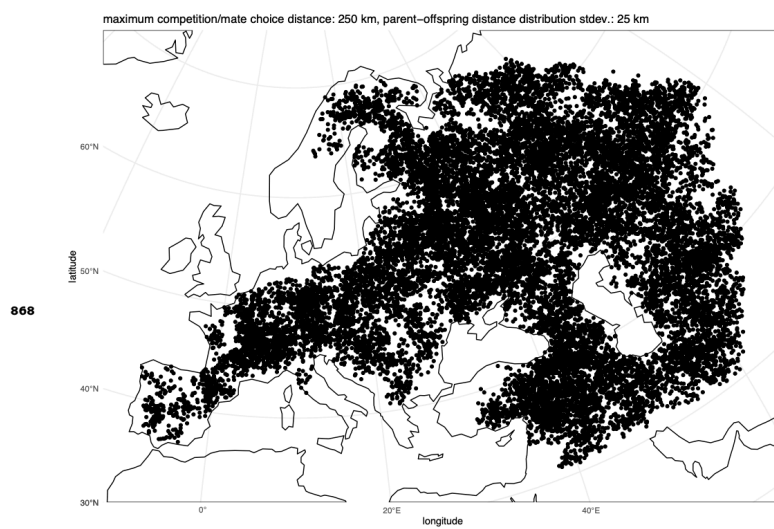


**Figure 3-Figure supplement 7.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “extreme 75%/25%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.

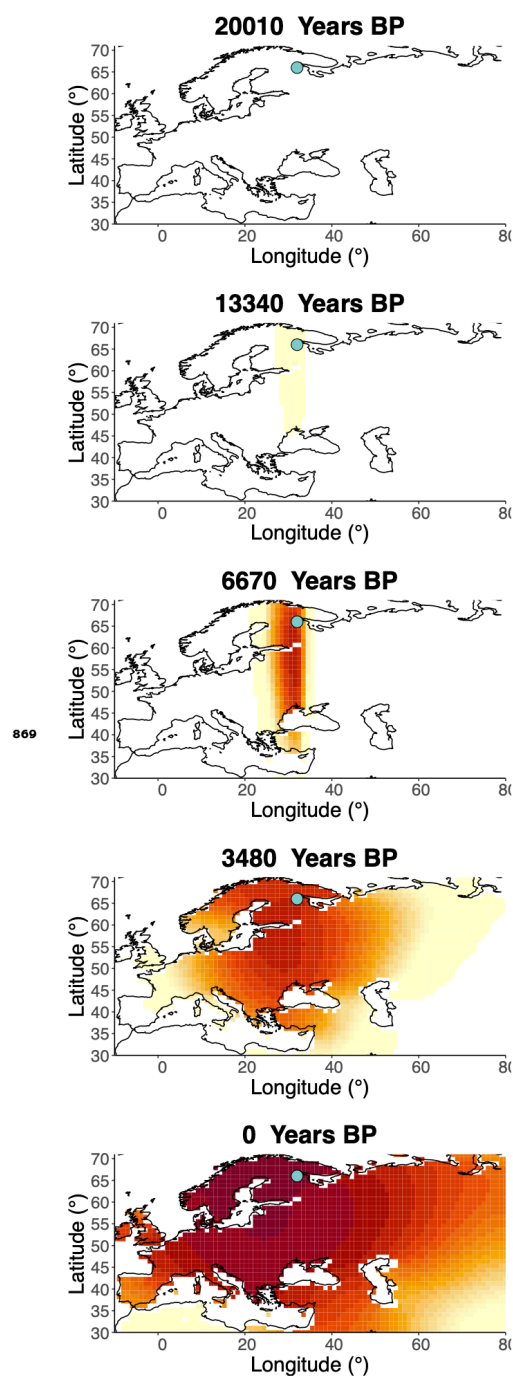




**Figure 3-Figure supplement 9.** Left - Allele frequency map generated using true parameter values. Right - Allele frequency map generated using parameter estimates for “extreme 25%/75%” clustering scheme. Parameter values used to generate the maps are summarised in **Table A3**.

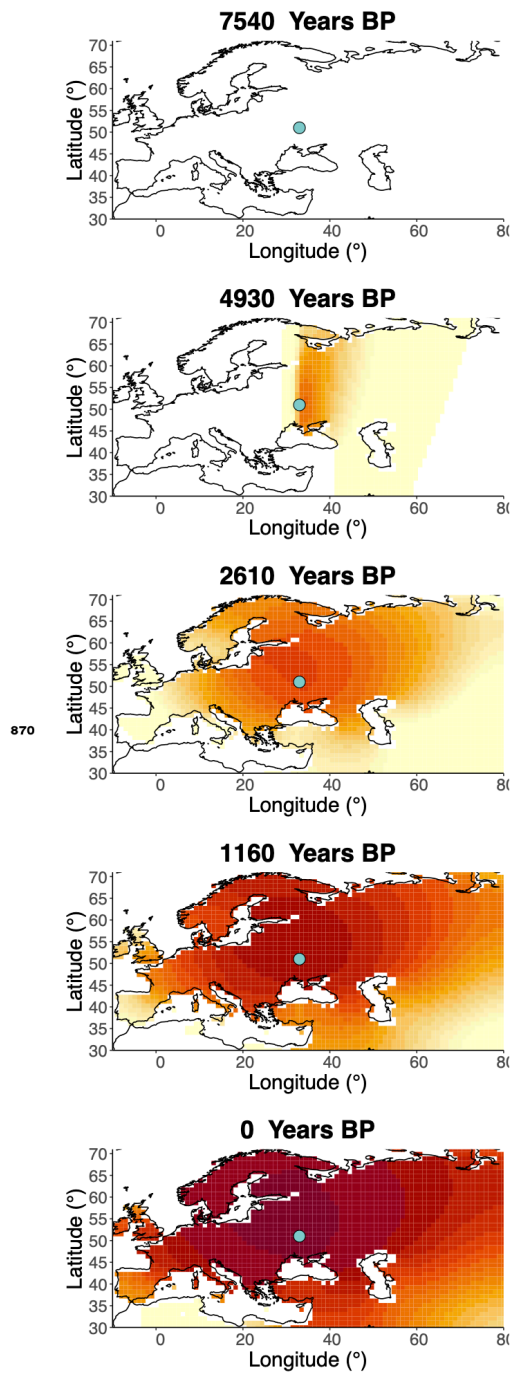


**Figure 4-Figure supplement 1.** Distribution of individuals across the map under neutrality, showing the tendency of individuals to cluster together.

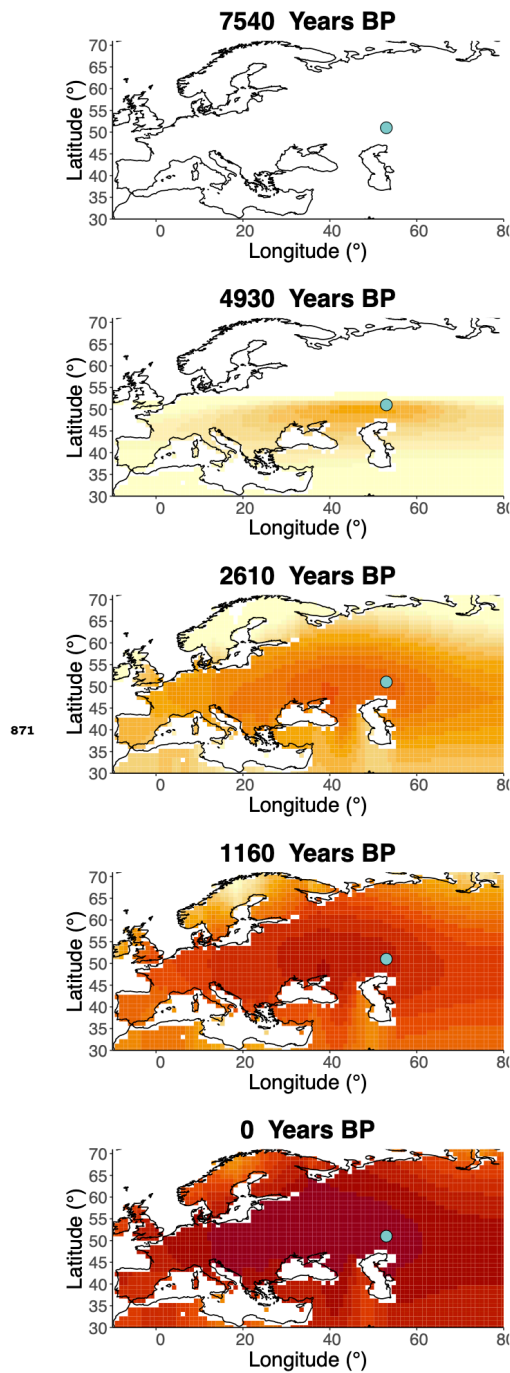


**Figure 6–Figure supplement 1.** Inferred frequency dynamics of rs4988235(T) using the allele age that was inferred in *Albers and McVean (2020)*.

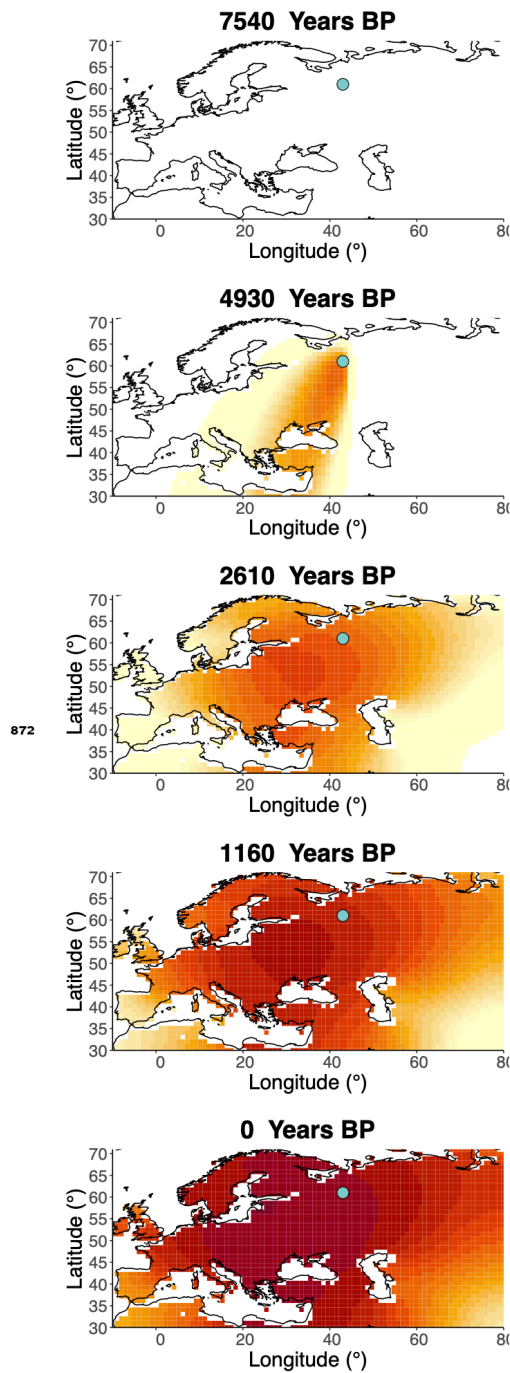




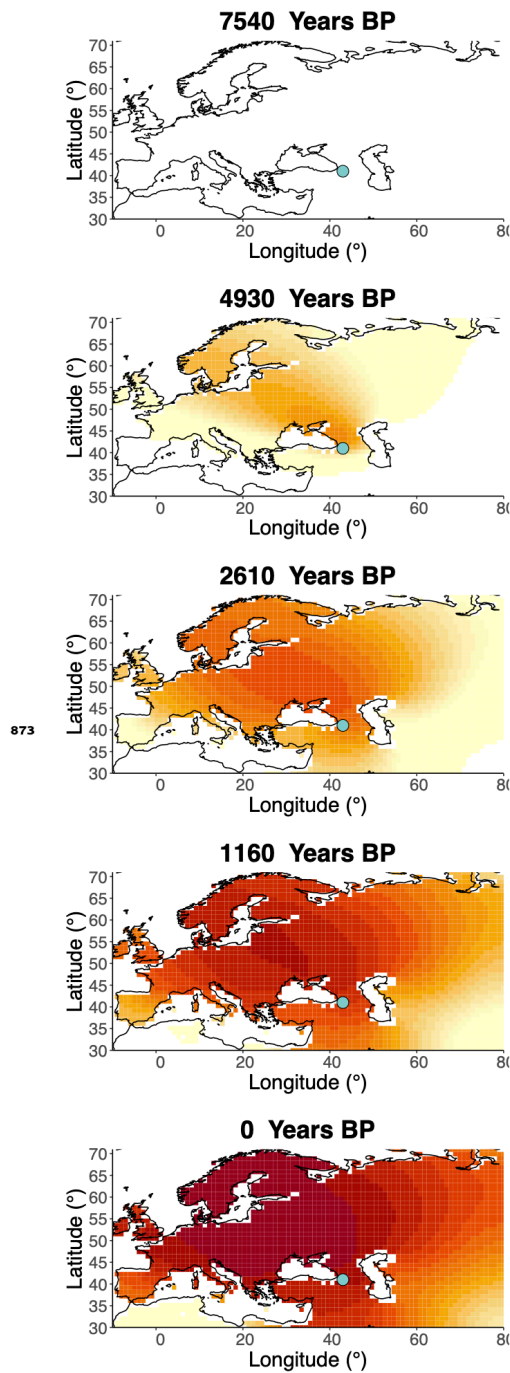
**Figure 6–Figure supplement 2.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees west from the original estimate.



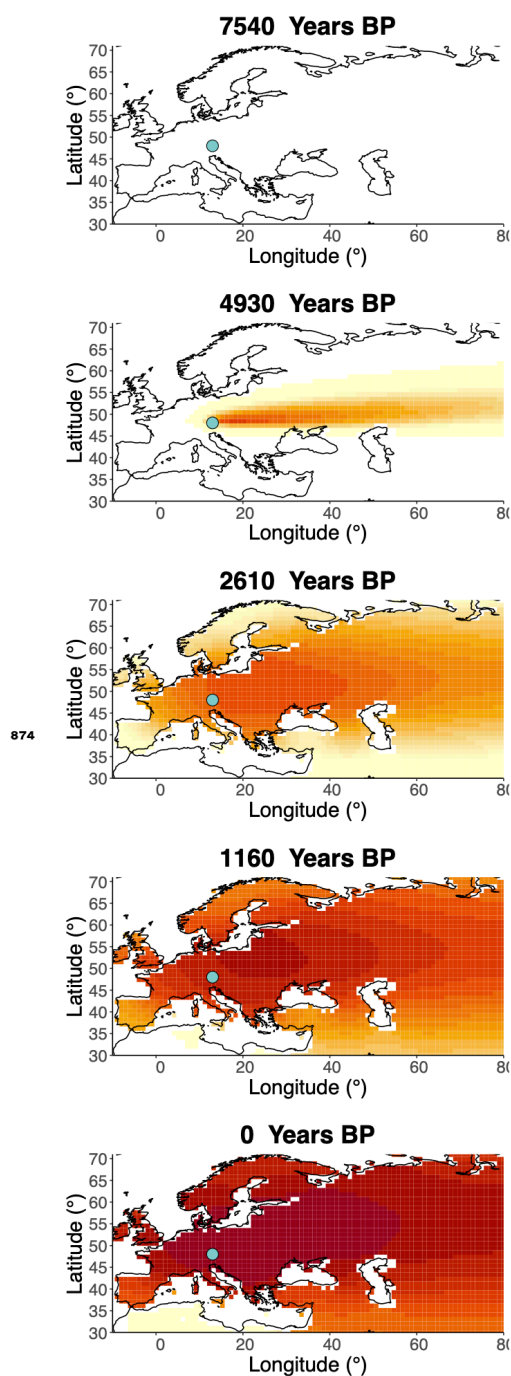
**Figure 6–Figure supplement 3.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees east from the original estimate.



**Figure 6–Figure supplement 4.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees north from the original estimate.

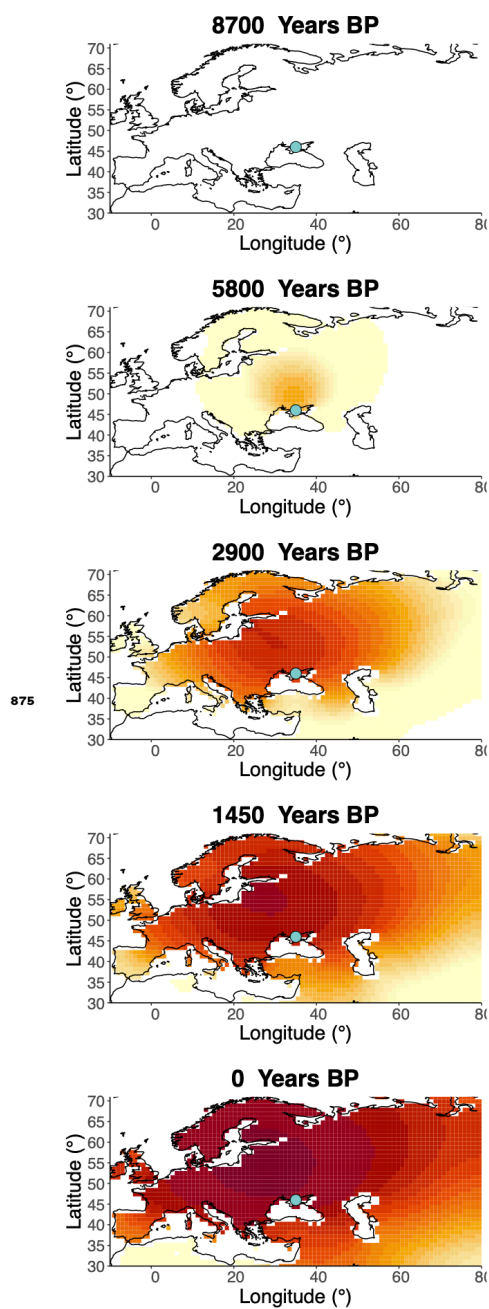


**Figure 6–Figure supplement 5.** Inferred frequency dynamics of rs4988235(T) when the origin of the allele is moved 10 degrees south from the original estimate.

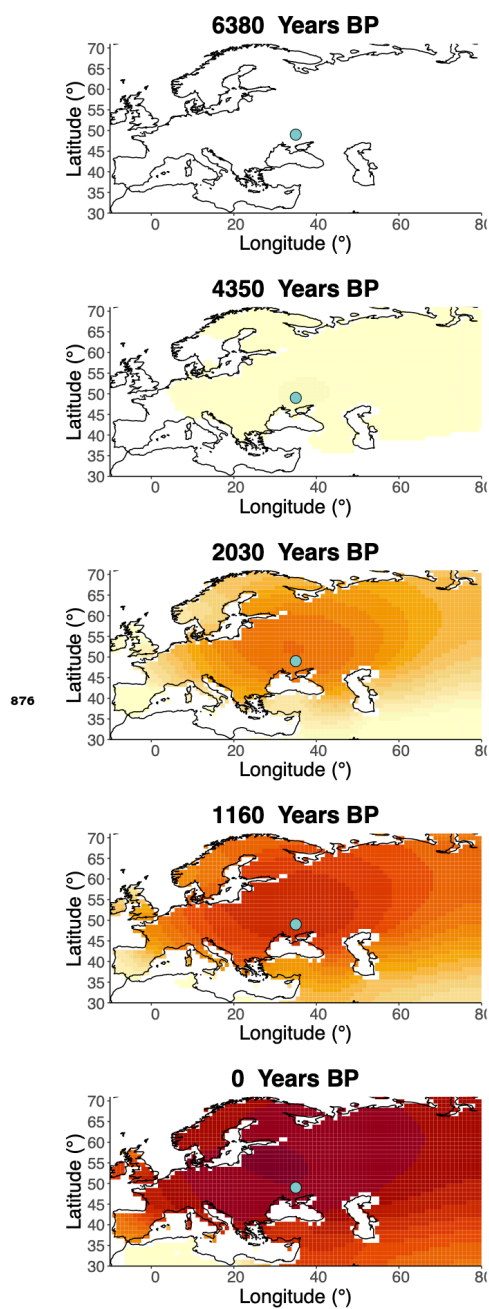


**Figure 6–Figure supplement 6.** Inferred frequency dynamics of rs4988235(T) forcing the geographic origin of the allele to be at the location inferred in *Itan et al. (2009)*.

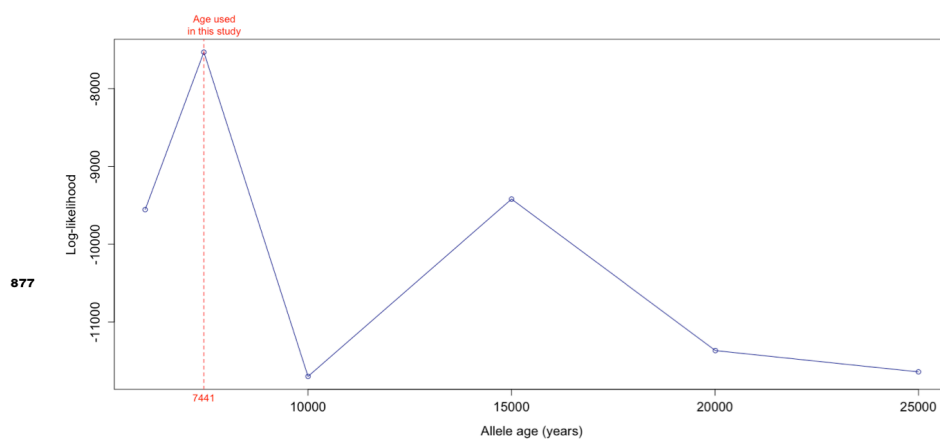




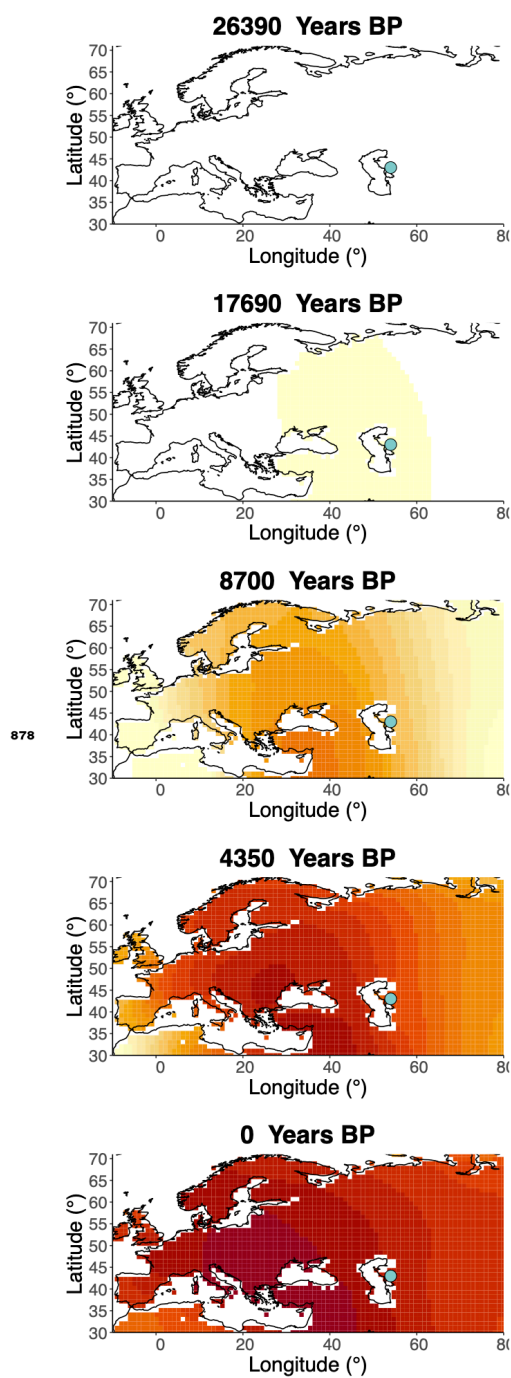
**Figure 6-Figure supplement 7.** Inferred frequency dynamics of rs4988235(T) assuming the allele age to be the lower end of the 95% credible interval for the allele age inferred in *Itan et al. (2009)*.



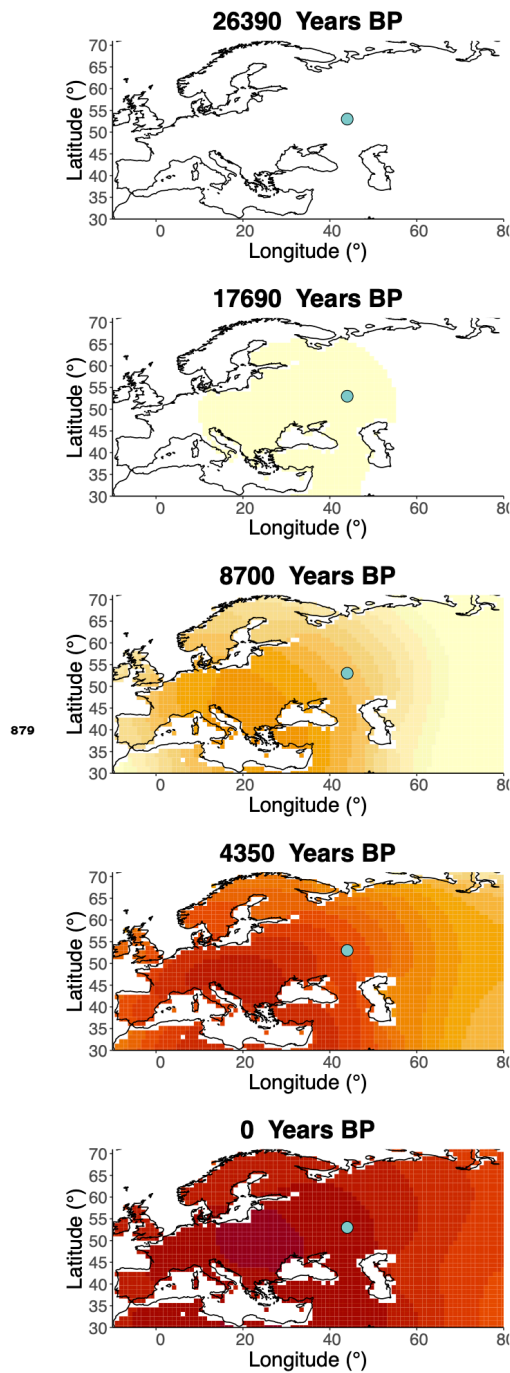
**Figure 6-Figure supplement 8.** Inferred frequency dynamics of rs4988235(T) assuming the allele age to be the higher end of the 95% credible interval for the allele age inferred in *Itan et al. (2009)*.



**Figure 6–Figure supplement 9.** Log-likelihood values for model runs using different ages of the rs4988235(T) allele as input, with the age inferred by *Itan et al. (2009)* we use as fixed input highlighted in red.

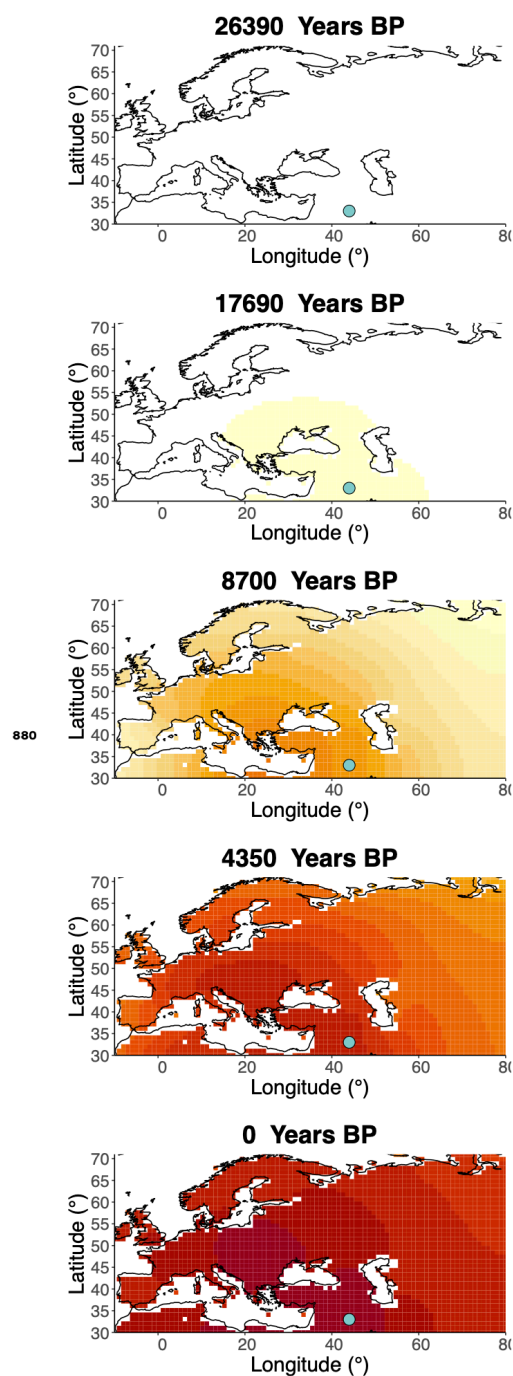


**Figure 8–Figure supplement 1.** Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees east from the original estimate.

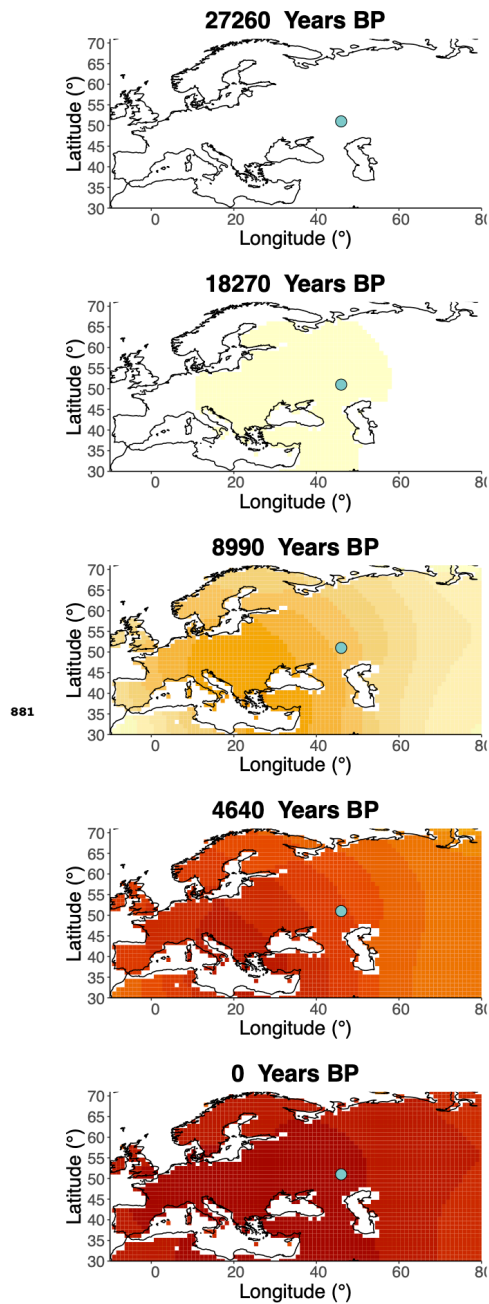


**Figure 8–Figure supplement 2.** Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees north from the original estimate.

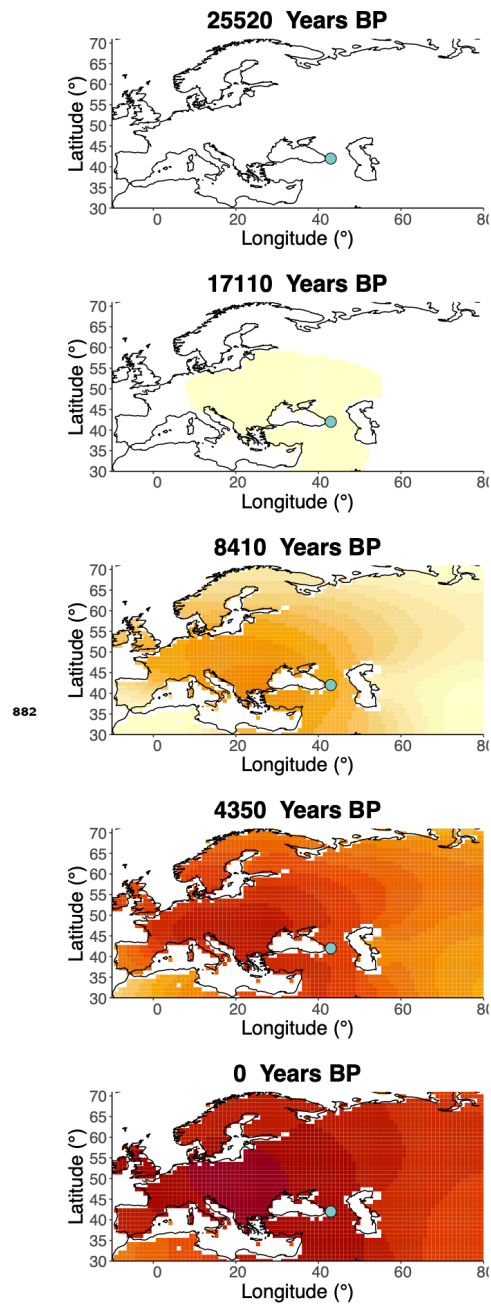




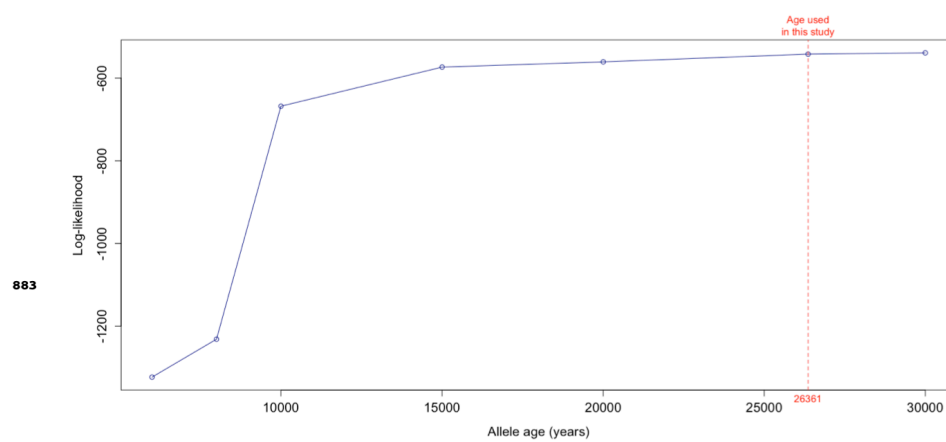
**Figure 8-Figure supplement 3.** Inferred frequency dynamics of rs1042602(A) when the origin of the allele is moved 10 degrees south from the original estimate.



**Figure 8-Figure supplement 4.** 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 and McVean (2020)*.



**Figure 8-Figure supplement 5.** 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 and McVean (2020)*.



**Figure 8-Figure supplement 6.** Log-likelihood values for model runs using different ages of the rs1042602(A) allele as input, with the age inferred by *Albers and McVean (2020)* we use as fixed input highlighted in red.