FINNISH CENTRE OF EXCELLENCE IN TAX SYSTEMS RESEARCH

FIT Working Paper 27

Nurfatima Jandarova and Aldo Rustichini

Selection and the Roy Model in the Neolithic Transition

Tampere University
UNIVERSITY OF HELSINKI
VATT INSTITUTE FOR ECONOMIC RESEARCH
CENTRES OF EXCELLENCE IN RESEARCH

Selection and the Roy Model in the Neolithic Transition

Nurfatima Jandarova* Aldo Rustichini†

October 4, 2024

Abstract

We analyze the evolution of the distribution of genotypes in European populations over the past 14,000 years. In our model, evolution is driven by selection operating after a shift in the productivity of agriculture, induced by the post-Younger Dryas climate change, in a Roy model where individuals self-select into one of two sectors, foraging and farming.

The model extends a standard Wright-Fisher model to include two technologies and sexual reproduction. We test the model in two data sets, ancient and modern DNA, matching the observed distributions of genetic variables (allele frequencies and lineages). We show that a shift in the distribution of allele frequencies in a direction favoring higher cognitive ability, occurred when climate warming changed the relative productivity of agriculture and foraging.

The general implication we draw is that historical transformations (e.g., climate change and technological change) may affect the distribution of genotype and thus economic equilibria and institutions.

JEL: E71, J24, O33

Keywords: technological change, occupational choice, individual characteristics, genetic transmission, population genetics

*Department of Economics, University of Minnesota; nurfatima.jandarova@gmail.com

†Department of Economics, University of Minnesota; aldo.rustichini@gmail.com

0. The authors thank audiences in the conference on *Frontiers in Economic Analysis with Genetic Data*, in Madison, WI, the 2022 New Delhi Winter School of the Econometric Society, audiences in Paris, Bologna, Toronto. In particular they thank for conversations on these and related themes with Nicola Barban, Jonathan P. Beauchamp, Tobias Edwards, Jason M. Fletcher, Oded Galor, Alexandros Giannelis, Andrea Ichino, Matt McGue, Nick Papageorge, Arthur Robson, Giulio Zanella, Ryan Webb.

1 Introduction

The general hypothesis tested in this paper is that, in the development of human societies, events such as technical change and climate shifts may have (in addition to the obvious direct impact on economic activity) also an indirect one via changes in the demographic composition of these societies. These changes occur in the long run, and operate through selection. Thus to understand them it will be necessary to consider a very long historical period, in the theory and in the data.

The specific historical phenomenon we consider is the adoption of agriculture in the period following the Younger Dryas (which ended 11,600 years Before Present (YBP)). This profound transformation of economic activity has been labelled Neolithic Agricultural Transition (NAT). It marked the transformation from foraging to farming, as well as profound cultural and institutional changes: from nomadism to sedentism, together with the creation and diffusion of private property. We focus here on the associated change in the genetic properties of the populations.

Evidence of selection in the more recent past (ranging from 2-3 thousand to 25 thousand years in the past) has been recently studied extensively (see Berg and Coop 2014, Racimo, Berg, and Pickrell 2018, Guo, Yang, and Visscher 2018, Uricchio 2020, Mathieson and Mathieson 2018, Mathieson 2021, Song et al. 2021, Stern et al. 2021, Song et al. 2021, Yair and Coop 2022). The addition offered in our study to this impressive line of research is the link with a standard economic model of activity choice, and the explicit connection with well documented historical transformations, in climate and technology.

The role of genetic factors in the explanation of long-run economic growth is proposed and analyzed theoretically in Galor and Moav 2000, Galor and Moav 2002. Even closer to our topic, Ashraf and Galor 2011 show an effect on population density of a country in 1500 of the number of years since the NAT. At the outset of their extensive survey of the the depth of roots of economic development, (Spolaore and Wacziarg 2013) noted in 2013 that "A growing body of new empirical work focuses on the measurement and estimation of the effects of historical variables on contemporary income by explicitly taking into account the ancestral composition of current population." Their conclusion is that "The evidence suggests that economic development is affected by traits that have been transmitted across generations over the very long run." But the immense volume of information made possible after the Human Genome Project (Gibbs 2020) is systematically absent in this literature.

More recently, research in economics has integrated information on genotypes of individuals in the analysis of wealth accumulation and inter-generational mobility (Barth, Papageorge, and Thom 2020, **rustichini2023polygenic**), both in the theory and in the data. The integration requires a new class of models in which, for example, the transmission of skills across generations takes explicitly into account the underlying process of

genetic transmission (rather than a simplified and incorrect process, such as the $AR(1)$ model). We build on this early investigations to consider the question of the roots of long-run growth, rather than short run (two or few generations) wealth accumulation and transmission.

If we consider more specifically the NAT, recent literature in economics has studied a related but very different issue (Bowles 2011a, Bowles and Choi 2013a, Robson 2010 Rowthorn 2011, Rowthorn and Seabright 2010). The focus of this research was the puzzle of the adoption of agriculture over hunting and gathering. The puzzle consists in some evidence that the health condition of the first farmers was not necessarily better than that of their contemporary foragers; in fact there is some evidence that the health condition may have deteriorated. A second piece of the puzzle was that productivity of cereals among foragers may not have been lower than among farmers (Bowles 2011b). The question then is natural: why the switch from foraging to farming, when it should have been clear that it was not providing better living conditions and higher productivity?

The solution to the puzzle proposed in this literature has emphasized the institutional aspects of this transformation, centering on the fact that the switch from foraging to farming makes necessary the rise and strengthening of private property, which is necessary to provide the appropriate incentive to farmers so they provide the necessary investment. For example, in Rowthorn and Seabright 2010 the adoption of foraging and farming is modelled as a game between two groups (groups are the players in the game) that can simultaneously choose between two actions. Both players would be better off if they both chose foraging rather than farming. However, if the other chooses foraging, a player will be better off by choosing farming; perhaps because farming makes creation of a military apparatus that can be used defensively (to insure the return on the agricultural investment) but also offensively (thus allowing the society with an active military to rob the other group). Facing this new situation the other player will be better off by also switching to farming, therefore insuring the bad equilibrium. The logic is that of the Prisoner's Dilemma game.

Our approach emphasizes, rather that the institutional transformations, the changes in the population structure that we conjecture were associated (and in large part induced) by the change in the nature of economic activity. A substantial part of the change was the change in the distribution of the genotype in the population (that is the change in the frequency of alleles associated with characteristics relevant for the productivity). To put this aspect of the NAT in perspective, it is important to note that a similar transformation of the genotype occurred in populations of plants and animals which was produced by the process of domestication (Darwin 1868, Trut, Oskina, and Kharlamova 2009, Larson and Burger 2013; see also Lord et al. 2020 for a critical evaluation of the speed and nature of the domestication syndrome); so it should not be surprising that a process of similar nature involved human population.

We model the chain of cause and effects following climate change with the introduction of a Roy model of choice of activity by individuals as a fundamental characteristic of the mechanism of selection of individuals and thus as a way of inducing a transformation of the distribution of characteristics in human populations. The starting point is model where two technologies, foraging and farming, coexist in the same geographical space. A populations of individuals who may differ in their abilities and preferences choose the activity which is more convenient. In an initial stage, before the warming and the climate change following the Younger Dryas, climatic conditions were less favorable to farming than to foraging; so a large fraction of the population chose foraging over farming. After a sufficiently long period, an equilibrium was reached where the characteristics of the population were stable, and in particular the distribution of the genotype was at a long run equilibrium. The increase in temperature following the Younger Dryas raised the productivity of agriculture. This increase was not necessarily the same for all individuals, but potentially was more substantial for those who were endowed with some characteristic.

A higher productivity for these people was associated with a larger number of children (that is an increase in fertility) and successively an increase in the mortality rate. This demographic transformation (known as the Neolithic Demographic Transition or NDT) is a well documented change in the demographic structure of the population. In addition to it, a switch to agricultural diet is also associated with higher survival rate of children. Thus, a shift in productivity can be linked to changes in individual fitness levels (probability of producing successful offspring) and, consequently, in the trajectory of the genetic evolution of the entire population.

We begin our analysis with a review of well established facts which are essential components of out general hypothesis, beginning with climate change around 15 thousand years ago in Section 2. We then introduce our model in Section 3 and study its implications in Section 4. In section 5 we outline our estimation strategy. In Section 6 we describe the datasets used in the analysis and present the estimation results in Section 7. Finally, we conclude in Section 8.

2 Climate Change and Evolution of Agriculture

We begin by recalling some well established facts about climate change in the last 20 thousand years. which are going to play a crucial role in our analysis.

2.1 Climate Change in the recent 20K BP

We review here broad patterns of world climate in the 20,000 years before present (BP). This time encompasses two major geological epochs: it captures the end of the Pleistocene which started roughly 2.5M years BP and finished around 11.5K years BP and

the beginning of the Holocene, the ongoing epoch. The late Pleistocene in itself can be broken down into several periods: end of the Last Glacial Maximum ($\sim$ 30K -20K years BP), abrupt cooling during the Oldest Dryas ($\sim$ 15K years BP), return to warming during the Bølling-Allerød interstadial ($\sim$ 15K-13K years BP) that was interrupted with the cooling in the Younger Dryas ($\sim$ 13K - 11.5K years BP)[1]. Consistent with these descriptions, the samples obtained from the ice cores (see figure Figure A-1) suggest a general pattern of a shift from colder to warmer climate in the Northern Hemisphere over the past 20K years BP. The warming slowed down during the Oldest Dryas, gained pace again during the Bølling-Allerød interstadial, reversed in the Younger Dryas period and finally continued to pick up in the Early Holocene. More importantly, these patterns are observed in the Fertile Crescent area as well, shown by reconstructed temperatures in the Eastern Mediterranean sea and Soreq cave in Israel at the bottom of Figure A-1. At the same time, the sea levels were on the rise. Although sea levels may be influenced by other factors such as tectonic shifts or changes in the density of water following temperature changes, it is also correlated with increased precipitation. Furthermore, we find confirmation for these observations in the paleoclimate literature. In particular, Robinson et al. (2006) review thoroughly the literature analysing marine and terrestrial sediments, as well pollen records, in the Eastern Mediterranean and Levant regions.

We are specifically interested in the changes taking place around 11K-10K years BP when the spreading of agriculture started. This period corresponds to some of the most marked shifts in climate over the last 20K years (see plots in Figure A-1). Robinson et al. (2006) characterize the Younger Dryas period as "extremely arid and, most likely, cold" (p. 1536). According to Rossignol-Strick (1995), the Younger Dryas in the Levant region was the driest climate period over the past 24K years. In contrast, the early Holocene appears to have been "the wettest phase over the last 25,000 years across much of the Levant and Eastern Mediterranean" (Robinson et al. 2006, p. 1536). Another important change documented by Shennan (2018) and Feynman and Ruzmaikin (2007) is that climate conditions became more stable with the onset of the Holocene. with fewer significant climate swings over the course of centuries. Along with the changes in average annual conditions, there is also evidence that the early Holocene saw a spike in seasonality within a year. [2] Baldini et al. (2019) reconstruct seasonal temperatures and rainfall amounts in North Iberia and find that the early Holocene saw a marked rise in rainfall seasonality compared to Younger Dryas, with hotter, drier summers, and milder, wetter winters .

1. The timelines according to Stuiver, Grootes, and Braziunas (1995), rounded.

2. Matranga (2019) argues that seasonality rose as a result of a change in planetary tilt relative to the sun.

2.2 Evolution of Agriculture

The earliest direct evidence of agriculture dates as far back as 20K years BP (Piperno et al. 2004). The authors analyse the archaeological findings from the Upper Paleolithic site called Ohalo II in Israel. It consisted of a group of huts at the shore of the Sea of Galilee. One of the well-preserved huts contained a large grinding stone. Analysing the stone and the hut, the authors were able to conclude that it was used to process wild grains, "including barley and possibly wheat" (p. 670), with the resulting flour being mixed with water and baked before consumption.

Nevertheless, agriculture only started spreading around 10K years BP. Existing research suggests that changing climate conditions played a crucial role in initiating this process. Feynman and Ruzmaikin (2007) argue that not only warmer and wetter climate, but climate stability during Holocene, has made agriculture a viable occupation. Predictable climate conditions from year to year, from century to century are necessary to build physical and social infrastructure involved in growing grains, growing domesticated animals, as well storing and sharing the produce. Feynman and Ruzmaikin (2007) argue that these processes require at least 2K years of a stable climate. Furthermore, changes in seasonality are thought to have contributed to the rise of agriculture as grains were better suited for storage and helped smooth consumption throughout the year (Matranga 2019; McCorriston and Hole 1991).

All of these factors can help explain the initial adoption of farming driven by the increased marginal productivity of agriculture (Bowles 2011a). The continued spread of agriculture (Figure A-2) points to evolutionary advantages (Shennan 2018; Bowles 2011a). The literature suggests that adoption of agriculture increases fertility, and possibly decreases childhood mortality (Kramer and Greaves 2007; cf. Helle et al. 2014), thanks to calorie-dense diet and settled lifestyle (Helle et al. 2014; Kramer and Greaves 2007; Bocquet-Appel 2008; Winterhalder and Goland 1993).

The spread of farming was accompanied by changes in socio-economic infrastructure. The most notable among these changes is the emergence of private property rights. Bowles and Choi (2013b) and later Gallagher, Shennan, and Thomas (2015) argue that private property rights and farming must have evolved jointly since each was necessary for the development of the other. The growing importance of property rights can also be seen in archaeological findings. For example, the literature documents transition from communal storage facilities (Kuijt and Finlayson 2009) to storage within houses (Kuijt 2008; Bogaard et al. 2009) between 11K and 10K years BP. Similarly, domesticated animals could be viewed as a form of property important for agriculture, and the existing evidence suggests that their rearing was closely controlled by farmers (Scheu et al. 2015). Apart from property rights, the spread of farming is also associated with the expansion of a "way of village life" (Özdoğan 2011).

This review suggests that the adoption of farming provided an evolutionary advantage to human society and new farming-friendly infrastructures were spreading together with agriculture. In this paper, we ask whether characteristics that affect individual productivity in agriculture would, as a result, be subject to positive selective pressure.

2.3 Selection with Two Technologies

We combine a model of selection with one of choice of occupation (here, either farming (AG) or hunter-gathering (HG)). It is important to emphasize from the start that we do not assume that any special advantage of a higher level of a phenotype in one of the two activities. In particular we do not assume that higher level of cognitive ability are more advantageous in farming.

Individuals have different level of skills, determined in part by their genetic endowment. Selection makes an individual with certain characteristics more likely to produce surviving offspring (Kingsolver and Pfennig 2007). A fitness function describes the relationship between individual characteristics and outcomes (for example, quantity of food produced), and thus ultimately on the expected number of surviving offsprings. The choice of occupation is described by a simple Roy model. The product of an individual depends on the chosen activity and the skill of the individual. Individuals choose the occupation that provides the highest quantity of output net of costs.

This model implies the existence of a fitness associated with a given level of skills. Thus, an occupation choice determined by relative fitness is equivalent to the choice driven by relative productivity. The evidence presented earlier shows that some people did engage in farming as far back as 20K years BP, but in limited number. Around 10K years BP, climate change and accompanying shifts in social infrastructure, raised individual productivity in AG. This change affected the relative returns fro the two occupations, possibly in a different way for different skills, and more people find AG attractive.

3 Model

We begin with a detailed description of the genotype and its structure. These details are essential to be able later to test the hypotheses with the available data. A detailed model as the one produced here is necessary to address confounding factors. For example in a possible alternative explanation of the relationship between changes in the frequency of alleles and the effect on phenotype is proposed in Charlesworth 2013, Zeng and Charlesworth 2009, focusing on the higher relative frequency of negative mutations. To exclude these explanations a model, substantially more detailed than the Wright-Fisher model, is necessary.

3.1 Setup

We begin with basic notation in genetic analysis (Crow and Kimura 1970, Nagylaki 2013). We call K the number of true causal loci, all taken to be bi-allelic. We denote 1 and 0 the two alleles at the k^{th} locus. It is convenient to adopt the convention for a finite sets that the letter for the cardinality in bold indicates the set, so $\mathbf{K} \equiv \{k : 1 \leq k \leq K\}$. In our analysis we want to keep track of the haplotype structure. So we first introduce a set $\mathbf{C}$ of chromosomes, and use $c : 1 \leq c \leq C$ as the index for the chromosome. The set of loci $\mathbf{K}$ has a natural partition into the subsets of loci $\mathbf{K}_c$ that belong to a same chromosome c, that is $\mathbf{K} = \cup_{c=1}^{C} \mathbf{K}_c$, with pairwise disjoint intersection.

We call an haplotype an element in the product space $\{0,1\}^K$. An individual in the population is characterized by a haplotype pair, that is an element in $\{0,1\}^K \times \{0,1\}^K$. The set of haplotype pairs is $\mathbf{H} \equiv \{0,1\}^K \times \{0,1\}^K$; and an haplotype pair can be written $h = (l, r)$. We write l_c the c chromosome component of the "left" haplotype. [3] A description of the genotype of an individual which ignores the haplotype structure is convenient, so we introduce it by letting:

$$\mathbf{G} \equiv \{0,1,2\}^K, \tag{1}$$

so $\mathbf{G}$ has cardinality $G = 3^K$. To illustrate, here $g(k) = 1$ indicates that the individual with g genotype is heterozygote at at the k^{th} locus, that is an individual with the pair of alleles $(1,0)$ at locus k. For every haplotype pair $h = (l, r)$ there is an associated genotype $S(h) \equiv l + r$. For any $h \in \mathbf{H}$ we denote $m(h)$ the number of individuals in the population with that haplotype pairs, $m(h,t)$ to denote that number at time t; similarly for $n(g)$ (respectively $n(g,t)$) for $g \in \mathbf{G}$.

The distribution of genotypes in the population may be described by $\pi \in \Delta(\mathbf{H})$ if we want to keep track of the distribution on haplotypes, or by $\mu \in \Delta(\mathbf{G})$ for a less detailed description. For a given finite value of the population N we denote $\Delta_N(\mathbf{H})$ the set of probability distributions on $\mathbf{H}$ that can be induced by a population of size N; similarly for $\Delta_N(\mathbf{G})$. [4] One can derive from these two distributions a basic tool which is the vector p of allele frequencies $p \in [0,1]^K$. Time indicated by t is discrete and measured in generations, that we can approximately take to be in 25 years. When necessary we

3. Of course, two haplotype pairs h and h' are equivalent, written $h \sim h'$, if at each chromosome either they are identical ($h_c = h'_c$) or one can obtained from the other by switching left and right ($l_c = r'_c$ and $r_c = l'_c$). In a more precise sense, the set of haplotypes is really the set of equivalence classes of $\{0,1\}^K \times \{0,1\}^K$ under this equivalence relation. However, for our purposes of computing the allele frequency at different times, the two formulations are equivalent, and the formulation of $\mathbf{H}$ offered here is simpler, so it is the one we adopt.

4. That is:

$$\Delta_N(\mathbf{H}) \equiv \left\{ \pi \in \Delta(\mathbf{H}) : \exists (n(h) : h \in \mathbf{H}), \sum_h n(h) = N, \pi(h) = \frac{n(h)}{N} \right\}$$

will make the dependence of the variables on time; for example $p(t)$ indicates the allele frequency at t. If we let the $K \times G$ dimensional matrix A, defined by

$$A_k(g) = 0 \text{ if } g(k) = 0, \frac{1}{2} \text{ if } g(k) = 1, 1 \text{ if } g(k) = 2, \tag{2}$$

then the allele frequency given a population $n = (n(g) : g \in G)$ is $p = \frac{1}{N}An$. We consider populations consisting of a fixed even number N of individuals, with $N = 2M$. An individual in the population will be indexed by i.

3.2 Process on allele Frequencies

In the following sections from 3.2.1 to 3.2.3 we describe the components of the process on the allele frequencies. These components will be used to define the process on frequencies haplotypes (section 3.3) and allele frequencies (section 4.1).

3.2.1 Mutation

The outcome of mutation is described as follows. For $x \in [0, 1]$ define:

$$Q(x, N, \alpha) = x + \frac{n - m}{2N} \tag{3}$$

where n and m are random variables, and $n \simeq \mathbf{bn}((1 - x)2N, \alpha), m \simeq \mathbf{bn}(x2N, \alpha)$, where $\mathbf{bn}(M, \alpha)$ denotes the binomial with M outcomes and probability of a success equal to α. When the population size N and the mutation rate α are fixed we write simply $Q(x)$. For any k, if $p(k)$ is the frequency of the 1 allele, then $Q(p(k))$ is the random variable describing the frequency after mutation. We call $q(k)$ a realization of this random variable.

3.2.2 Recombination

We assume that recombination at meiosis can occur between locus k and $k + 1$ independently. In particular there is a vector $\mathbf{r} \in [0, 1]^K$ where the value $\mathbf{r}(k)$ describes the probability that an odd number of crossovers occurs between the k and $k + 1$ locus, so that after meiosis the alleles in different parental haplotypes are in the same gamete. The probability of a crossover is modeled as usual: it follows a non-homogeneous Poisson process (generalizing the original formulation Haldane 1919) with rate $\lambda(k)$ at position k on the chromosome, starting from the initial position $k = 1$. This allows us to model the existence of recombination hotspots and coldspots (Lichten and Goldman 1995, Myers et al. 2005). The number of crossovers between position i, called $NC(l)$ and $j > i$ is assumed to have the distribution $NC(l) - NC(j) \sim \text{Poisson}\left(\sum_{i=j}^{l} \lambda(i)\right)$.

If we define, for any position l, $\Lambda(l) \equiv \sum_{i=1}^{k} \lambda(i)$, then the number of crossovers between any two positions satisfies: $NC(l) - NC(j) \sim \text{Poisson}\left(\Lambda(l)) - \Lambda(j)\right)$ and therefore

$$P(N(l) - N(j) \text{ is odd}) = \frac{1 - e^{-2(\Lambda(l) - \Lambda(j))}}{2} \tag{4}$$

which is the formula we use to compute the vector r.

3.2.3 Selection

Selection operates on phenotypes, so we describe them first. There is a finite set L of phenotypes, usually those of interest to our analysis, such as cognitive skills, attitude to risk or physical strength; a phenotype is indicated by the the vector $z \equiv (z_l : 1 \leq l \leq L)$. The map $z : G \to Z^L$ where Z is the real line, describes the vector of phenotypes associated with a genotype. We assume it has the linear form:

$$\forall l : z_l(g) = \sum_k \beta_l(k) g(k). \tag{5}$$

where the matrix $\beta \equiv (\beta_l(k) : 1 \leq L, 1 \leq K)$ is the matrix of genetic contributions. These coefficients are estimated by the $GWAS$ coefficients. A similar expression, denoted $z(h)$ gives the phenotype as function of the haplotype.

We next describe technologies available in an environment. There is a finite set of technologies that are available at every point in time, with generic element τ. In our analysis below we will focus on the set of two technologies $\{HG, AG\}$ (for Hunter-gathering and Agriculture respectively. We will assume that an individual with genotype g and thus phenotype $z(g)$ chooses the technology that gives the highest yield in some consumption good. There is an underlying causal sequence, that we do not need to model, from technology chosen to yield, and then from yield to fitness; we will consider the reduced form in which a fitness function associates, for a given technology, fitness to phenotype. This function is described by:

$$f(z, \tau) \equiv R_\tau \exp(-(z - \hat{z}_\tau)^T \Omega_\tau (z - \hat{z}_\tau)). \tag{6}$$

In equation (6), $\hat{z}_\tau$ is the ideal phenotype, and Ω_τ measures potential complementarities among different phenotypes. The R_τ is a scaling factor measuring the relative fitness of the technologies at the ideal phenotype for each technology. We assume that every individual (that is, every phenotype or equivalently every genotype) choose the fitness maximizing technology, so assuming this choice the fitness associated to a phenotype z is

$$\hat{f}(z) \equiv \max_\tau f(z, \tau). \tag{7}$$

The essential feature of the assumption made on the fitness function is that the fitness is initially (that is, for small values) increasing in the phenotype for both technologies, and that the maximum value is different for the two technologies and may be, depending on environmental conditions, higher for one (say HG) and in a different environment larger for the other.

Fitness Function

The assumption that fitness may decline for values higher than the ideal point $\hat{z}$ is usually made in the evolutionary literature to exclude an indefinite increase in the value of the phenotype. This feature is not essential for our conclusions, and some reader may object to the assumption, so a discussion of possible alternative forms of the technology is appropriate.

Assumptions on the technology of hunter-gatherers may appear hard to formulate, but some recent research from anthropologists may throw some light on it. In particular, the following conclusions may be worth considering

1. Contrary to what seems natural in an economic analysis approach to the problem, sharing among hunter-gatherers may be due more to the difficulty in excluding others from consumption than to insurance reasons (I share today after a good hunt because tomorrow I might have to rely on your catch) (see e.g.Hawkes et al. 1993)

2. Distribution of income and wealth [5] among hunter-gatherers was not egalitarian, hence it is natural to assume that some individual characteristic is likely to influence the output (Testart et al. 1988, Smith et al. 2010, Smith 2004, Smith and Codding 2021).

In view of this evidence, it may be reasonable to assume that character traits such as intelligence, *everything else being equal* increase the average product. Thus it is reasonable to consider the hypothesis that the fitness increases as some measure of cognitive skill increases. On the other hand, it may be harder to see why fitness should decline beyond some threshold, as the standard Gaussian fitness function assumes.

As we mentioned, the assumption of a decline is only made to be consistent with an existing literature. An alternative form of fitness is the logistic function:

$$f(z,\tau) \equiv R_\tau \frac{\exp((z-\hat{z}_\tau)^T \Omega_\tau)}{1+\exp((z-\hat{z}_\tau)^T \Omega_\tau)}. \tag{8}$$

5. Wealth in these societies is "defined broadly as factors that contribute to individual or household well-being, ranging from embodied forms, such as weight and hunting success, to material forms, such as household goods, as well as relational wealth in exchange partners". (Smith et al. 2010).

which is monotonic in the phenotype, but with decreasing returns after the flex point. This formulation is closer to the linear form we have used in the simple approximation:

$$f(z,\tau) \equiv R_\tau z^T \Omega_\tau \tag{9}$$

Of course a formulation as in equation (8), and the associated (7) predicts that the phenotype will tend to grow with no limit, but the rate will decline in later periods.

3.3 The Process on Haplotype Pairs

We will adjust the level of detail in the notation as needed. For example, $h(k;i,t)$ will denote the k^{th} value of the genotype of the i^{th} individual at time t, and we will write $h(i,t)$ the vector $(h(k;i,t) : 1 \leq k \leq K)$, and so on. Similarly for the other variables l, r and g. The most detailed description of the genotypic structure of the population is provided by the vector h, which keeps the distinction of identity by state and by descent. Thus the process below models the evolution of the distribution of the vector $h(t)$ in the population, for each time t.

We describe now the steps in the process taking from $h(t)$ to a distribution on the variable $h(t+1)$ in the next period. The transition probability of an h-individual depends (though the mating process and the fitness function) on the entire distribution of haplotypes in the population.

1. (**Initial Condition**) At time t there is a vector $h(t)$ of haplotypes, one for each of the N individuals. Of course the only relevant information is the number of individual with a given haplotype pair.

2. (**Mutation**) Every allele in the genome of every individual mutates randomly with probability α. Mutations are independent across loci and individuals. If a mutation occurs, every allele changes state from ancestral to derived or from derived to ancestral. [6]

3. (**Recombination**) Recombination events are drawn randomly with probability given by the **r** vector introduced in section 3.2.2, between each locus.

4. (**Sexual mating**) The N individuals are assigned to two disjoint subsets of men and women, and matched in pairs. Matching is random and independent of the genotypes of the individuals. Each individual in the pair produces a gamete, choosing one of the two with equal probability.

6. That is, we do not impose any infinite allele assumption and allow mutations to occur twice in the same location.

5. (**Reproduction**) Each individual in the pair produces, for each child, a gamete, that is, in our notation, the individual with $h_c(t) = (l_c(t), r_c(t))$ contributes either $l_c(t)$ or $r_c(t)$ with equal probabilities independently to each child, for each c chromosome. Random matching is repeated independently C times. Overall, a number of children $C \geq 2$ for each pair of new parents is generated.

6. (**Selection in Wright-Fisher**) Only a total number N of children reaches reproductive age, as in Wright-Fisher, so the population size stays constant. Every child reaches reproductive age with a probability that depends on the *relative* fitness of the child (compared to that of the other children), according to the fitness function. Specifically, the next generation is drawn from the multinomial distribution over the current set of haplotypes with probabilities given by the existing frequency on haplotypes of the set of children, as modified by the fitness function.

7. (**Next stage**) The set of children reaching the reproductive age gives a new vector of haplotypes $h(t+1)$ and the process starts again until we reach modern-day generation.

3.3.1 The Invariant Measure

We note that the process described here is a Markov chain characterized by some function T:

$$T : \Delta_N(\mathbf{H}) \to \Delta(\Delta_N(\mathbf{H}), \mathcal{B}(\Delta_N(\mathbf{H})), \tag{10}$$

where $\mathcal{B}(\Delta_N(\mathbf{H})$ are the Borel sets. We study in the following analysis the invariant measure of the process on the state space, which is $\Delta_N(\mathbf{H})$:

Theorem 3.1. *There exists an invariant measure on the set of haplotype frequencies* $\Delta_N(\mathbf{H})$ *induced by the process described in section 3.3.*

Note that to each step there is an associated distribution of allele frequency in the relevant population (of haplotype pairs in the mutation, recombination steps, of gametes after matching and so on). Some of the steps in the process change the mean allele frequency in the population, and some do not. For example, the mutation step and the selection step may change it, while the recombination and matching steps do not. Thus, when there is no selection, the only change in frequency is produced by the mutation.

Proposition 3.2. *If there is no selection, at the invariant measure the correlation between two loci separated by a positive recombination value is zero.*

Since at the invariant measure the mean phenotype is close to the optimal one, the derivative of the fitness function that enters into the dynamic of the mean allele frequency is close to zero and so the selective pressure is small, thus we can expect the correlation

induced by linkage disequilibrium to be small. Thus approximating the main model discussed in this section with a model that assumes the loci to uncorrelated is reasonable. We develop this in the next section and we compare the paths in the two models later.

4 Allele Frequencies under Independence

In this section we analyse the process on genotypes as described in section 3.3, under the additional independence assumption that for all k, $\mathbf{r}(k) = 1/2$. In this case the analysis can be simplified by considering directly the allele frequency at every locus.

4.1 The Process on Frequencies

The process we describe takes the state variable of interest to be the allele frequency, so the starting point is some $p \in [0,1]^K$. In fact, since the population is finite, for every k, $p(k)$ is in the subset of $[0,1]_N$ of the unit interval made of multiples of $\frac{1}{2N}$. This is the initial condition. We call $p(1)$ the vector of allele frequencies in the next period and $\Delta p = p(1) - p$. The next lemma describes the mean change in frequency induced by the process described in section 3.3 on allele frequencies, for a specific realization of the mutation vector.

The object of study is as in the more general case considered earlier the invariant measure, this time on allele frequencies because they contain all the relevant information:

Theorem 4.1. *There exists an invariant measure on allele frequencies.*

We now proceed to give a characterization of the process:

Lemma 4.2. *Under the independence assumption, for any realization q of the mutation vector and population size N, the process on allele frequencies is such that:*

$$E(\Delta p|q,N)(k) = \tag{11}$$

$$q(k)(1-q(k))\left((1-q(k))(F_1^k - F_0^k) + q(k)(F_2^k - F_1^k)\right) + q(k) - p(k)$$

where F_i^k is defined in equation (18).

4.2 The Role of Migration

An important assumption of the model is random mating. This assumption may conflict with evidence of genetic separation between populations; for example, Bramanti et al. 2009 find using mitochondrial DNA (mtDNA) in a sample of european hunter-gatherers and early farmers, that farmers were not descendant from local hunter-gatherers but from

migrants reaching central Europe at the beginning of the Neolithic. The role of the random mating assumption is strictly related to that of migration in the selection process. A group of individuals which benefits from an increase in number, due to characteristics favorable to fitness in a specific environment, may decide to stay in place or instead move to a different location within, say, a continent or a larger region. From the point of view of the process we are considering, namely that of the shift in the distribution of genotype in this region, whether the larger number migrates or not within that region is irrelevant. The crucial question from the point of view of our model, and in particular the random mating assumption, is whether the migrants interbreed with the local population or not. If genetic discontinuity (see Crosby et al. 1970 for a study of theoretical conditions for the selection of genes preventing interbreeding among subspecies) or separation between the two subspecies persisted, then the driving force of the operation of selection was migration rather than selection operating on a homogeneous population.

The evidence of genetic discontinuity within human populations in the period we are considering is not clear. For example Bollongino et al. 2013 show evidence in Mesolithic populations (with individuals identified with mtDNA) of the parallel existence of two groups, with different diets: one which maintained a foraging and freshwater fishing diet, the other which instead had adopted an agriculturalist diet, and presumably a farming lifestyle. This evidence is not necessarily contradicting the earlier findings Bramanti et al. 2009, but shows that those findings may not be general for all the regions. [7] Overall, the existing literature provides evidence that the diffusion of farming techniques is the result of a combination of movement of people, ideas, and of differential selection (see González-Fortes et al. 2017). Studies of selection for several phenotypes have already noted the difficulty of separating evolution over time or replacement (for the example of height, see Cox et al. 2019a, page 21488; Cox et al. 2022, page 169; for skin hair and eye pigmentation, Wilde et al. 2014, page 4833).

5 Estimation

The process described in section 3 takes a vector of parameters and produces a probability distribution on observable variables of interest. Specifically:

1. The vector of parameters is:

$$\lambda \equiv (N, \beta, (\hat{z}_\tau, \Omega_\tau, R_\tau)_\tau) \tag{12}$$

   of population size, phenotype weights and parameters describing the fitness function;

7. Verdu et al. 2013 study the mating patterns between Pygmy and non-Pygmy (farmers) populations, Aimé et al. 2013.

2. The observed variables is the frequency of genotypes, that is a distribution on **H**, and more in detail the history over time of the distribution and on the the tree of lineages.

We will see in our examples that the initial condition on the distribution of haplotypes is essential, because even the steady state depends on it. We focus on the initial condition in which all alleles are ancestral.

5.1 Fixation probability

Consider as an illustration the case in which a phenotype is influenced by few derived alleles with relatively large coefficients. We may refer to the ancestral (that is, the allele in the original form) as the a allele, and the derived (that is, the one arising after mutation from the ancestral form) one as the A allele. Since selective pressure after the change in the fitness function increases the frequency of the allele, some of those alleles might eventually reach fixation in the derived form A. This event is particularly important because although the allele has a large positive effect on the phenotype, it would be undetected by the GWAS analysis, hence it would not be among those considered in our analysis.

6 Data

6.1 Ancient DNA

We use version 44.3 of Allen Ancient DNA Resource (AADR) (David Reich Lab 2021) covering 5,225 ancient and 3,720 present-day individuals genotyped at 1,233,013 sites. Due to low coverage, most of the ancient DNA is pseudohaploid, meaning that at each site one of the alleles has been chosen independently at random. The dataset comes with rich individual-level information, including estimated dates, geographical coordinates, genotyping coverage and family membership. Most of the samples were radiocarbon dated or dated using archaeological context.

We restricted the sample to individuals from Western Eurasia (west of 60°E and north of 35°N) dated between 38,000 and 1,100 years before present (BP), similar to Cox et al. 2019b. We also remove neanderthal samples and samples with fewer than 15,000 SNP hits on autosomal targets Mathieson et al. 2018. We identify duplicate individuals by uniqueness of their master ID and by pairwise identity-by-state (IBS) value Ju and Mathieson 2021. In the latter case, for each individual we select a person with highest IBS value and denominate pair as duplicates if IBS z-score is above 7. Between all duplicated individuals, we choose one with highest number of SNP hits on autosomal targets. Furthermore, among samples identified as members of a single family, we select

one with the highest number of autosomal SNP hits. At the end, our working sample consists of 2,328 ancient individuals.

The original data contains 1,233,013 variants, out of which 49,704 and 32,670 are on chromosomes X and Y, respectively. We remove variants with missing rates above 0.99. The working dataset contains 1,232,725 variants. Share of missing genotypes per SNP ranges from 0.19 to 0.99 with average 0.59 (sd = 0.14). The missing rate tends to be larger in the sex chromosomes.

We use the AADR in order to construct a vector of initial allele frequencies at around 14,000 years BP. For this purpose, we run a supervised ADMIXTURE (Alexander and Lange 2011) algorithm described in Mathieson and Mathieson (2018) and Mathieson et al. (2018). In particular, using the sample labels presented in the Supplementary Table 1 of Mathieson et al. (2018), we assign individuals to one of four groups: western hunter-gatherers (WHG), eastern hunter-gatherers (EHG), Anatolyan Neolithic farmers (AN) and Yamnaya Samara steppe individuals (YS). Together with ancestry shares in the genotypes of the sample members, the ADMIXTURE also returns the estimated allele frequencies in each of these populations. We use the estimated allele frequencies among WHG as one of the initial allele frequency vectors.

6.1.1 Descriptive statistics

Results from ancient DNA have to be considered with great care. In particular, there may be concerns that AADR contains a biased sample of ancient populations. For example, differences in burial procedures could result in over- or under-representation of some socio-economic groups. The archaeological literature documents a number of burial practices coexisting around 11K years BP, including the earliest direct evidence of cremation dated at $\sim$ 9K years BP (Bocquentin et al. 2020). Thus, unless the choice of burial method was random, skeletal samples available today for sequencing may represent a biased view of ancient populations. The AADR dataset does not contain information about the social position of sample members but does contain information about gender and age at death (Figure 1). There is little information about very old samples (older than 10K years BP). The sex composition is balanced: about half of skeletal remains from 10K years BP and earlier are female.

There may also be discrepancies between the archaeological context and skeletal remains used in DNA sequencing as bones may have been displaced over time (Charlton, Booth, and Barnes 2019). This can introduce measurement error in dates, if skeletal remains were dated based on archaeological context only (37.5% of full sample). However, Figure A-3 in section A-2 shows a similar pattern in diffusion of AN ancestry both over time and space as the spread of farming presented in Shennan (2018) and reprinted here in Figure A-2. Figure A-4 in section A-2 shows more closely the evolution of ancestral

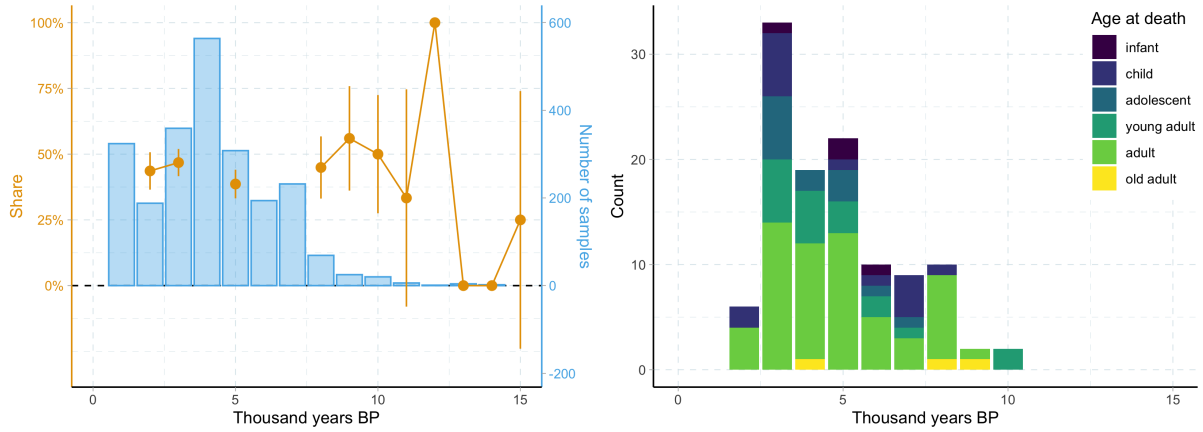


Figure 1: Sex and age of skeletal remains

Note: the left panel plots number of samples with non-missing sex (99.6% of full sample) information and share of females among them over time. The right panel plots counts of ancient individuals by age at death over time. This information is only available for 113 individuals out of 2,328.

genotypes over time among the AADR samples. The WHG genotypes were dominant among samples dated prior to 10,000 years BP. Around 8,000 years BP we can see that AN ancestries became dominant ancestry, almost entirely displacing the WHG.

6.2 1000 Genome Project

Our primary analysis dataset is the 1000 Genomes Project Consortium et al. (2015) (1000GP) containing data from 2,504 individuals from 26 populations and more than 88 million variants. In our analysis, we consider 503 individuals from EUR populations. We subset biallelic SNPs, remove SNPs with minor allele frequency below 1% and set p-value threshold for Hardy-Weinberg equilibrium exact test at 10^{-6}. From the remaining SNPs we select the subset that is present in the AADR. This leaves us with 909,809 SNPs in the working dataset.

6.3 GWAS estimates

We use GWAS summary statistics for educational attainment from Lee et al. 2018. The original table contains results for more than 10 million SNPs. In order to identify the set of lead SNPs, we apply clumping algorithm of Lee et al. (2018). This is an iterative algorithm, that initiates at SNP with the lowest p-value and assigning all SNPs in that chromosome with the coefficient of squared correlation $r^2 > 0.1$ to the respective clump. The process is repeated in the set of unassigned SNPs, until no SNPs remain with the p-value below 5×10^{-8}. We use the working 1000GP dataset described in Section 6.2 as the LD reference panel. After the clumping, we have 475 lead SNPs.

Finally, we normalize the alleles in a trait-increasing direction. For example, the SNP

rs12735232 is biallelic with reference allele T and alternative allele C. The original data reports the GWAS coefficient of allele T at -0.0131. Therefore, allele T is trait-decreasing and conversely allele C is trait-increasing. Our working dataset reports allele frequencies and GWAS coefficients of trait-increasing alleles, which in this example is frequency of allele C and the GWAS coefficient +0.0131. We repeat this process for all lead SNPs.

6.4 Reconstruction of polygenic score history

Ancient DNA extracted from the archaeological remains may present a non-random sample of ancient populations and, therefore, lend a biased view into the genetic history. An alternative source of information about genetic past is already contained in the genotypes of modern individuals. The coalescent theory provides a graphical representation of this history in the form of a coalescent tree. Edge and Coop (2019) propose three estimators based on the coalescent tree to reconstruct the path of allele frequencies and, hence, phenotypes in the past.

We apply the algorithm presented by Edge and Coop (2019) to the set of 475 analysis SNPs. For each of these SNPs in the analysis sample, we use haplotypes of European populations in the 1000GP within 100K basepair window around the SNP. Using RENT+ software and these haplotypes, we construct a coalescent tree. The topology of the tree and branch lengths provide information about lineages across time as well as the timing of coalescent events. Finally, this information is plugged into the three estimators presented in Edge and Coop (2019): proportion-of-lineages, waiting-time and lineages-remaining estimators. Here, we provide a brief overview of these estimators and refer to Edge and Coop (2019) for a complete discussion of estimator properties.

The proportion-of-lineages estimator is the most straightforward estimator of the three. It counts the number of lineages carrying an allele of interest at a given point in time. The caveat of this estimator is that in the presence of selection, lineages carrying the favourable allele are over-represented among the ancestors of modern individuals compared to a general ancestral population at the time. The two other estimators, waiting-time and lineages-remaining, relax the neutrality assumption (no selection). Both of these estimators estimate population size in each subtree separately and use them to construct an estimate of allele frequency that does not assume neutrality. The waiting-time estimator is motivated by the inverse relationship between effective population size[8] and the span of time between two coalescent events. The lineages-remaining estimator uses the prediction about the number of coalescent events expected to occur between two points in time as a function of effective population size.

The original algorithm in Edge and Coop (2019) returns the reconstructed histories

8. The effective population size is the number of unique haplotypes in the population that can be passed on to the next generation and may be smaller than the count of individuals in the population.

with time measured in approximate coalescent units. The coalescent time τ accrues according to the following formula: $\tau(t) = \int_0^t \frac{1}{N(z)} dz$ where t is time in generation units and $N(t)$ is the population size at time t[9]. We convert the coalescent-unit time to generation-unit using this formula and the estimated population size over time published in Speidel et al. (2019)[10].

6.5 Descriptive evidence

The goal of this paper is to examine whether the widespread adoption of agriculture gave positive impulse to selection of genotypes that are associated with higher level of skills. In this paper we use modern measurement of educational attainment (EA) as such measure of skills. Figure 2 plots the EA polygenic score (PGS) using the genotypes of ancient individuals across time to which the archaeological remains belong. The figure shows that, indeed, the EA PGS started rising around 8,000 years BP, after the adoption of agriculture. Prior to that it remained essentially constant over millennia.

Figure 3 provides an alternative source of genetic history, following the algorithm presented in Edge and Coop (2019). The three panels plot the average path of the phenotype according to each of the three estimators: proportion-of-lineages, waiting-time and lineages-remaining. The two horizontal lines also show the average phenotype in WHG and modern EUR populations. By and large, the reconstructed paths also suggest that EA phenotype was stable in the past and started rising approximately within the past 14K years BP. The proportion-of-lineages estimator suggests that the phenotype started rising earlier. Given that this estimator overestimates the frequency of selected allele, it provides the first hint that our phenotype of interest was indeed subject to positive selection in the last 14K years BP.

7 Estimation Results

7.1 Parameter Estimation in Full Model

In this section we present the results of the parameter estimation when the data generating process is the full model, as presented in section 3, and in particular the model in section 3.3, which describes the process on haplotype pairs.

The fitness function we adopt here is the linear function (so the selective shift on probability next generation in step is exponential), for both production functions, foraging

9. Note that time moves backwards. Thus, $t > 0$ means t generations in the past.

10. The demographic history is esitmated from the coalescent tree using an iterative algorithm. First, Speidel et al. construct coalescent trees using constant population size. Then, given the branch lengths of this tree, they find an ML estimate of population size over time. Third, they re-estimate the coalescent tree using the estimated demographic history. The process is repeated until convergence is achieved. For a more detailed discussion, see Speidel et al. (2019).

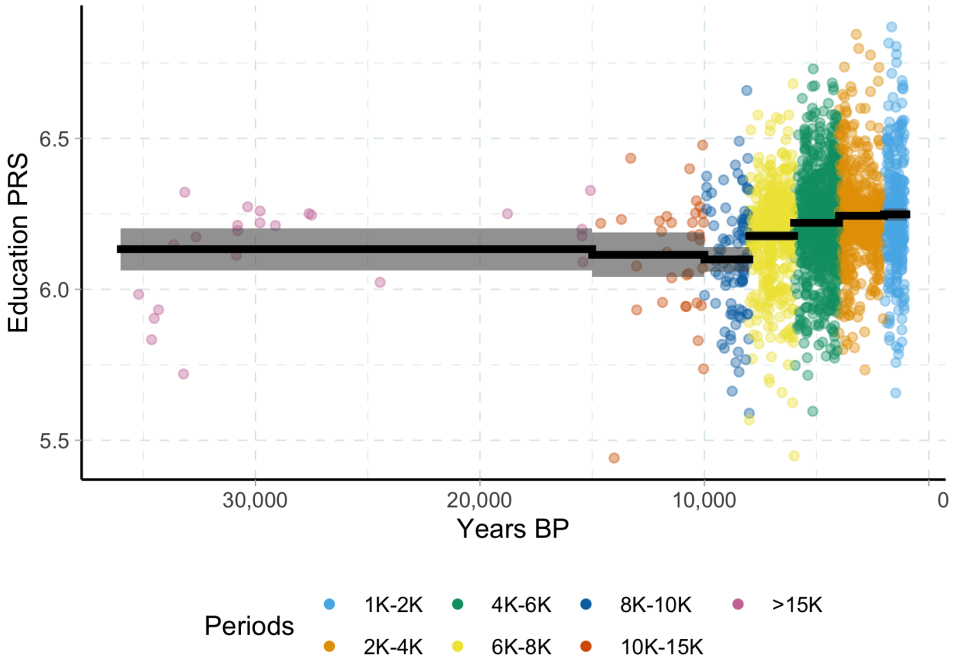


Figure 2: Education PGS over time: AADR

Note: the figure plots the polygenic scores (PGS) of ancient individuals in the analysis sample (points) and average PGS score in each period (solid black line). The shaded area around the black line corresponds to 95% confidence interval based on t-distribution.

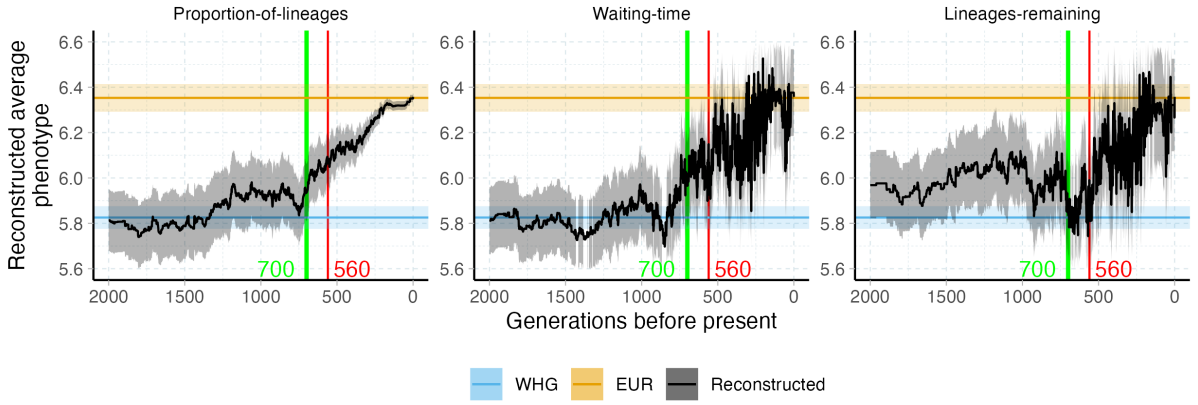


Figure 3: Education PGS over time: EC reconstructions

Note: the figure plots the reconstructed polygenic scores using the three estimators presented in Edge and Coop (2019) as well as average phenotypes in WHG and modern EUR populations (two horizontal lines). The shaded areas correspond to the 95% probability interval of the phenotype distribution. The approximate coalescent-unit time is converted to generation-unit time on the x axis using the demographic history of CEU population published in Speidel et al. (2019). The vertical red line marks 560 generations period that roughly corresponds to approximately 14K years BP.

and farming. This is a departure from a more common assumption, in the literature on selection, of a Gaussian fitness. The departure is motivated by the consideration that it is not clear why the effect of characteristics associated with the ability to attain education should have a maximum effect on productivity and fitness, after which the effect changes direction, being first increasing and then, for higher values, decreasing .

The set of parameters in our hypothesis is the set of vectors (ω, M) of strength of the selection coefficient in the farming technology and the population size. We assume that the distribution of the genotype before the climate change following the Younger Dryas was at the steady state for an economy in which all individuals chose foraging. After the climate change, the productivity of farming may increase, to a new value ω. We consider as possible new values of ω both positive and negative, so we do not assume that in the new situation these characteristics are more favorable in farming than in foraging. In particular we do *not* assume that higher cognitive skills are more favorable in farming then in foraging. The set of fixed parameters, that is those that are given by biological constraints (for example the recombination rate) are as specified in section 3.3. The value of the mutation rate (1.2e-8) was chosen in line with the results in Campbell et al. (2012), Scally and Durbin (2012), Shendure and Akey (2015), and Tian, Browning, and Browning (2019). We now present the results of our analysis.

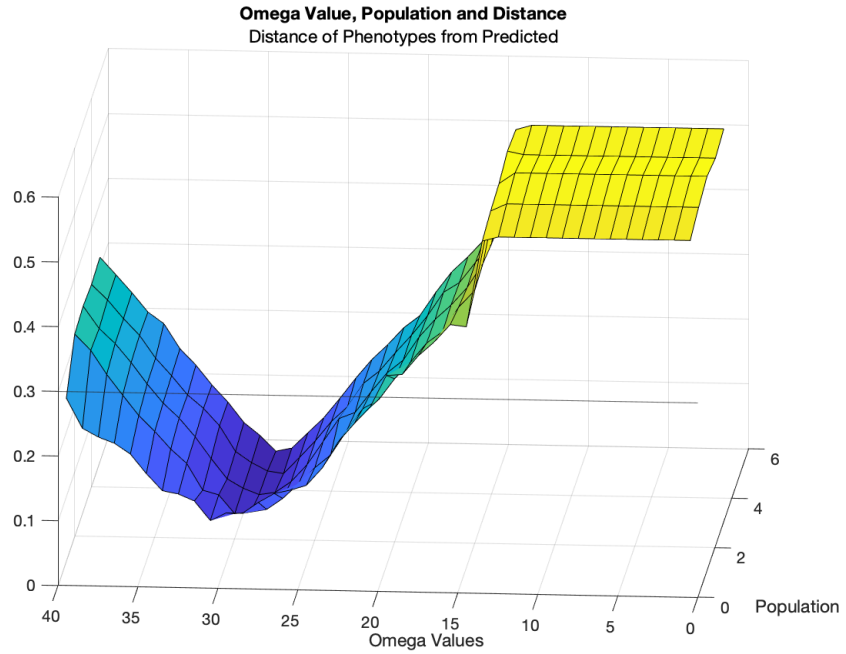
Figure (4) reports the values of the distance from the true final phenotype of the final phenotype predicted by the full model at different values of the parameters (ω, M). The figure shows that the function we minimize is well behaved, and has a minimum approximately at value $\omega = 0.04$. The precise value at which the minimum is achieved only depends weakly on the population size. The flat portion of the surface for low (negative) values of ω corresponds to fitness functions for farming that are inferior, in the relevant region, to foraging. In this region the value of the predicted phenotype is approximately constant, equal to the steady state value of the phenotype in an economy in which only foraging is practiced, and thus the difference from the true value does not depend on ω. As figures (5) and (6) will show, the population size affects the variance of the predicted values and of the estimators, rather than their means, as should be expected.

We then test the the null hypothesis that $\omega = 0$ in the full model. Figure 5 reports the distribution of mean phenotype over computations of the allele paths at the null hypothesis, for several values of the population size.

The variance of the mean phenotype over replications decreases as population size increases, as expected. The two panels of figure (5) prove that even at small population sizes the null hypothesis is rejected.

Using the computed paths we can estimate of the value of ω that maximizes the likelihood over an initial sample of simulated paths following the full model. Figure (6), top panel, reports the values of the likelihood of observing the true final phenotype at

Figure 4: **Distance of predicted phenotype from true final phenotype.** Predicted values are generated by the full model, as presented in section 3.3. Values of ω and M on the horizontal axes, distance on the vertical axis.



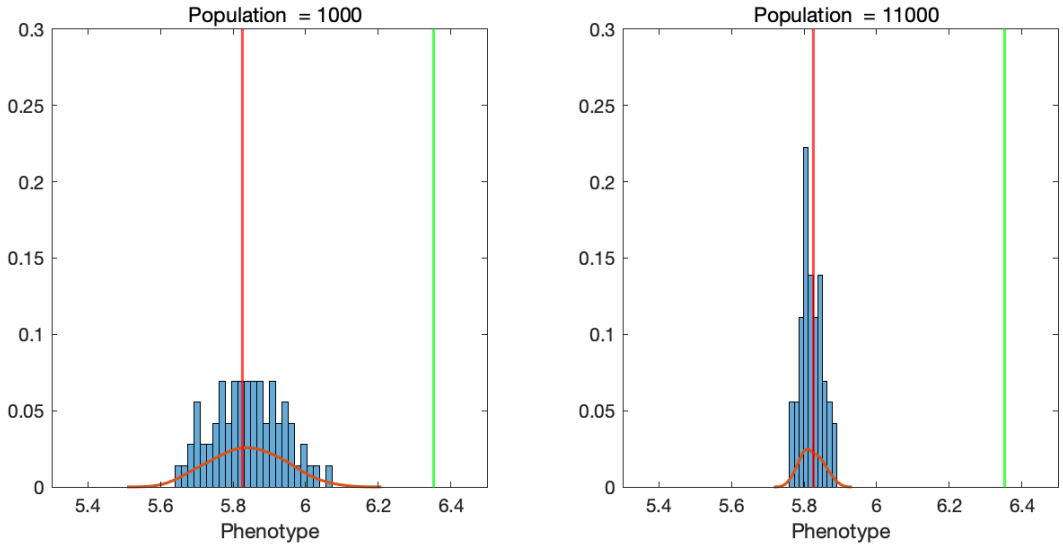
different values of the parameter vector. Values of ω are reported on the horizontal axis. Each line in the right side of the figure refers to a specific population size. Consider any of the lines. For each value of ω, at that fixed population size, we have an empirical distribution on final values of the final phenotype given by the set of realizations produced by the full model with that value of ω and that population size.

The figure shows that again the variance of the distribution varies with the population size, and decreases as expected with the size. Thus, the optimal value of ω depends on the population size. It is however close to the value $\omega = 0.04$.

7.1.1 Truncation of alleles at the boundary

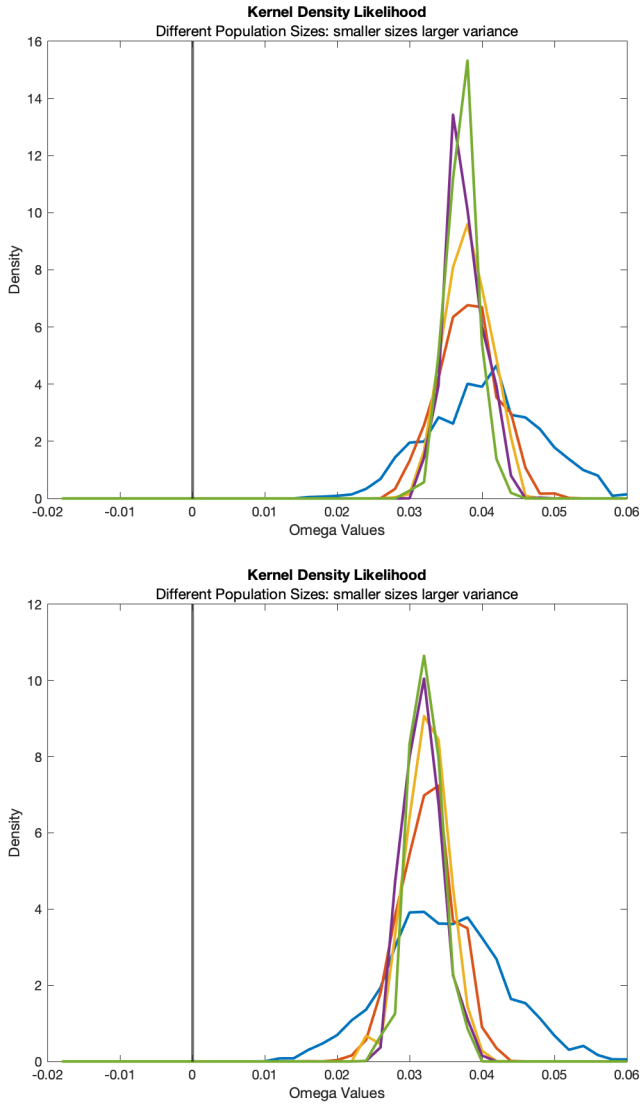
Approximately 7 % of the alleles had initial value frequency below the frequency of .025, or larger than .975, and a final frequency within the interval $(.025, .975)$. The model we propose takes into account this possibility allowing for a mutation to occur that introduced a new allele in the population. Another natural possibility is that the allele was present in low frequency in the population. Allowing only the possibility that new alleles are introduced by mutation rather than considering the possibility that they start from low frequency, and thus initially unobserved, might produce a higher estimate of the coefficient to allow the frequency to raise from 0 to higher values. A conservative way to deal with this possibility is to eliminate these alleles from the analysis. The ω estimated by MK in this way is likely to be smaller than the one obtained with an estimation

Figure 5: **Mean of Final Phenotype at Null Hypothesis.** Predicted values are generated by the full model, as presented in section 3.3. In both panels, the histograms reports the frequency of mean phenotype over computations of the alleles path, with selection parameter $\omega = 0$. Red vertical line: Initial Phenotype. Green vertical line: Final Phenotype. Red Curve: kernel Density. Left panel: small population size (M=1000); Right panel: larger population size (M=11000).



using the entire sample. Figure (6), bottom panel, reports the values of the likelihood after these alleles are eliminated. As expected, the estimated maximum likelihood shift slightly to the left; however, it is positive.

Figure 6: **Maximum Likelihood Estimator** Kernel density of the likelihood function for different values of the population. Lines indicating larger variance of the distribution corresponds to smaller population size.



8 Conclusions

We have modeled the evolution of the distribution of genotype in European populations' recent past (within 14 thousand years before present). We hypothesized that the evolution was driven by selection operating after a shift in the productivity of agriculture, induced by a well documented climate change, in a standard Roy model in which individuals self-select into one of two sectors (foraging or farming).

We tested the model in two data sets, one of ancient and one of modern DNA, matching the observed distributions of genetic variables of interest (allele frequencies and lineages). The model extends a standard Wright-Fisher model, allowing a more general fitness function that models the role of the two technologies, and includes sexual reproduction. We estimate the model and find support for the hypothesis that a major shift in distribution of allele frequencies (in a direction favoring higher cognitive skills) occurred after the climate warming at the end of the Younger Dryas (11,600 years BPE) made agriculture more productive than hunter-gathering. The relative size of the effect of cognitive skills in the two sectors is estimated and is not assumed *a priori*. The general implication we draw is that historical transformations (in our case climate change and technological change) can affect the distribution of genotype and thus institutions, rather than the other way round.

Important future research and extensions are possible. As we discuss in the section on the role of migration (4.2), an important open question is the way in which the shift from foraging to farming occurred. One possibility is that descendants of the same population changed the chosen activity; another possibility is that a different population, migrating from another region, replaced the original one. In the first case, the selective pressure operated in a given population, favoring some of the individuals; in the second, it favored a population over another. A combination of these two possibilities is most likely, and the consensus of research seems to support this conclusion. Deciding this will ultimately provide a precise estimate of how far the random mating assumption is from reality.

A Proofs

Proof of theorem 3.1

The process described in section 3.3 has several steps, starting from an initial state in the state space $\Delta_N(\mathbf{H})$ and concluding with a probability distribution over the same state space. An equivalent formulation that we use to prove the statement considers a sequence of finite state spaces of the general form, with M and L two positive integers:

$$\Sigma_L^J \equiv \{m = (m(1), \ldots, m(J)) : \sum_{j=1}^{J} m(j) = L\}. \tag{13}$$

It is clear that each step in the process described in section 3.3 defines a map from a set of the form Σ_N^L to probability distributions on a set of the same form, although possibly different values of N and L. For example, the mutation step maps from the set $\Sigma_N^{\#\mathbf{H}}$ to probability distributions over the same set; and the same does the recombination step. The matching step maps an element in $\Sigma_N^{\#\mathbf{H}}$ into probability distributions over pairs of gametes, for a total number of possible pairs equal to M, that is $\Sigma_M^{\#(\mathbf{H}\times\mathbf{H})}$. The overall process is thus a composition of stochastic maps, all over finite spaces, with initial and final state equal to $\Sigma_N^{\#\mathbf{H}}$. This initial and final state corresponds to $\Delta_N^{\mathbf{H}}$. Hence an invariant measure on this set exists.

Proof of proposition 3.2

Take two loci such that the probability of recombination across them is $\rho > 0$, call X and Y (recall these are $\{0, 1\}$ valued random variables). At the invariant measure the allele frequency is $\frac{1}{2}$ for both, and does not change in the next period. When recombination occurs at one or both of the loci, the new allele is chosen independently, and has mean equal to 0. It follows from this and the definition of recombination that $cov(X, Y) = (1 - \rho)^2 cov(X, Y)$, from which our claim follows.

Proof of theorem 4.1

For every finite value of the population size, the state space in every stage is finite. For example, the allele frequency $p \in ([0, 1]_N)^K$. The function defined in equation (10 is the composition of several functions, all each one corresponding to a stage in the process described in section 3.3. For example, the fist step mutation maps the initial allele frequency p to a probability on $([0, 1]_N)^K$ induced by mutation (given by equation (16) below). Similarly the random mating defines a map from the q frequency in $([0, 1]_N)^K$ to $\Delta_N\left(\{0, 1, 2\}^K\right)$.

Proof of lemma 4.2

The mutation step takes any haplotype pair (l, r) and mutates each allele with a probability α. The new frequency is given as described in section 3.2.1.

The next population is chosen according to the Wright-Fisher model, as the realization of a multinomial random variable with N outcomes and probability over outcome induced by the current frequency, distorted by the selection in direction of the more favorable genotypes. We describe this in steps.

We denote q the realization of the random variable vector Q. Random mating produces the probability distribution on children of random mating to be

$$(\otimes_{k=1}^{K} Hq)(g) = \Pi_{k=1}^{K} Hq(g(k), k) \tag{14}$$

where we have defined the Hardy-Weinberg probability

$$\forall k : Hq(\cdot, k) \equiv \left((1 - q(k)^2, 2q(k)(1 - q(k)), q(k)^2\right). \tag{15}$$

Repetition of the random mating produces a set of children, from which the next population of size N is going to be selected. We apply the fitness function to determine the next generations of children reaching reproductive age. This is the selection step in Wright-Fisher model. The vector of population numbers of the different genotypes is:

$$n(\cdot, t+1) = \mathbf{mn}(N, F \otimes Hq) \tag{16}$$

where $F \otimes Hq \in \Delta(\mathbf{G})$ is defined by:

$$\text{for every } g \in \mathbf{G}, F \otimes Hq(g) \equiv \frac{e^{F(z(g))}(\otimes_{k=1}^{K} Hq)(g)}{E_{(\otimes_{k=1}^{K} Hq)} e^{F(z(\cdot))}} \tag{17}$$

where $E_{(\otimes_{k=1}^{K} Hq)} e^{F(z(\cdot))}$ is the average fitness in the population. Apply the definition of next period p defined by $p = \frac{1}{N} An$ (see section (3.1)) to the new population $n(\cdot, t+1)$ to get the new frequency

$$Ep(\cdot, t+1) = \frac{1}{N} AE(n(\cdot, t+1)).$$

Now use the fact that the expectation of the multinomial $n(\cdot, t+1)$ is $F \otimes HqN$. We subtract $p(\cdot, t)$, add and subtract q and rearrange. If we denote for $k \in \mathbf{K}$ and $i \in \{0, 1, 2\}$ the expected contribution to the selection coefficient if the genotype at k is i:

$$F_i^k \equiv \frac{E_{(\otimes_{-k} Hq)} e^{F(z(i,\cdot))}}{E_{(\otimes_{k=1}^{K} Hq)} e^{F(z(\cdot))}}. \tag{18}$$

where $E_{(\otimes_{-k} Hq)}$ denotes the product measure on all the coordinates except k, and $F(z(i, \cdot))$ is defined on all the coordinates except k. From this we obtain (11).

B Limit Models

In this section we consider useful limit processes the population size (section B.1), the number of loci (section B.2) tend to infinity, and finally (section B.3) the continuous time limit.

B.1 Large Population Limit

The expected value of the frequency change at a given N, $E(\Delta p|N)(k)$, is the expectation over realizations of q of of the conditional value $E(\Delta p|q,N)(k)$. The limit when the population becomes large can now be derived. The new frequency is derived from the cumulated effect of the selection on all the individuals. Since the selection effect is non-linear, it depends on the rest of the genotype of the individual.

We denote the expected value of Q, defined in equation (3):

$$Mp = (1-2p)\alpha + p. \tag{19}$$

Corollary B.1. *For every locus k, and initial frequency p:*

$$\lim_{N\to\infty} E(\Delta p|N)(k) = Mp(k) - p(k) + \tag{20}$$

$$Mp(k)(1-Mp(k))\left((1-Mp(k))(F_1^k - F_0^k) + Mp(k)(F_2^k - F_1^k)\right).$$

where the F_i^k are computed at the frequency Mp.

Proof. Take the expectation over the realization of the mutation process of the equation (11) to get $E(\Delta p|N)$. The statement follows from the weak convergence of the variable Q in N to Mp, and the continuity of the function of p. □

B.2 Large K dynamics

If we consider the fact that, in a highly polygenic phenotype, the contribution of each allele to the phenotype is small compared to the overall contribution of others, the we can derive a simpler and easier to interpret condition.

Corollary B.2. *For every locus k, and initial frequency p:*

$$\lim_{N\to\infty} E(\Delta p(k)|N) = Mp(k) - p(k) + \tag{21}$$

$$\beta(k)Mp(k)(1-Mp(k))\frac{E_{(\otimes_{-k}Hq)}\left(e^{F(z(0,\cdot))}F'(z(0,\cdot)\right)}{E_{(\otimes_{k=1}^{K}HM(p))}e^{F(z(\cdot))}}.$$

Equation (21) states that the change in frequency is proportional to the product $q(k)(1-q(k))$ of current frequencies, times the product of two terms. One term is term

specific to k, $\beta(k)$, which determines the effect of the phenotype of that allele, with a positive sign if the effect is positive. The other is the weighted average of the slope of the fitness function over the genotype. The size and sign of this second term is common to all alleles, and determines whether, at the current distribution over the phenotype, an increase in the phenotype is beneficial to fitness. In the limit of large population, the dynamic process is deterministic, and the steady state at each locus is given by the balance of the two terms on the right hand side of equation (21) the first describing the effect of selection, mediated by the term, common to all loci, of the average derivative of the fitness function, the second the effect of mutation.

B.3 Continuous Time Limit

It is useful to understand the continuous time limit of the discrete time model developed earlier. In this section, time is continuous and indexed by t. To derive a time limit we consider the standard construction that maps the change in frequency occurring in one discrete time unit into a change in continuous time in the time interval Δt while the population size also increases tending to infinity. In particular we take: $\Delta t = \frac{1}{N}$. For the limit to be well defined we require

Assumption B.3. *In the time interval Δt, the effect of the selection function is $e^{F(z)\Delta t}$ and the mutation probability is $\alpha_0 \Delta t$.*

To make the distinction between discrete and continuous time clear we denote the vector of allele frequencies with $x \in [0,1]^K$, writing $x(k,t)$ as the frequency of the k^{th} allele of type 1 as $x(k,t)$. We link the two processes with the following definition:

$$x(k,t) \equiv p(\lfloor Nt \rfloor) \tag{22}$$

and linear interpolation in the intervals.

Theorem B.4. *The continuous time limit process under the assumption (B.3) on fitness and mutations rate has a variance term defined by (29) below, and it is approximated by:*

$$dx(k,t) = \tag{23}$$

$$\left(\beta(k)x(k,t)(1-x(k,t))E_{\otimes_{-k}Hx}F'(z(0,\cdot)) + \alpha_0(1-2x(k,t))\right)dt$$

$$+\frac{1}{2}\left(x(k,t)(1-x(k,t))\right)^{\frac{1}{2}} dW(t).$$

The reason why we state that equation (23) is an approximation is discussed when we analyze the diffusion term. An intuitive understanding of the equation (23) follows the lines of the corollary (B.2). The term due to mutation, $\alpha_0(1-2x(k,t))$, affects the

frequency by increasing the less frequent allele. The term due to selection has sign and size determined as in corollary (B.2) by the coefficient of the allele on the phenotype ($\beta(k)$) times the term, which is common to all alleles, given by the average derivative of the fitness over the genotypes.

Proof. Consider first the drift term. As in the derivation of the equation (18) we add and subtract q, and take the limit of the first term (the selection term) and the second term ($q - p$, the mutation term) separately.

Consider first the selection term. We use the correspondent of the equation (11) obtained using the assumption (B.3), so that in the current case:

$$F_i^k \equiv \frac{E_{(\otimes_{-k} Hq)} e^{F(z(i,\cdot))\Delta t}}{E_{(\otimes_{k=1}^K Hq)} e^{F(z(\cdot))\Delta t}}. \tag{24}$$

Recall that also the distribution of the q variable depends on Δt according to assumption (B.3), and we denote it using the cdf $F(q|\Delta t)$. We now compute the selection component of the drift term as:

$$\begin{aligned} & \lim_{\Delta t \to 0} \frac{1}{\Delta t} E((\Delta p|\Delta t) \\ = & \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{[0,1]^K} E((\Delta p|q, \Delta t) dF(q|\Delta t) \\ = & \lim_{\Delta t \to 0} \int_{[0,1]^K} q(1-q) G(q, \Delta t) dF(q|\Delta t) \end{aligned}$$

where

$$G(q, \Delta t) \equiv (1-q)\frac{(F_1 - F_0)}{\Delta t} + q\frac{(F_2 - F_1)}{\Delta t}$$

Now note that for $i = 0, 1$:

$$\lim_{\Delta t \to 0} \frac{(F_{i+1} - F_i)}{\Delta t} = E_{(\otimes_{-k} Hq)} \left(F(z(i+1, \cdot)) - F(z(i, \cdot))\right) \tag{25}$$

because clearly

$$\lim_{\Delta t \to 0} E_{(\otimes_{k=1}^K Hq)} e^{F(z(\cdot))\Delta t} = 1.$$

Note that the convergence is uniform in q. It follows that

$$\begin{aligned} \lim_{\Delta t \to 0} G(q, \Delta t)(k) = (1 - x(k))(E_{\otimes_{-k} Hq} \left(F(z(1, \cdot)) - F(z(0, \cdot))\right) + \\ x(k)(E_{\otimes_{-k} Hq} \left(F(z(2, \cdot)) - F(z(1, \cdot))\right) \end{aligned} \tag{26}$$

and hence the drift term in equation (23).

We consider now the mutation term. Since q is the realization of the random variable

$Q \equiv p + \frac{m-n}{N}$ with $m \sim B((1-p)N, \alpha_0 \Delta t)$ and $n \sim B(pN, \alpha_0 \Delta t)$, clearly

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \left(E_{\Delta t} Q - p \right) = (1 - 2p)\alpha_0. \tag{27}$$

We now consider the diffusion term. Recall that the matrix A has been defined in equation (2); $n(1) \equiv (n(g,1) : g \in \mathbf{G})$ is the next period population.

$$\begin{aligned} Var(\Delta p) &= Var(\frac{1}{N} An(1)) \\ &= \frac{1}{N} Var(A\xi) \end{aligned}$$

where ξ is a random variable with genotype g occurring with probability $F \otimes Hq$.

We now consider for a locus k the random variable $A_k \xi = \frac{1}{2}\xi(k)$, which is distributed for $x \in \{0, 1, 2\}$ with value:

$$\frac{x}{2} \text{ with probability } \sum_{g_{-k} \in \mathbf{G}^{-k}} \left(\frac{e^{F(z(x, g_{-k}))\Delta t} \otimes_{-k} Hq(x, g_{-k})}{\sum_{g' \in \mathbf{G}} e^{F(z(g')\Delta t} \otimes Hq(g')} \right)$$

We now consider two loci k and j and consider the limit

$$\lim_{\Delta t \to 0} \frac{E\left((A_k \xi - E(A_k \xi))(A_j \xi - E(A_j \xi))\right)}{\Delta t} \tag{28}$$

The numerator of the ratio in (28) tends to 0 because in the limit the two variables $A_k \xi$ and $A_j \xi$ are independent, so the limit in (28) is the derivative of the numerator. One can check that the terms of the derivative are zero except one that gives:

$$\sum_{x,k \in \{0,1,2\}} Hq(x,k) Hq(y,j) \Phi(x,y) \left(\frac{x}{2} - E(A_k \xi) \right) \left(\frac{y}{2} - E(A_j \xi) \right) \tag{29}$$

where the distortion factor Φ is defined by:

$$\Phi(x,y) \equiv \sum_{g^{-k,-j} \in \mathbf{G}^{-k,-j}} F(z(x, y, g^{-k,-j})) \tag{30}$$

for any x, y.

□

In sections below we have shown that. although it is an approximation, the simple model is a good approximation of the complete model developed earlier. In this section we present the estimates of the parameters in the simple model.

B.4 Simple model

In this section we estimate the selection parameters in the diffusion process presented in Equation (23). In particular, we consider a linear fitness function where $\mathbb{E}_{\otimes_{-k}Hx}F'(z(0,\cdot)) = \omega$:

$$dx(k,t) = [x(k,t)(1-x(k,t))\beta(k)\omega + \alpha_0(1-2x(k,t))]\,dt + \sqrt{x(k,t)(1-x(k,t))}dW(t) \qquad (31)$$
$$t \in [0,T]$$
$$x(k,0), \beta(k) \text{given}$$

Here, ω is the parameter of interest, the sign and magnitude of which captures the direction and strength of selection. We set reference alleles to trait-increasing alleles, which implies that $\beta(k) \geq 0, \forall k \in \{1,\ldots,K\}$. Therefore, positive values of ω imply positive selection of education-enhancing alleles.

It is also worth noting that the Equation (31) depends on the population size N via the terminal time T parameter. Recall that the model presented in Section 3.3 measures time in terms of generations. For example, in this paper we consider the distance between the WHG population and modern EUR population to be approximately equivalent to 560 generations[11]. The derivation of the continuous-time approximation clearly shows that the time duration is inversely scaled by the population size. That is, simulating the Equation (31) until $T = \frac{560}{N}$ is equivalent to simulating the main model for 560 generations.

B.4.1 Estimators

We propose two estimators for the parameters $\theta = (\omega, T)$: the non-linear least squares estimator $\hat{\theta}^{\text{NLS}}$ and the maximum-likelihood estimator $\hat{\theta}^{\text{MLE}}$.

$$\hat{\theta}^{\text{NLS}} = \arg\min_{\theta} \frac{1}{K}\sum_{k=1}^{K}\left(x(k,T) - \mathbb{E}_{\Psi}[x(k,T)|x(k,0),\beta(k),\theta]\right)^2 \qquad (32)$$

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta} \sum_{k=1}^{K} \ln \Psi\left(x(k,T)|x(k,0),\beta(k),\theta\right) \qquad (33)$$

where $\Psi\left(x(k,T)|x(k,0),\beta(k),\theta\right)$ is the probability density function of the final allele frequency $x(k,T)$ conditional on initial conditions $(x(k,0),\beta(k))$ and parameter vector θ.

The distribution of final allele frequencies Ψ can be found by integrating the diffusion process in Equation (31) over time given the initial conditions. However, we do not have an analytic solution to this integral. Instead, we use two approaches to approximate Ψ: one based on additive regression with normally distributed error term (normal) and one based on numerical evaluation of the diffusion process (numerical). Appendix A-4

11. Assuming average generation length is 25 years, this is equivalent to 14,000 years.

discusses these approximations in detail. We distinguish between the two sets of estimators using subscripts $\mathcal{N}$ for normal approximation and B for numerical approximation[12]. That is, estimators obtained under normal approximation are denoted $\hat{\theta}^i_{\mathcal{N}}$ and estimators obtained under numerical approximation are denoted $\hat{\theta}^i_B$, $\quad \forall i \in \{\text{NLS}, \text{MLE}\}$.

The estimators presented above weigh SNPs equally. We can make the NLS estimator a little closer to the estimator used in the main model by setting the importance of each SNP proportional to its GWAS coefficient. Namely, the estimator now solves the following problem

$$\check{\theta}^{\text{NLS wt}} = \arg\min_{\theta} \frac{1}{K} \sum_{k=1}^{K} \left[\beta(k) \left(x(k,T) - \mathbb{E}_{\Psi}[x(k,T)|x(k,0), \beta(k), \theta]\right)\right]^2 \tag{34}$$

B.4.2 Consistency of estimators

We examine the properties of the estimators using Monte Carlo simulations. To do so, we first generate synthetic datasets given some true parameter vectors θ_0. We generate these datasets such that (i) for each parameter vector θ_0 we have a sample of 475 SNPs that belong to $\mathbf{V}_T(\nu) \cap \mathbf{S}_T(\alpha)$, and (ii) for each SNP and parameter vector θ_0 we have $P = 100$ random samples. Thus, for each true parameter vector θ_0, we obtain P estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(P)}$, which allows us to analyse the bias $\hat{\theta} - \theta_0$ statistically.

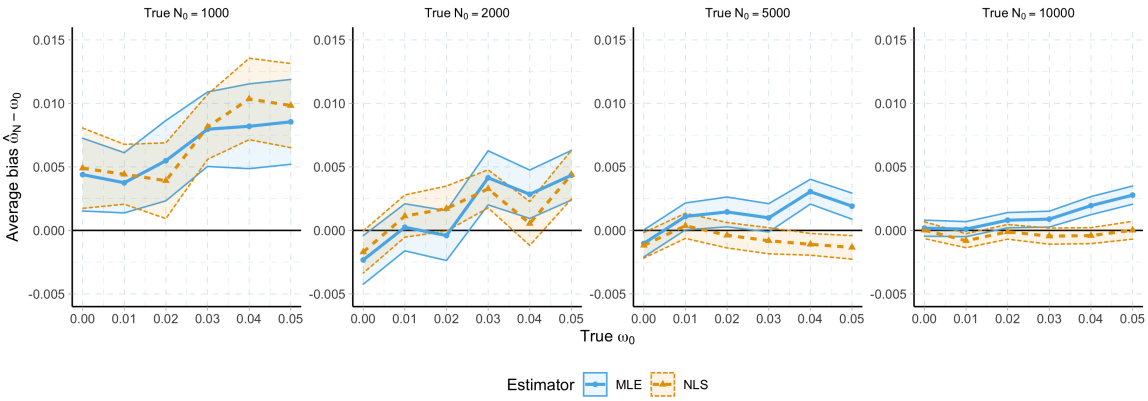
Figure 7 plots average bias of $\hat{\theta}^{\text{NLS}}_{\mathcal{N}}$ and $\hat{\theta}^{\text{MLE}}_{\mathcal{N}}$ given the truncated normal approximation. Both estimators of ω seem to be performing fairly well in terms of average bias, with the NLS esitmator having slightly lower bias than the ML estimator. Both estimators tend to overestimate the true parameter ω_0 at low population size $N = 1,000$. This could be attributed to the relatively poor fit of the normal approximation whe population size is low (see Appendix A-4.1). In the bottom panel, we see that both estimators tend to underestimate the population size. Here, the NLS estimator has much lower bias compared to the ML estimator. In relative terms, the bias is the strongest when true population size is low.

Similarly, Figure 8 plots average bias of $\hat{\theta}^{\text{NLS}}_B$ and $\hat{\theta}^{\text{MLE}}_B$ given the numerical approximation of final allele frequencies[13]. The results suggest that the NLS estimator $\hat{\omega}^{\text{NLS}}_B$ performs well for any population size N, whereas the ML estimator tends to overestimate the selection parameter when N is low. Similarly, the NLS estimator of the population size demonstrates very small bias, especially compared to larger bias in the ML estimator of the population size.

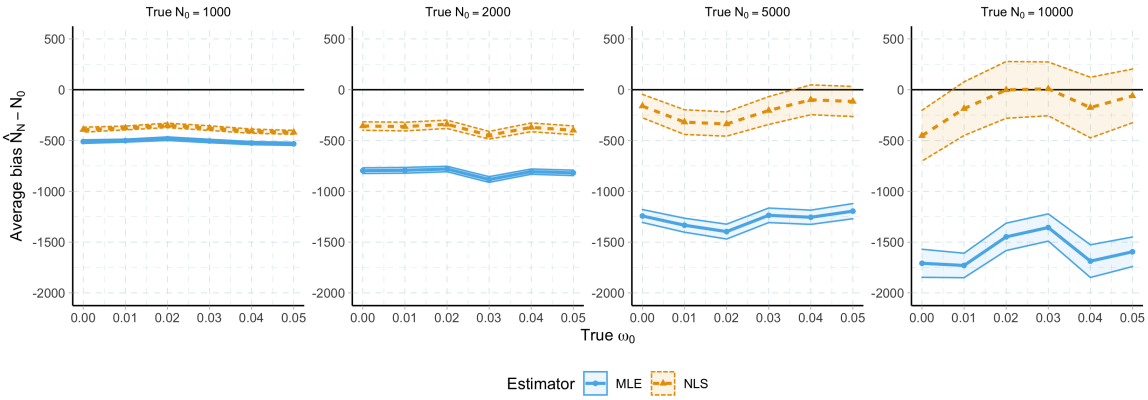
Thus, on average the estimators based on numerical approximation demonstrate lower bias than the ones based on normal approximation. This could be attributed to the better

12. In our analysis, we use $B = 200$.

13. We use grid-search algorithm to find estimates under numerical approximation. That is, we evaluate the objective function at every parameter θ in a grid and find the value associated with either the lowest squared distance from conditional mean or highest log-likelihood.



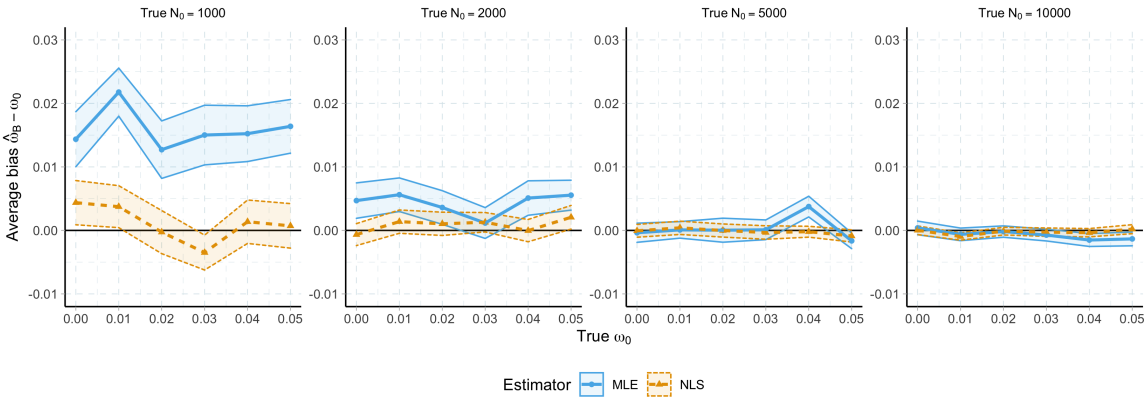
(a) Selection parameter ω



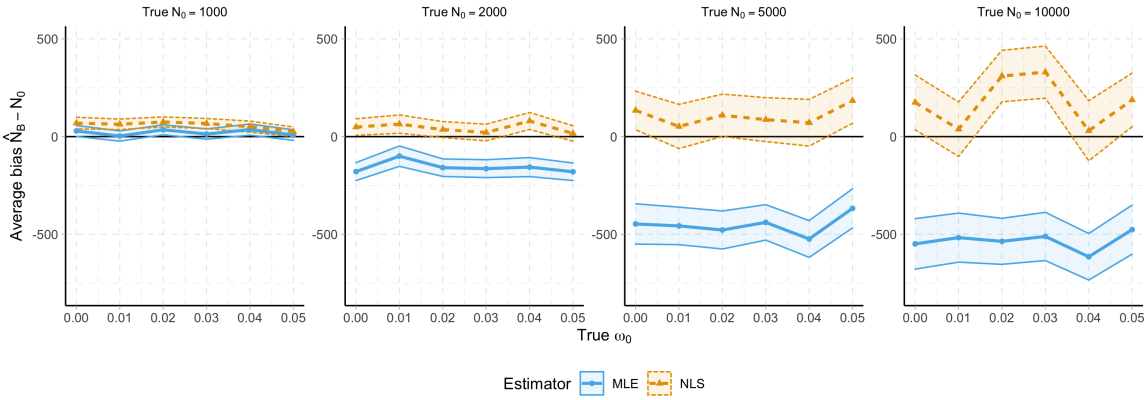
(b) Population size N

Figure 7: Average bias of estimators based on normal approximation

Notes: the figure plots average bias of estimates based on normal approximation obtained in synthetic datasets at each combination of true parameters.



(a) Selection parameter ω



(b) Population size N

Figure 8: Average bias of estimator based on numerical approximation

Notes: the figure plots average bias of estimates based on numerical approximation obtained in synthetic datasets at each combination of true parameters.

fit of numerical approximation for the true distribution of allele frequencies Ψ.

B.4.3 Estimates in the data

We now turn to the parameter estimations in the analysis sample (Table 1). The main results are reported in the first two columns: NLS and ML estimates based on truncated normal approximation of the simple model.

Table 1: Estimates of the simple model based on truncated normal approximation

	Full sample			Excl. bounds		
	NLS	MLE	NLS wt	NLS	MLE	NLS wt
$\hat{\omega}_{\mathcal{N}}$	0.041	0.152	0.042	0.040	0.105	0.042
	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)
$\hat{N}_{\mathcal{N}}$	2553.6	3.5	2721.8	2563.6	1804.1	2727.0
	(343.5)	(0.2)	(384.9)	(327.5)	(142.6)	(374.1)
Obs.	475	475	475	440	440	440
LR test of : $\omega = 0$						
χ^2_1 stat	88899.7	421.1	49001.7	14.7	231.9	16.0
χ^2_1 p-value	0.0000	0.0000	0.0000	0.0001	0.0000	0.0001

Note:
The table reports estimation results for the parameters in the diffusion process. Conventional standard errors are reported in parentheses. The third column $NLSwt$ reports the estimates with weighted SNP's as in equation (34).

The NLS estimate suggests that the education-enhancing alleles were subject to positive selection with a parameter $\omega = 0.041$. The estimate is statistically significant according to both conventional t-test and likelihood ratio test. The NLS also estimates the population size at about 2,500 individuals. The ML estimator returns considerably higher estimate of ω and much lower estimate of the population size. This is in line with the results reported in the previous section, which show that ML estimator tends to overestimate ω and underestimate N. But the ML estimates are further biased due to the characteristics of the dataset. The analysis sample includes SNPs with allele frequencies in WHG population very close to the boundaries since we did not impose any restrictions on the initial values. The ML estimator favours smaller N because it allows the allele frequencies to jump by a larger step at any point in time. Thus, smaller N increases the chance of observing these SNPs in the modern sample. At the same time, this can increase the distance between the implied and observed final allele frequencies. To compensate for larger distance, the ML estimator also increases ω.

For this reason, we repeat the estimations in the subsample without the SNPs with initial allele frequencies closest to the boundaries[14]. These results are reported in the last

14. The initial allele frequencies in the full sample range between 0.00001 and 0.99999, with 25 SNPs

two columns of Table 1. The NLS point estimates, both simple and weighted, remain unchanged. The ML estimate of ω is slightly lower than in the full sample, and the ML estimate of N is now much closer to the NLS estimate.

We are mainly interested in whether the parameter ω is non-zero, which would imply that the education-enhancing alleles were subject to selection. The conventional t-statistic suggests that all points estimates reported in Table 1 are statistically different from zero. In addition, we perform the likelihood ratio test reported at the bottom of the table, which strongly rejects the null hypothesis of $\omega = 0$. Thus, the results in Table 1 suggest that there was a strong selective push favouring education-enhancing alleles over the past 14,000 years.

below 0.01 and 10 SNPs - above 0.99.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526, no. 7571 (2015): 68–74.

Aimé, C., Laval, G., Patin, E., et al. 2013. "Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming." *Molecular biology and evolution* 30 (12): 2629–2644.

Alexander, D. H. and Lange, K. 2011. "Enhancements to the ADMIXTURE Algorithm for Individual Ancestry Estimation." *BMC bioinformatics* 12 (2011): 246.

Ashraf, Q. and Galor, O. 2011. "Dynamics and stagnation in the Malthusian epoch." *American Economic Review* 101 (5): 2003–2041.

Baldini, L. M., Baldini, J. U. L., McDermott, F., et al. 2019. "North Iberian Temperature and Rainfall Seasonality over the Younger Dryas and Holocene." *Quaternary Science Reviews* 226 (2019): 105998.

Barth, D., Papageorge, N. W., and Thom, K. 2020. "Genetic endowments and wealth inequality." *Journal of Political Economy* 128 (4): 1474–1522.

Berg, J. J. and Coop, G. 2014. "A Population Genetic Signal of Polygenic Adaptation." Edited by M. W. Feldman. *PLoS Genetics* 10, no. 8 (2014): e1004412.

Bocquentin, F., Anton, M., Berna, F., et al. 2020. "Emergence of Corpse Cremation during the Pre-Pottery Neolithic of the Southern Levant: A Multidisciplinary Study of a Pyre-Pit Burial." *PLOS ONE* 15, no. 8 (2020): e0235386.

Bocquet-Appel, J.-P. 2008. "Explaining the Neolithic Demographic Transition." *The Neolithic Demographic Transition and its Consequences,* 35–55.

Bogaard, A., Charles, M., Twiss, K. C., et al. 2009. "Private Pantries and Celebrated Surplus: Storing and Sharing Food at Neolithic Çatalhöyük, Central Anatolia." *Antiquity* (Cambridge).

Bollongino, R., Nehlich, O., Richards, M. P., et al. 2013. "2000 years of parallel societies in Stone Age Central Europe." *Science* 342 (6157): 479–481.

Bowles, S. 2011a. "Cultivation of Cereals by the First Farmers Was Not More Productive than Foraging." *Proceedings of the National Academy of Sciences* 108, no. 12 (2011): 4760–4765.

———. 2011b. "Cultivation of cereals by the first farmers was not more productive than foraging." *Proceedings of the National Academy of Sciences* 108 (12): 4760–4765.

Bowles, S. and Choi, J.-K. 2013a. "Coevolution of farming and private property during the early Holocene." *Proceedings of the National Academy of Sciences* 110 (22): 8830–8835.

———. 2013b. "Coevolution of Farming and Private Property during the Early Holocene." *Proceedings of the National Academy of Sciences* 110, no. 22 (2013): 8830–8835.

Bramanti, B., Thomas, M. G., Haak, W., et al. 2009. "Genetic discontinuity between local hunter-gatherers and central Europe's first farmers." *Science* 326 (5949): 137–140.

Campbell, C. D., Chong, J. X., Malig, M., et al. 2012. "Estimating the Human Mutation Rate Using Autozygosity in a Founder Population." *Nature Genetics* 44, no. 11 (11): 1277–1281.

Charlesworth, B. 2013. "Stabilizing selection, purifying selection, and mutational bias in finite populations." *Genetics* 194 (4): 955–971.

Charlton, S., Booth, T., and Barnes, I. 2019. "The Problem with Petrous? A Consideration of the Potential Biases in the Utilization of Pars Petrosa for Ancient DNA Analysis." *World Archaeology* 51, no. 4 (2019): 574–585.

Cox, S. L., Moots, H. M., Stock, J. T., et al. 2022. "Predicting skeletal stature using ancient DNA." *American Journal of Biological Anthropology* 177 (1): 162–174.

Cox, S. L., Ruff, C. B., Maier, R. M., et al. 2019a. "Genetic contributions to variation in human stature in prehistoric Europe." *Proceedings of the National Academy of Sciences* 116 (43): 21484–21492.

———. 2019b. "Genetic Contributions to Variation in Human Stature in Prehistoric Europe." *Proceedings of the National Academy of Sciences* 116, no. 43 (2019): 21484–21492.

Crosby, J. L. et al. 1970. "The evolution of genetic discontinuity: computer models of the selection of barriers to interbreeding between subspecies." *Heredity* 25:253–97.

Crow, J. F. and Kimura, M. 1970. *An introduction to Population Genetics Theory.* Harper / Row.

Darwin, C. 1868. *The variation of animals and plants under domestication.* Vol. 2. J. murray.

David Reich Lab. 2021. "Allen Ancient DNA Resource (AADR): Downloadable Genotypes of Present-Day and Ancient DNA Data (v.44.3)." David Reich Lab, 2021. Accessed March 2, 2021. https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data.

Edge, M. D. and Coop, G. 2019. "Reconstructing the History of Polygenic Scores Using Coalescent Trees." *Genetics* 211, no. 1 (2019): 235–262.

Feynman, J. and Ruzmaikin, A. 2007. "Climate Stability and the Development of Agricultural Societies." *Climatic Change* 84, no. 3 (2007): 295–311.

Gallagher, E. M., Shennan, S. J., and Thomas, M. G. 2015. "Transition to Farming More Likely for Small, Conservative Groups with Property Rights, but Increased Productivity Is Not Essential." *Proceedings of the National Academy of Sciences* 112, no. 46 (2015): 14218–14223.

Galor, O. and Moav, O. 2000. "Ability-biased technological transition, wage inequality, and economic growth." *The quarterly journal of economics* 115 (2): 469–497.

———. 2002. "Natural selection and the origin of economic growth." *The Quarterly Journal of Economics* 117 (4): 1133–1191.

Gibbs, R. A. 2020. "The human genome project changed everything." *Nature Reviews Genetics* 21 (10): 575–576.

González-Fortes, G., Jones, E. R., Lightfoot, E., et al. 2017. "Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin." *Current Biology* 27 (12): 1801–1810.e10.

Guo, J., Yang, J., and Visscher, P. M. 2018. "Leveraging GWAS for complex traits to detect signatures of natural selection in humans." *Current opinion in genetics & development* 53:9–14.

Haldane, J. 1919. "The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series." *Journal of Genetics* 8 (4): 291–297.

Hawkes, K., Altman, J., Beckerman, S., et al. 1993. "Why hunter-gatherers work: An ancient version of the problem of public goods [and comments and reply]." *Current anthropology* 34 (4): 341–361.

Helle, S., Brommer, J. E., Pettay, J. E., et al. 2014. "Evolutionary Demography of Agricultural Expansion in Preindustrial Northern Finland." *Proceedings of the Royal Society B: Biological Sciences* 281, no. 1794 (2014): 20141559.

Ju, D. and Mathieson, I. 2021. "The Evolution of Skin Pigmentation-Associated Variation in West Eurasia." *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 1 (2021).

Kingsolver, J. G. and Pfennig, D. W. 2007. "Patterns and Power of Phenotypic Selection in Nature." *BioScience* 57, no. 7 (2007): 561–572.

Kramer, K. L. and Greaves, R. D. 2007. "Changing Patterns of Infant Mortality and Maternal Fertility among Pumé Foragers and Horticulturalists." *American Anthropologist* 109 (4): 713–726.

Kuijt, I. 2008. "Demography and Storage Systems During the Southern Levantine Neolithic Demographic Transition." In *The Neolithic Demographic Transition and Its Consequences,* edited by J.-P. Bocquet-Appel and O. Bar-Yosef, 287–313. Dordrecht: Springer Netherlands.

Kuijt, I. and Finlayson, B. 2009. "Evidence for Food Storage and Predomestication Granaries 11,000 Years Ago in the Jordan Valley." *Proceedings of the National Academy of Sciences* 106, no. 27 (2009): 10966–10970.

Larson, G. and Burger, J. 2013. "A population genetics view of animal domestication." *Trends in Genetics* 29 (4): 197–205.

Lee, J. J., Wedow, R., Okbay, A., et al. 2018. "Gene Discovery and Polygenic Prediction from a Genome-Wide Association Study of Educational Attainment in 1.1 Million Individuals." *Nature Genetics* 50, no. 8 (8): 1112–1121.

Lichten, M. and Goldman, A. S. 1995. "Meiotic recombination hotspots." *Annual review of genetics* 29 (1): 423–444.

Lord, K. A., Larson, G., Coppinger, R. P., et al. 2020. "The history of farm foxes undermines the animal domestication syndrome." *Trends in ecology & evolution* 35 (2): 125–136.

Mathieson, I. 2021. "The omnigenic model and polygenic prediction of complex traits." *The American Journal of Human Genetics* 108 (9): 1558–1563.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., et al. 2018. "The Genomic History of Southeastern Europe." *Nature* 555, no. 7695 (2018): 197–203.

Mathieson, S. and Mathieson, I. 2018. "FADS1 and the Timing of Human Adaptation to Agriculture." *Molecular Biology and Evolution* 35, no. 12 (2018): 2957–2970.

Matranga, A. 2019. "The Ant and the Grasshopper: Seasonality and the Invention of Agriculture." Review and Resubmit at Quarterly Journal of Economics, 2019.

McCorriston, J. and Hole, F. 1991. "The Ecology of Seasonal Stress and the Origins of Agriculture in the Near East." *American Anthropologist* 93 (1): 46–69.

Myers, S., Bottolo, L., Freeman, C., et al. 2005. "A fine-scale map of recombination rates and hotspots across the human genome." *Science* 310 (5746): 321–324.

Nagylaki, T. 2013. *Introduction to theoretical population genetics.* Vol. 21. Springer Science & Business Media.

Okbay, A., Baselmans, B. M. L., De Neve, J.-E., et al. 2016. "Genetic Variants Associated with Subjective Well-Being, Depressive Symptoms, and Neuroticism Identified through Genome-Wide Analyses." *Nature Genetics* 48, no. 6 (6): 624–633.

Özdoğan, M. 2011. "Archaeological Evidence on the Westward Expansion of Farming Communities from Eastern Anatolia to the Aegean and the Balkans." *Current Anthropology* 52 (S4): S415–S430.

Piperno, D. R., Weiss, E., Holst, I., et al. 2004. "Processing of Wild Cereal Grains in the Upper Palaeolithic Revealed by Starch Grain Analysis." *Nature* 430, no. 7000 (7000): 670–673.

Racimo, F., Berg, J. J., and Pickrell, J. K. 2018. "Detecting polygenic adaptation in admixture graphs." *Genetics* 208 (4): 1565–1584.

Robinson, S. A., Black, S., Sellwood, B. W., et al. 2006. "A Review of Palaeoclimates and Palaeoenvironments in the Levant and Eastern Mediterranean from 25,000 to 5000 Years BP: Setting the Environmental Background for the Evolution of Human Civilisation." *Quaternary Science Reviews* 25, no. 13 (2006): 1517–1541.

Robson, A. J. 2010. "A bioeconomic view of the Neolithic transition to agriculture." *Canadian Journal of Economics/Revue canadienne d'économique* 43 (1): 280–300.

Rossignol-Strick, M. 1995. "Sea-Land Correlation of Pollen Records in the Eastern Mediterranean for the Glacial-Interglacial Transition: Biostratigraphy versus Radiometric Time-Scale." *Quaternary Science Reviews* 14, no. 9 (1995): 893–915.

Rowthorn, R. 2011. "A bioeconomic view of the transition to agriculture: a comment." *Canadian Journal of Economics/Revue canadienne d'économique* 44 (3): 1044–1047.

Rowthorn, R. and Seabright, P. 2010. "Property rights, warfare and the neolithic transition." *TSE Working Paper.*

Scally, A. and Durbin, R. 2012. "Revising the Human Mutation Rate: Implications for Understanding Human Evolution." *Nature Reviews Genetics* 13, no. 10 (10): 745–753.

Scheu, A., Powell, A., Bollongino, R., et al. 2015. "The Genetic Prehistory of Domesticated Cattle from Their Origin to the Spread across Europe." *BMC Genetics* 16 (1): 54.

Shendure, J. and Akey, J. M. 2015. "The Origins, Determinants, and Consequences of Human Mutations." *Science* 349, no. 6255 (2015): 1478–1483.

Shennan, S. 2018. *The First Farmers of Europe: An Evolutionary Perspective.* Cambridge World Archaeology. Cambridge: Cambridge University Press.

Smith, E. A. 2004. "Why do good hunters have higher reproductive success?" *Human Nature* 15 (4): 343–364.

Smith, E. A. and Codding, B. F. 2021. "Ecological variation and institutionalized inequality in hunter-gatherer societies." *Proceedings of the National Academy of Sciences* 118 (13): e2016134118.

Smith, E. A., Hill, K., Marlowe, F. W., et al. 2010. "Wealth transmission and inequality among hunter-gatherers." *Current Anthropology* 51 (1): 19–34.

Song, W., Shi, Y., Wang, W., et al. 2021. "A selection pressure landscape for 870 human polygenic traits." *Nature Human Behaviour* 5 (12): 1731–1743.

Speidel, L., Forest, M., Shi, S., et al. 2019. "A Method for Genome-Wide Genealogy Estimation for Thousands of Samples." *Nature Genetics* 51, no. 9 (9): 1321–1329.

Spolaore, E. and Wacziarg, R. 2013. "How deep are the roots of economic development?" *Journal of economic literature* 51 (2): 325–369.

Stern, A. J., Speidel, L., Zaitlen, N. A., et al. 2021. "Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies." *The American Journal of Human Genetics* 108 (2): 219–239.

Stuiver, M., Grootes, P. M., and Braziunas, T. F. 1995. "The GISP2 δ18O Climate Record of the Past 16,500 Years and the Role of the Sun, Ocean, and Volcanoes." *Quaternary Research* 44 (3): 341–354.

Testart, A., Arcand, B., Ingold, T., et al. 1988. "Some major problems in the social anthropology of hunter-gatherers [and Comments and Reply]." *Current Anthropology* 29 (1): 1–31.

Tian, X., Browning, B. L., and Browning, S. R. 2019. "Estimating the Genome-wide Mutation Rate with Three-Way Identity by Descent." *The American Journal of Human Genetics* 105, no. 5 (2019): 883–893.

Trut, L., Oskina, I., and Kharlamova, A. 2009. "Animal evolution during domestication: the domesticated fox as a model." *Bioessays* 31 (3): 349–360.

Uricchio, L. H. 2020. "Evolutionary perspectives on polygenic selection, missing heritability, and GWAS." *Human genetics* 139 (1): 5–21.

Verdu, P., Becker, N. S., Froment, A., et al. 2013. "Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies." *Molecular biology and evolution* 30 (4): 918–937.

Wilde, S., Timpson, A., Kirsanow, K., et al. 2014. "Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y." *Proceedings of the National Academy of Sciences* 111 (13): 4832–4837.

Winterhalder, B. and Goland, C. 1993. "On Population, Foraging Efficiency, and Plant Domestication." *Current Anthropology* 34 (5): 710–715.

Yair, S. and Coop, G. 2022. "Population differentiation of polygenic score predictions under stabilizing selection." *Philosophical Transactions of the Royal Society B* 377 (1852): 20200416.

Zeng, K. and Charlesworth, B. 2009. "Estimating selection intensity on synonymous codon usage in a nonequilibrium population." *Genetics* 183 (2): 651–662.

Selection and the Roy Model in the Neolithic Transition

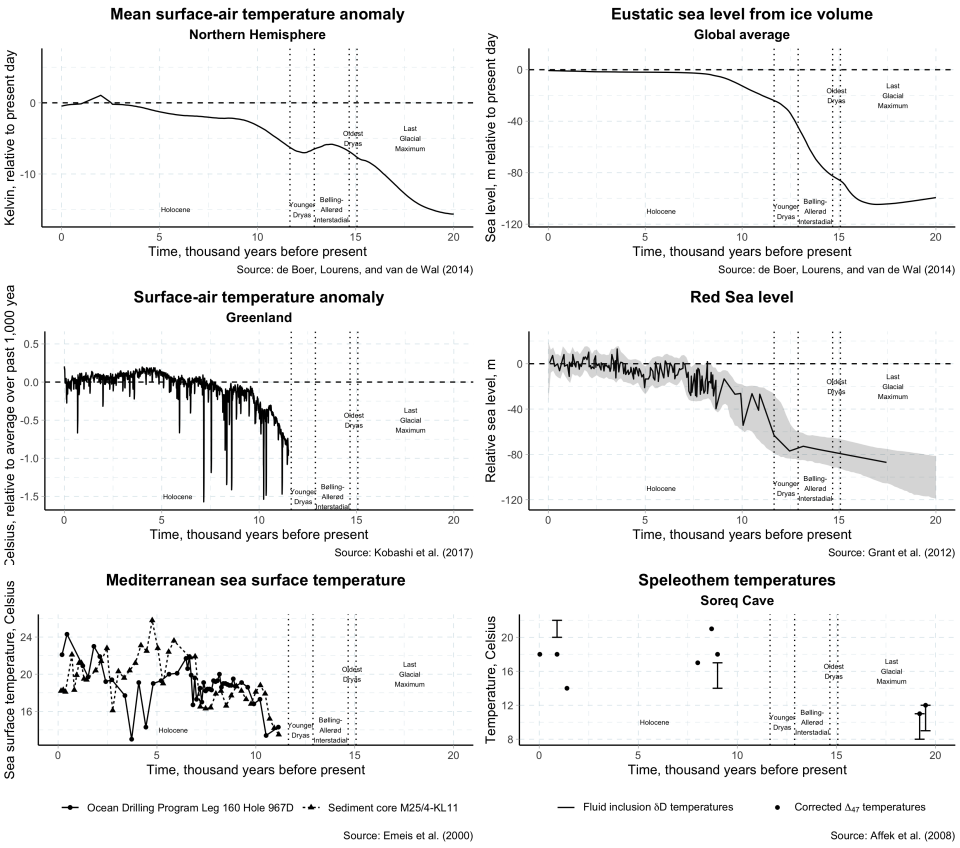
Online Appendix

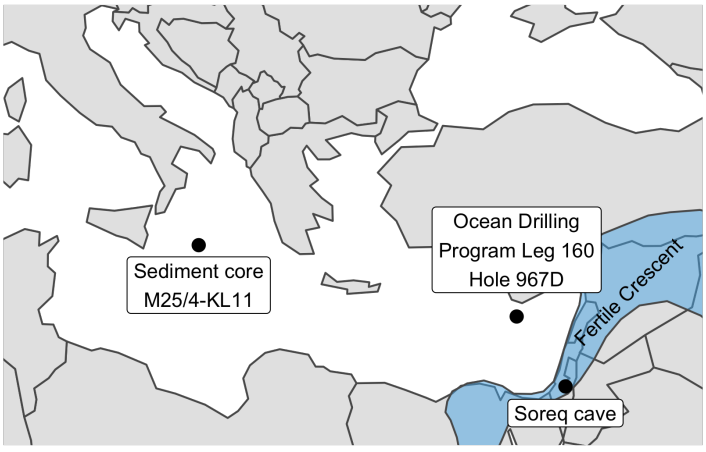
(Not meant to be part of the journal publication.)

Aldo Rustichini, Nurfatima Jandarova

October 4, 2024

A-1 Evidence on Climate Change





Sources: Emeis et al. (2000) and Bar-Matthews et al. (2003)

Figure A-1: Climate reconstructions over the past 20,000 years

Notes: temperature anomaly is the deviation of reconstructed temperature from reference or long-term average. Eustatic sea level is the distance from the centre of the Earth to the sea surface. Relative sea level is the sea level relative to a land-based reference point. For more details refer to the sources indicated at the bottom right corner of each plot. The climate periods are taken from Stuiver, Grootes, and Braziunas (1995).

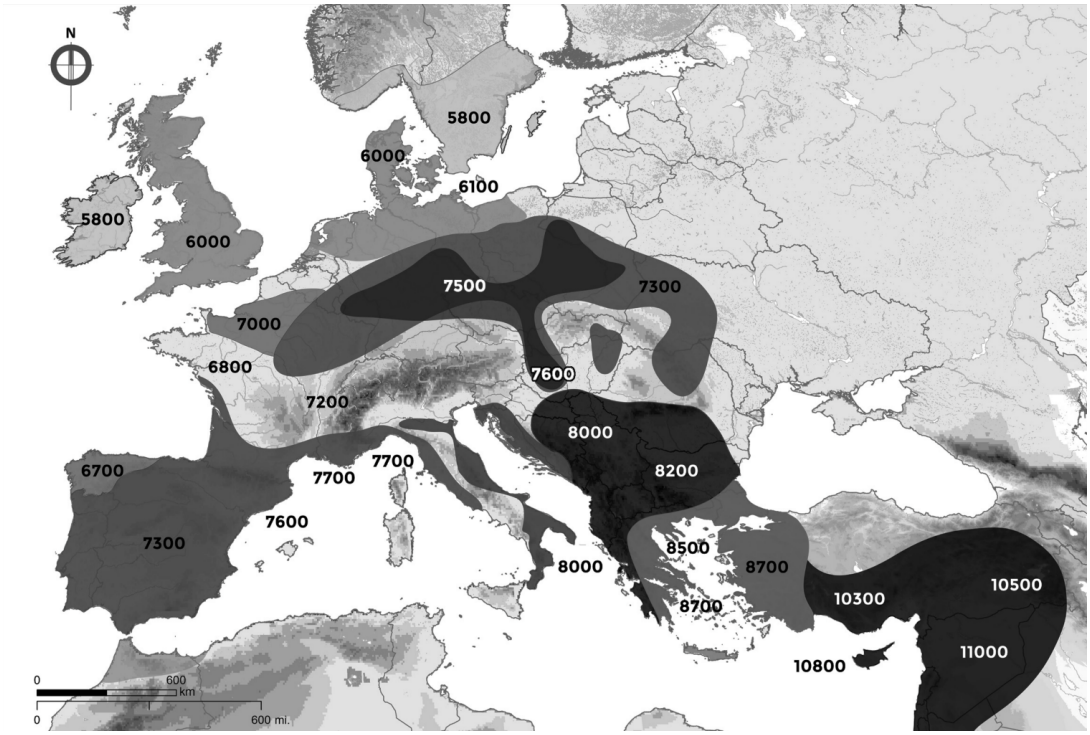


Figure A-2: Farming spread

Note: Reprinted Fig 1.1 from Shennan (2018). Dates are shown in years before present.

A-2 Distribution of Ancestries

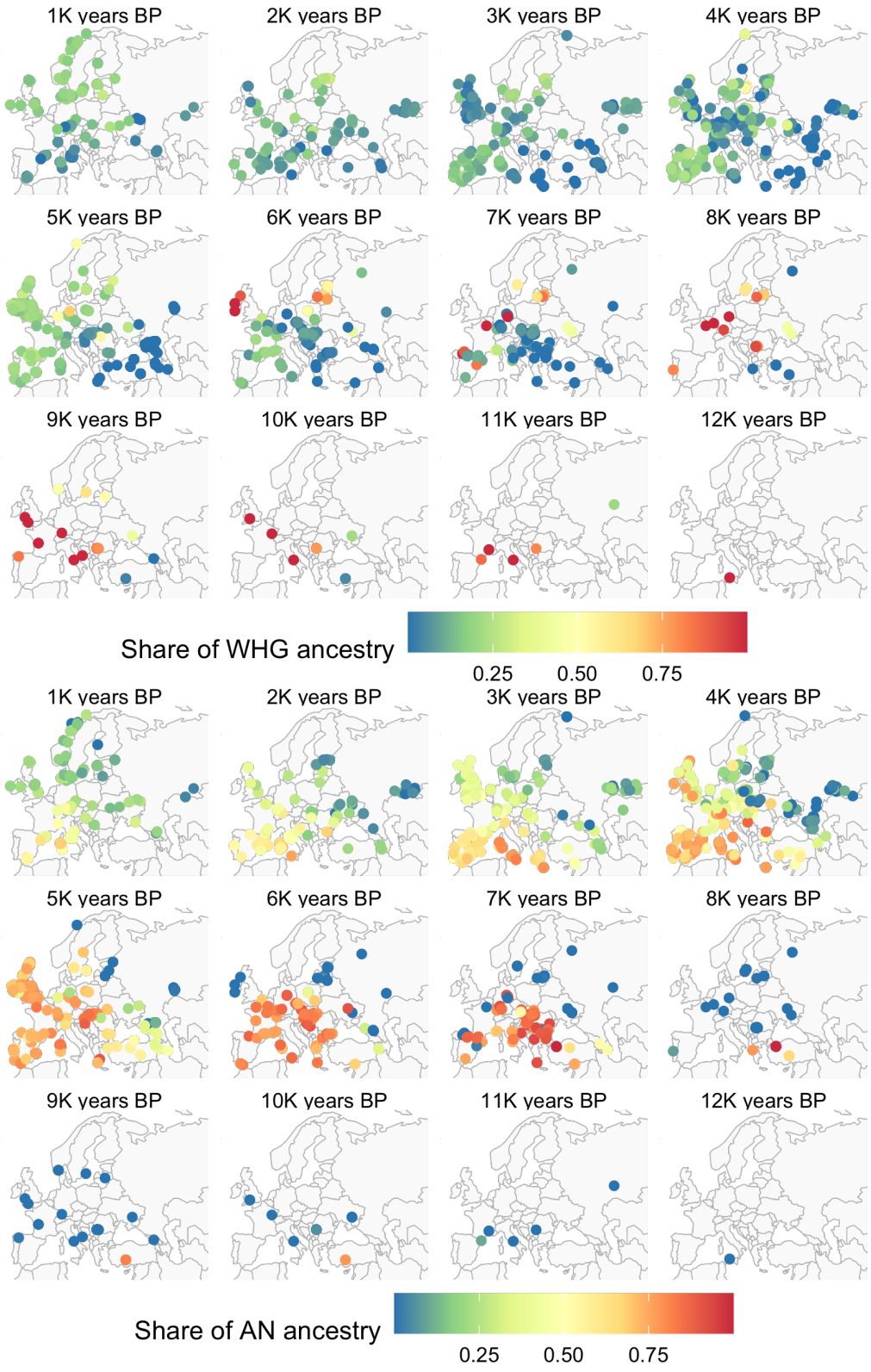


Figure A-3: Distribution ancestries over time and space

Note: the figure plots location of ancient individuals in sample over time. The colour of points is based on shares of WHG and AN ancestries in genotypes of each sample member computed using supervised ADMIXTURE algorithm.

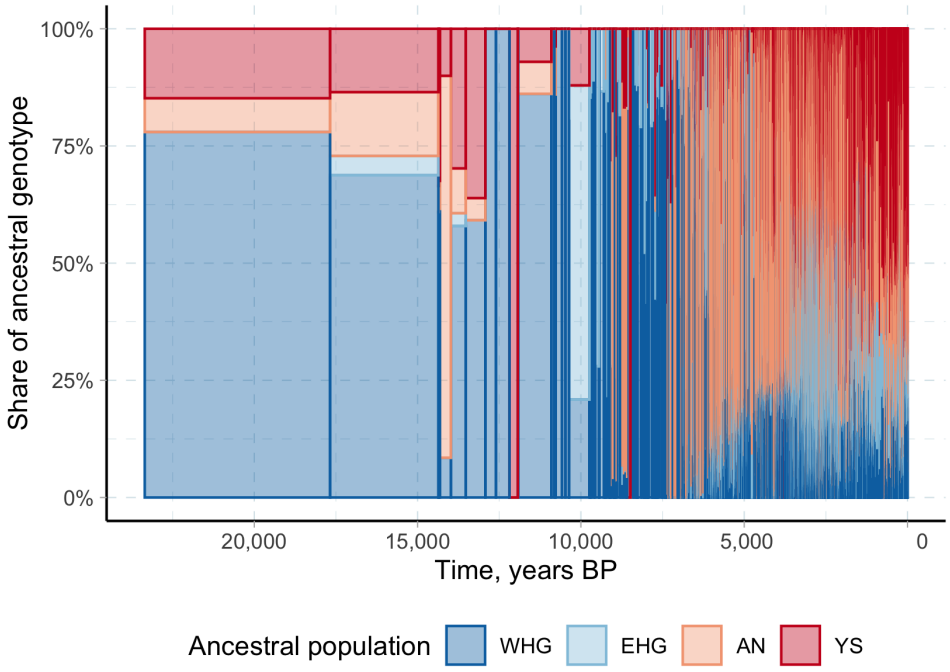


Figure A-4: Ancestral shares over time

Note: the figure plots shares of ancestral genotypes in each ancient individual in the working AADR dataset estimated using the supervised ADMIXTURE algorithm. The width of bars is inversely proportional to sample size available at that period of time.

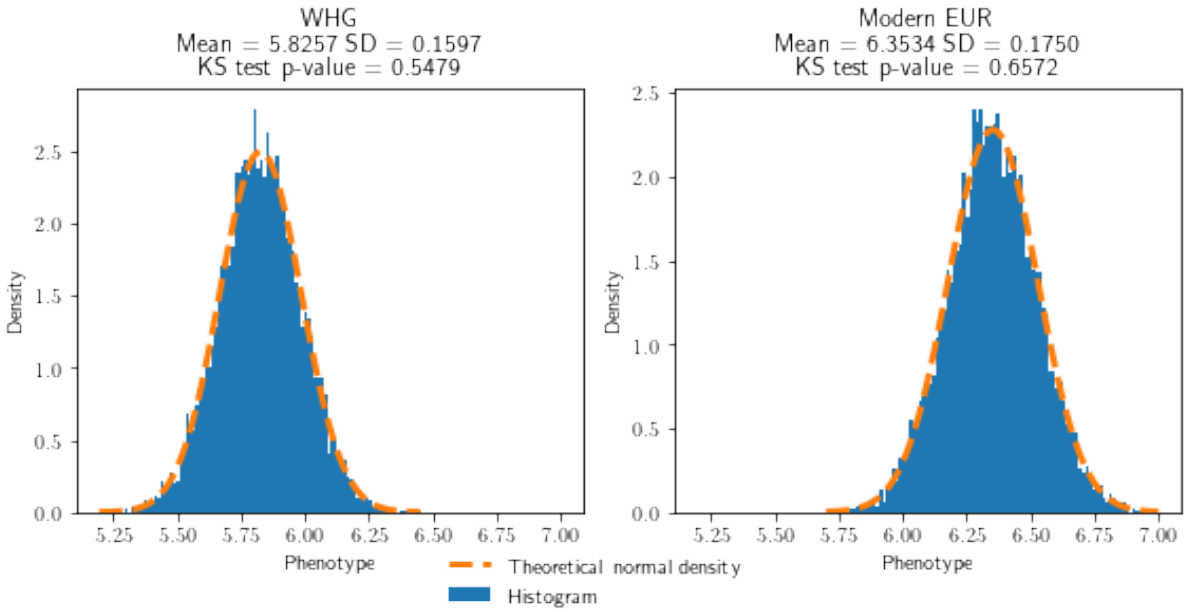


Figure A-5: Distribution of phenotypes implied by allele frequencies in the data

A-3 Normal approximation to phenotype distribution

In this section we show that phenotype distribution can be approximated by a normal distribution. Recall that we assume phenotypes to be linear functions of genotypes as shown in Equation (5). Therefore, the exact distirbution of the phenotypes stems from the distribution of genotypes in the population. We argue that unless GWAS weights decline too fast, the exact distribution of phenotypes is well approximated by a normal distribution.

We use the allele frequency estimates among modern EUR and WHG populations to compute the distribution of genotypes according to Hardy-Weinberg equation. That is, if an allele frequency at locus k is $p(k)$, then

$$\begin{aligned}\Pr(g(k)=0) &= (1-p(k))^2 \\ \Pr(g(k)=1) &= 2p(k)(1-p(k)) \\ \Pr(g(k)=2) &= p(k)^2\end{aligned}$$

Using these genotype distributions we randomly draw genotypes of a population of size $N = 10,000$ and compute their polygenic scores. We compare the resulting histogram of polygenic scores to the theoretical normal density function in Figure A-5. Furthermore, we perform Kolmogorov-Smirnov tests that also fail to reject the null hypothesis of normality. Thus, the phenotype distributions implied by allele frequencies of modern EUR and WHG populations are indeed well approximated by normal distribution.

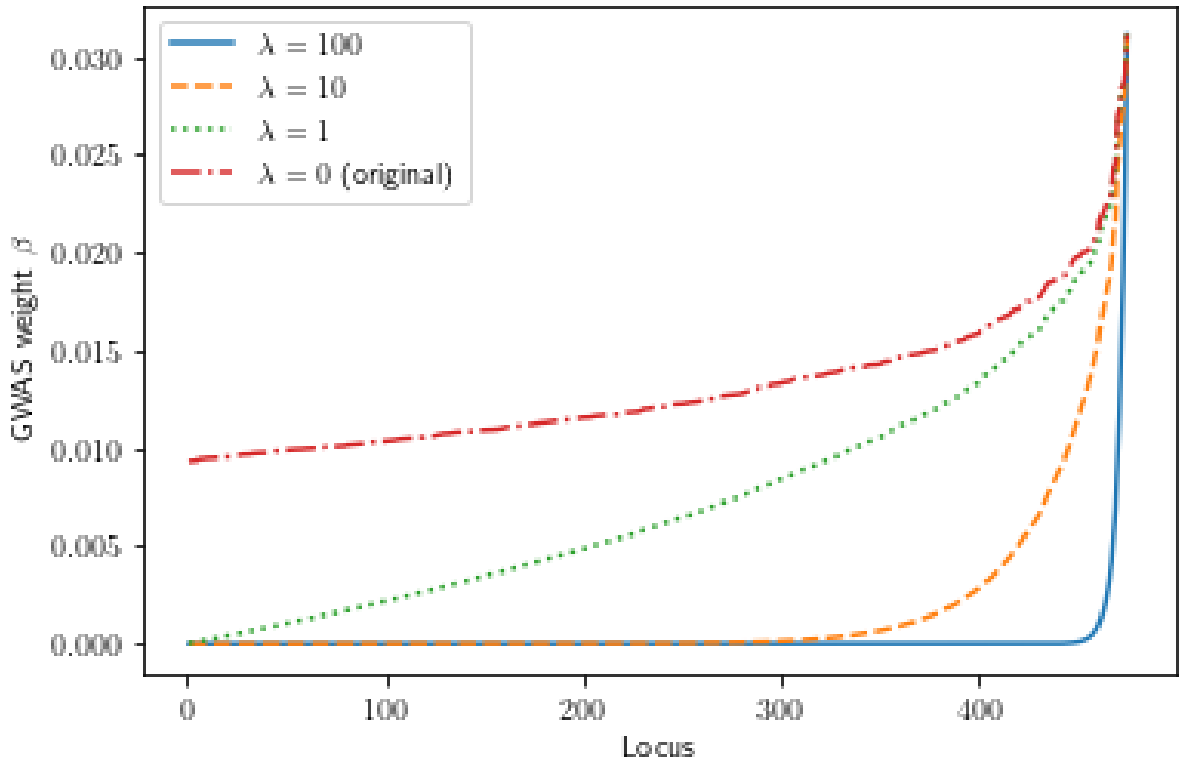


Figure A-6: GWAS decay

Next, we show that normal approximation to phenotype distribution works well unless the speed of decay of GWAS weights is extremely fast. We modify the vector of GWAS weights in the following way

$$\tilde{\beta}(k) = \beta(k) \left(\frac{k}{K} \right)^{\lambda} \tag{A-1}$$

The speed of decay is controlled by parameter λ: higher values mean faster decline. Figure A-6 plots the adjusted GWAS weights for various values of λ. For example, at $\lambda = 100$ only 25 SNPs out of 475 have GWAS weights above 10^{-4}.

Then, using the adjusted GWAS weights we plot the distribution of phenotypes (normalized to have mean one) and test for normality using Kolmogorov-Smirnov test. The results are plotted in Figure A-7 and suggest that the normal approximation starts failing at very extreme values of λ. Even at $\lambda = 10$ where barely a quarter of SNPs have GWAS weights above 10^{-4}, the phenotype is described well by a normal distribution.

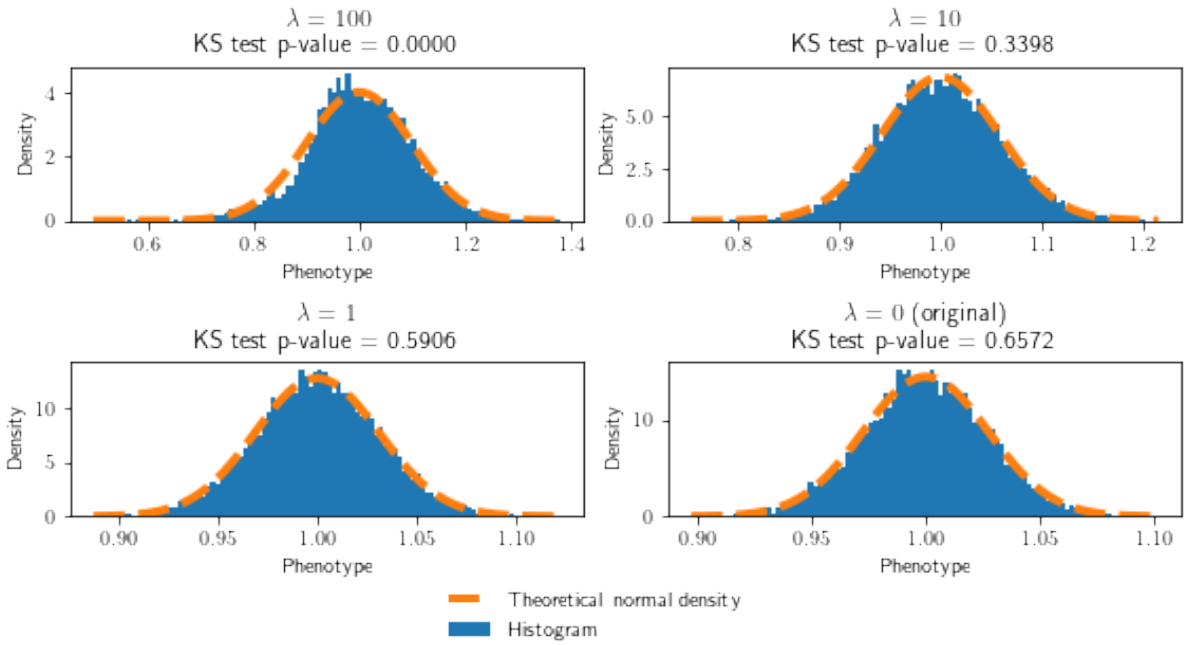


Figure A-7: Distribution of phenotypes by speed of decay of GWAS weights

A-4 Approximations to the final allele frequency distributions

The estimators proposed in Section 7 are based on the distribution of final allele frequencies Ψ from the diffusion process in Equation (31). We do not have analytical solution to the true distribution Ψ and rely on approximations $\hat{\Psi}$. In particular, we consider two approximations: normal and numerical. In this section, we describe in detail what these approximations imply for the NLS and ML estimators.

A-4.1 Truncated normal approximation

If Equation (31) were deterministic, the integral over time would yield

$$\mu(k,T) \equiv x(k,T) = \frac{x(k,0)\exp(\beta(k)\omega T)}{x(k,0)\exp(\beta(k)\omega T) + 1 - x(k,0)} \tag{A-2}$$

Thus, we conjecture that the diffusion process in Equation (31) can be approximated with the following regression equation

$$\begin{aligned} x(k,T) = \mu(k,T) + \varepsilon, \qquad \varepsilon \sim \text{Censored } \mathcal{N}\left(0, T\mu(k,T)(1-\mu(k,T))\right) \\ \text{s.t.} \quad x(k,T) \in [0,1] \\ x(k,0), \beta(k) \quad \text{given} \end{aligned} \tag{A-3}$$

That is, the error term is distributed normally and censored such that the final allele frequencies are bounded between 0 and 1.

We assess how well the censored normal approximation in Equation (A-3) fits the diffusion process in Equation (31) using the following simulation exercise. We consider a single SNP and set its initial conditions to $x(k,0) = 0.2$ and $\beta(k) = 0.0131$. Then, we simulate the diffusion process in Equation (31) multiple times given a pair of parameters (ω, N). Finally, we compare the empirical distribution of the final allele frequencies with those implied by the censored normal approximation in Figure A-8. The visual inspection suggests that the approximation works well as long as few SNPs hit the fixation boundaries (left panel). Variance of allele frequency paths is inversely related to the population size. Therefore, the smaller the population size, the faster the SNP reaches fixation. This can be seen in the right panel of Figure A-8. With population size $N = 1,000$, larger share of paths end at fixation boundaries. In this case, the censored normal approximation seems to be a poor fit both visually and statistically. On the right panel, the distribution of the final allele frequencies when $N = 10,000$ is mostly located at the

interior points. Visually, the censored normal approximation seems to be an adequate fit, although the Kolmogorov-Smirnov test strongly rejects the normality.

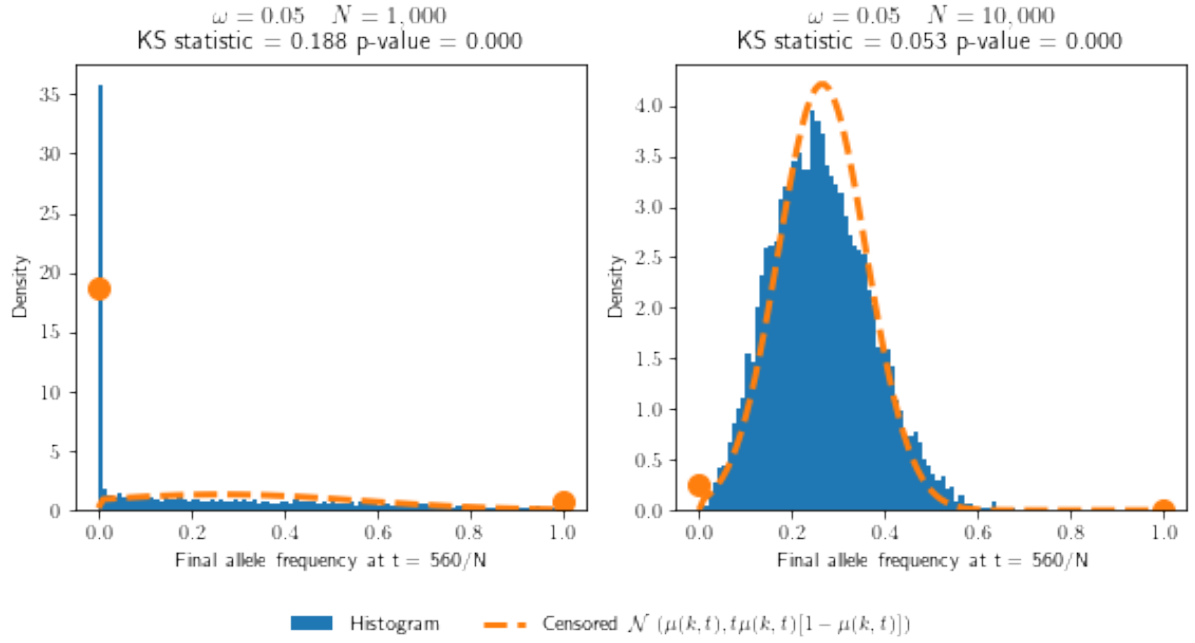


Figure A-8: Distribution of allele frequencies at $t = \frac{560}{N}$

Note: The figure plots the empirical distribution of the simulated diffusion process in Equation (31) together with the theoretical distribution implied by the censored normal approximation in Equation (A-3). The initial conditions of the simulations are given by $x(k,0) = 0.2$ and $\beta(k) = 0.0131$. The empirical distribution is based on sampling 10,000 paths from the diffusion process.

The censored normal approximation assumes that we observe a random sample of SNPs. In reality, our sample was selected based on two characteristics: the alleles are not fixated in modern European populations and have statistically significant effect sizes estimated in a GWAS of 766,345 individuals. Denote the set of non-fixated SNPs at time t as $\mathbf{V}_t(\nu) \equiv \{k : \nu \leq x(k,t) \leq 1-\nu\}$. It can also be shown that the set of statistically-significant SNPs[15] at time t given significance level α can be written as $\mathbf{S}_t(\alpha) \equiv \{k : \frac{1-\kappa(k)}{2} \leq x(k,t) \leq \frac{1+\kappa(k)}{2}\}$ where $\kappa(k) = \sqrt{1 - \frac{2c^2}{\beta(k)^2(n+c^2)}}$, n is the GWAS sample size and $c \equiv \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. The data processing steps imply fixation threshold $\nu = 0.01$ and significance level $\alpha = 5 \times 10^{-8}$. The approximation to the simple model taking into account the sample selection criteria can be written using the truncated normal distribution

15. Assuming GWAS effect sizes stay constant over time, or equivalently, linear allele effect across genotypes $g(k) = \{0, 1, 2\}$.

$$x(k,T) = \mu(k,T) + \varepsilon, \qquad \varepsilon \sim \text{Truncated } \mathcal{N}\left(0, T\mu(k,T)(1-\mu(k,T))\right) \qquad \text{(A-4)}$$
$$\text{s.t.} \quad k \in \mathbf{V}_T(\nu) \cap \mathbf{S}_T(\alpha)$$
$$x(k,0), \beta(k) \quad \text{given}$$

Figure A-9 compares the empirical distribution of allele frequencies satisfying the sample selection criteria with the theoretical distribution implied by the truncated normal approximation in Equation (A-4). This time, the visual inspection suggests that in both cases the truncated normal approximation matches the distribution relatively closely, even though the Kolmogorov-Smirnov tests strongly reject the normality.

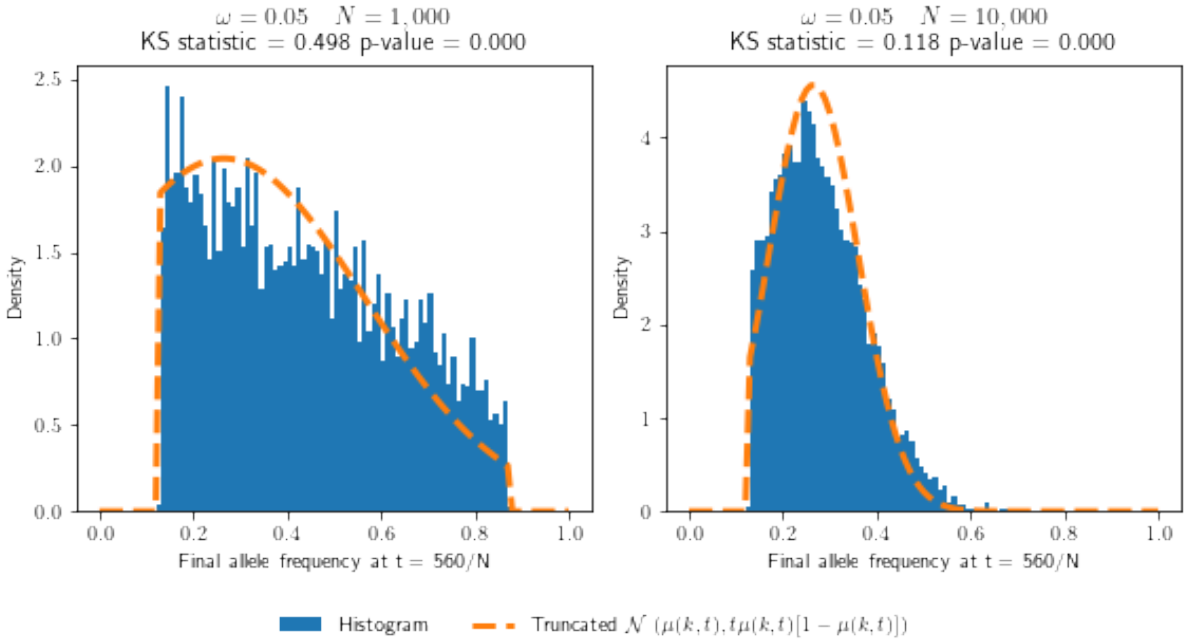


Figure A-9: Distribution of allele frequencies at $T = \frac{560}{N}$ in truncated samples

Note: The figure plots the empirical distribution of the simulated diffusion process in Equation (31) satisfying the sample selection criteria together with the theoretical distribution implied by the truncated normal approximation in Equation (A-4). The initial conditions of the simulations are given by $x(k,0) = 0.2$ and $\beta(k) = 0.0131$. The empirical distribution is based on sampling 10,000 paths from the diffusion process.

Given the truncated normal model in Equation (A-4), we can write

$$\mathbb{E}_\Psi[x(k,T)|x(k,0),\beta(k),\theta] = \mu(k,T) - \sqrt{T\mu(k,T)(1-\mu(k,T))}\frac{\phi(\bar{c}) - \phi(\underline{c})}{\Phi(\bar{c}) - \Phi(\underline{c})} \qquad \text{(A-5)}$$

$$\Psi(x(k,T)|x(k,0),\beta(k),\theta) = \frac{1}{\sqrt{T\mu(k,T)(1-\mu(k,T))}}\frac{\phi\left(\frac{x(k,T)-\mu(k,T)}{\sqrt{T\mu(k,T)(1-\mu(k,T))}}\right)}{\Phi(\bar{c}) - \Phi(\underline{c})} \qquad \text{(A-6)}$$

where

$$\bar{c} = \frac{\min\left\{1-\nu, \frac{1+\kappa}{2}\right\} - \mu(k,T)}{\sqrt{T\mu(k,T)(1-\mu(k,T))}} \qquad \text{and} \qquad \underline{c} = \frac{\max\left\{\nu, \frac{1-\kappa}{2}\right\} - \mu(k,T)}{\sqrt{T\mu(k,T)(1-\mu(k,T))}}$$

We obtain $\hat{\theta}^{\text{NLS}}_{\mathcal{N}}$ and $\hat{\theta}^{\text{MLE}}_{\mathcal{N}}$ by plugging in Equations (A-5) and (A-6) to Equations (32) and (33), respectively.

A-4.2 Numerical approximation

The second approach is based on numerical approximation of Ψ directly from the diffusion process in Equation (31). The algorithm is as follows

1. Select a SNP k with initial conditions $(x(k,0), \beta(k))$ and candidate parameter vector θ. We choose θ from the cartesian product of grid on $\omega \in \{\underline{\omega}, \bar{\omega}\}_{1\times W}$ and grid on $N \in \{\underline{N}, \bar{N}\}_{1\times S}$.

2. Simulate the diffusion process in Equation (31) B times and extract a vector $\left(x(k,T)^{(1)}, \ldots, x(k,T)^{(B)}\right)$ of finale allele frequency realizations.

3. Estimate the density function Ψ

   (a) Select the subset of strictly interior allele frequency realizations:

   $$\left\{x(k,T)^{(i)} : 0 < x(k,T)^{(i)} < 1\right\}$$

   (b) Using the subset of interior realizations, we can use kernel density estimation

   $$\hat{f}(x|x(k,0),\beta(k),\theta) = \frac{1}{B_{\text{int}}h}\sum_{i=1}^{B_{\text{int}}} K\left(\frac{x - x(k,T)^{(i)}}{h}\right)$$

   We use Gaussian kernel $K(\cdot) = \phi(\cdot)$ and optimal bandwidth h determined by Silverman's rule of thumb.

   (c) Renormalize the estimated $\hat{f}$ such that it correctly integrates to the mass of interior points. Denote the probability mass points at the boundaries as $p_l = \Pr(x(k,T) = l), \quad l \in \{0,1\}$. We estimate these probability mass points using the shares of realizations at the respective boundaries

   $$\hat{p}_0 = \frac{1}{B}\sum_{i=1}^{B} 1\left\{x(k,T)^{(i)} = 0\right\}$$

   $$\hat{p}_1 = \frac{1}{B}\sum_{i=1}^{B} 1\left\{x(k,T)^{(i)} = 1\right\}$$

Thus, after renormalization the following should hold true

$$\int_0^1 \hat{f}(x|x(k,0),\beta(k),\theta)dx = 1 - \hat{p}_0 - \hat{p}_1$$

(d) Now, the estimate of the density in the truncated sample can be written as

$$\hat{\Psi}(x(k,T)|x(k,0),\beta(k),\theta) = \frac{\hat{f}(x(k,T)|x(k,0),\beta(k),\theta)}{\int_{\max\{\nu,\frac{1-\kappa}{2}\}}^{\min\{1-\nu,\frac{1+\kappa}{2}\}} \hat{f}(x|x(k,0),\beta(k),\theta)dx} \tag{A-7}$$

4. Estimate the conditional mean $\mathbb{E}_\Psi[x(k,T)|x(k,0),\beta(k),\theta]$

(a) Select the subset of realizations satisfying the truncation condition:

$$\left\{x(k,T)^{(i)} : \max\left\{\nu, \frac{1-\kappa}{2}\right\} \leq x(k,T)^{(i)} \leq \min\left\{1-\nu, \frac{1+\kappa}{2}\right\}\right\}$$

Let B_{tr} denote the size of this subset.

(b) Using this subset, compute the sample mean

$$\widehat{\mathbb{E}_\Psi}[x(k,T)|x(k,0),\beta(k),\theta] = \frac{1}{B_{\text{tr}}}\sum_{i=1}^{B_{\text{tr}}} x(k,T)^{(i)} \tag{A-8}$$

Similarly, we obtain $\hat{\theta}_B^{\text{NLS}}$ and $\hat{\theta}_B^{\text{MLE}}$ by plugging in Equations (A-7) and (A-8) to Equations (32) and (33), respectively.

Even though this approximation is based on the exact diffusion process that generates Ψ, there may still be approximation error. First, due to sample size B used to estimate the final allele frequency properties. Setting B very high would ensure that the approximation error is minimized, but at the cost of exploring fewer candidate parameters for θ due to computational constraints. Second, the estimated pdf or cdf at the tails of the frequency distribution may collapse to zero or one due to machine precision constraints. Thus, the estimators based on numerical approximation are not immune to approximation bias. The question is whether the bias under numerical approximation is smaller than under normal approximation. To answer this question we study the bias of estimators in a Monte-Carlo simulation framework in Section 7.

A-5 Winner's curse

It is well known that by selecting the statistically-significant lead SNPs we get a biased information about the distribution of true effects in the population. In particular, Okbay et al. (2016) provide detailed derivations of the posterior distribution in the supplementary materials taking into account the criteria that GWAS effect sizes are non-null and statistically significant. We summarise their argument below.

Suppose the true GWAS effects follow mixture distribution: they are either normally distributed with probability π or null with probability $1-\pi$:

$$\beta \sim \begin{cases} \mathcal{N}(0, \tau^2) & \text{with probability } \pi \\ 0 & \text{otherwise} \end{cases} \tag{A-9}$$

The distribution of a GWAS estimator $\hat{\beta}$ can be approximately written as $\hat{\beta}|\beta \sim \mathcal{N}(\beta, \sigma^2)$, where β is the true effect and σ^2 captures sampling uncertainty of the estimator. Given this, the posterior distribution of true non-null GWAS effects can be written as

$$\beta|\hat{\beta}, \beta \neq 0 \sim \mathcal{N}\left(\frac{\tau^2}{\sigma^2+\tau^2}\hat{\beta}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) \tag{A-10}$$

Okbay et al. (2016) propose an ML estimator of the parameters of the prior distribution, namely τ and π, that solve the following problem

$$(\hat{\tau}, \hat{\pi}) = \arg\max_{\tau,\pi} \sum_{k=1}^{K} \log\left[\frac{1}{\sqrt{1+N_k t_k^2}}\phi\left(\frac{\hat{z}}{\sqrt{1+N_k t_k^2}}\right)\pi + \phi(\hat{z})(1-\pi)\right] \tag{A-11}$$

where N_k is the GWAS sample size; $t_k^2 = \tau^2 2x_k(1-x_k)$ and x_k is the reference allele frequency of SNP k; $\hat{z}$ is the estimated z-statistic from GWAS equivalent to $\frac{\hat{\beta}}{\hat{\sigma}}$; and $\phi(\cdot)$ is the standard normal pdf. Applying this algorithm, we estimate $\hat{\tau} = 0.0035$ (SE 3.06×10^{-6}) and $\hat{\pi} = 0.5927$ (SE 0.0009) using a full sample of more than 10 million SNPs. For comparison, following a similar estimation procedure applied to a subsample of more than 4.5 million SNPs with sibling cohorts excluded Lee et al. (2018) estimate $\hat{\tau}^2 = 4.5 \times 10^{-6}$ and $\hat{\pi} = 0.64$.

A-6 Full Model Estimation

A-6.1 Effect of Population size

In this subsection we explore how the two parameters, selection strength ω and population size, affect in different ways the final distribution of alleles. It is clear from the analysis developed in the section describing the full model (3.3) that a larger population size reduces the variance of the process; as the population size tends to infinity the process tends to a deterministic limit process. We define this process precisely. Note that the fucntion T defined in equation (10) depends on the population size N. We denote

Definition A-6.1. *The deterministic map (denoted by T_d) associated with the process defined in section (3.3) is obtained by replacing the set $\Delta_N(\mathbf{H})$ with the set $\Delta(\mathbf{H})$, and by replacing the random variables by their expectation.*

When the population size is small, the probability that an allele frequency hits one of the two boundary values (0 or 1), for a fixed value of all other parameters, is higher than it is at larger values of the population.

The next proposition is clear:

Proposition A-6.2. *The difference equation defined by the map T_d is the limit as $N \to +\infty$ of the maps T in equation (10).*

It is interesting to compare the effect of the two parameters on the two statistics we use to evaluate how well the model predicts the true process, namely the distance of predicted from simulated phenotype and the distance (weighted or not) between simulated and true allele frequency. The key consideration is that the same mean value of phenotype can be obtained with many different distributions of allele frequency. In view of this, the same mean phenotype can be obtained with a larger number of alleles with frequency at the boundary (when the population size is small), and that same phenotype value can be obtained with a smaller number at the boundary (when the population size is larger). Figure A-10 illustrates how predictions change keeping all the parameters fixed and only changing the population size.

The figure compares the true allele frequency with the simulated one, for two values of the population size. The value of the final phenotype is approximately the same in both cases, but it is clear that the model with a larger population size is a better predictor than the small population size. As it is natural to imagine and easy to verify, the scatter plot for intermediate values is approximately a combination of the figure obtained at these two extreme values. The convergence to the values portrayed in the right panel (large population) occurs quickly, at values of M between to 3 to 5 thousand.

Figure A-11 illustrates the content of proposition A-6.2 in the case of the two population sizes considered in proposition A-6.2. Each panel presents the allele frequency of

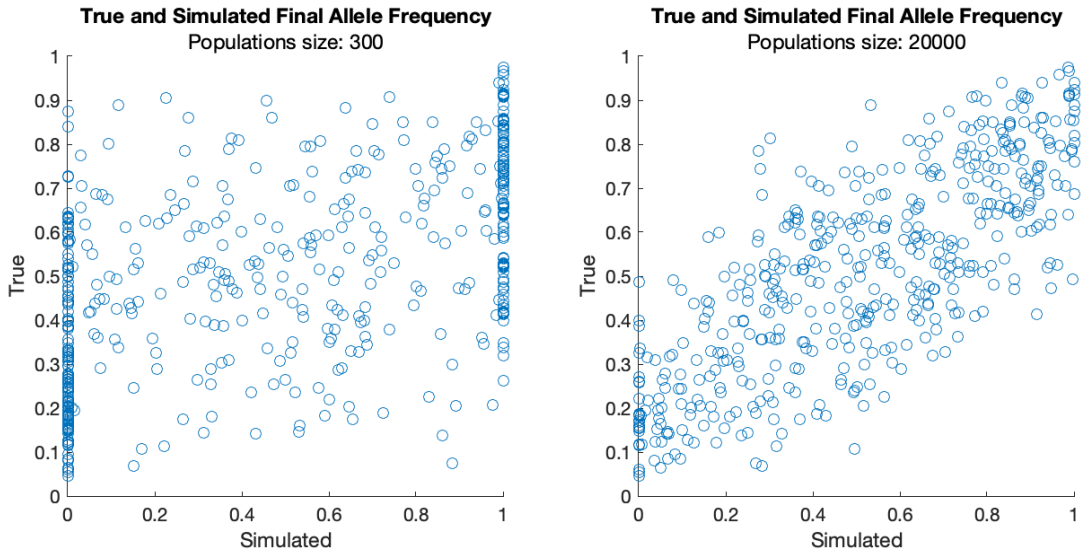


Figure A-10: **Distribution of simulated final allele frequencies compared to true.** Effect of population size on final allele frequency. Left panel: small population size. Right panel: large population size.

Note: The figure plots the simulated allele frequency using the full model as presented in section 3.3.

each allele in two independent realization of the stochastic process described by the full model. The only difference between right and left panel is the population size, smaller in the left panel and larger in the right panel. In the limit of large population sizes the scatter plot converges to the diagonal for every realization of the process, that is, the process becomes completely deterministic.

It is, therefore, not surprising that the variance of the phenotype distribution in the population is increasing with population size. The variance of the phenotype can be computed as

$$2\sum_{k=1}^{K}\beta(k)^2 x(k)(1-x(k)) \tag{A-12}$$

where $\beta(k)$ is the GWAS effect and $x(k)$ is the allele frequency of SNP k. Since $x(k) \in [0,1]$, the variance of allele frequency is maximised when $x(k) = 0.5$. As demonstrated in Figures A-10 and A-11, allele frequencies are more likely to converge to the boundaries when population size is small, thereby, lowering the phenotype variance. The upper boundary of the phenotype variance given the GWAS effects is, therefore, given by $2\sum_{k=1}^{K}\beta(k)^2 0.5^2$.

Figure A-12 demonstrates the relationship between phenotype variance, population size and fitness parameter ω_{AG}. First, it is clear that selection parameter has minimal effect on phenotype variance: lines corresponding to different values of $\omega_{AG} \in [0, 0.055]$ are essentially overlaid on top of each other. Second, the phenotype variance rises steeply with population sizes $M \leq 10K$ and nearly flattens as population size grows further. And

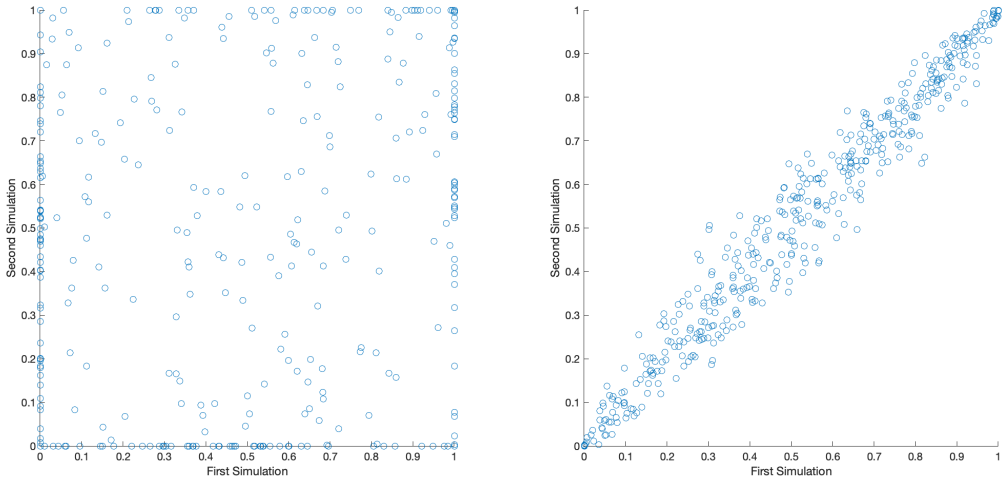


Figure A-11: **Scatter plot of two independent simulations of the process.** Effect of population size on final allele frequency. Left panel: small population size (300). Right panel: large population size (20,000).

Note: The figure plots the simulated allele frequency using the full model as presented in section 3.3.

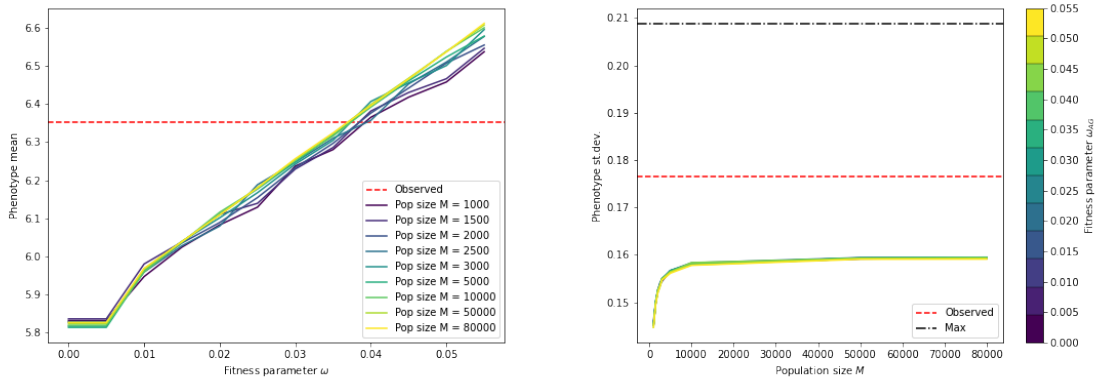


Figure A-12: **Phenotype distribution by population size and fitness parameter ω_{AG}.** The red dashed horizontal line indicate the observed mean and standard deviation of phenotype in 1000GP. The black horizontal line indicates the max phenotype standard deviation given the GWAS effect sizes.

Note: The figure plots the average phenotype mean and standard deviations over 50 independent simulations using the full model as presented in section 3.3.

finally, our main model presented in Section 3.3 is more likely to push the allele frequencies towards the boundaries compared to the observed allele frequencies in the data. This can be seen by the gap between the average simulated phenotype standard deviations and the observed standard deviation of the phenotype in 1000GP (red dashed line). This observation also holds in a subsample of 440 SNPs whose initial allele frequencies were in the range between 0.01 and 0.99. Thus, the tendency of the main model to steer the allele frequency paths towards the boundaries can not be entirely explained by the characteristics of the observed data.

Figure A-12 also demonstrates the near independence of the two parameters: ω_{AG} and M. The left panel shows that phenotype mean is almost exclusively a function of ω_{AG}. The lines corresponding to population sizes $M < 5K$ require slightly higher values of ω_{AG} to match the observed average phenotype in 1000GP. However, the lines corresponding to $M \geq 5K$ are almost identical to each other. The right panel shows that phenotype variance is almost exclusively a function of population size M. This suggests that we can estimate the two parameters separately: estimate ω_{AG} to match the observed average phenotype and estimate M to match the observed phenotype variance. However, the right panel also suggests that in the current setting, we cannot obtain a sharp estimate of M: there is likely no value of M large enough that would help simulations of the main model attain the observed phenotype variance in 1000GP. Nevertheless, the figure also suggests that setting $M = 5K$ would provide a satisfactory estimate of ω_{AG}.

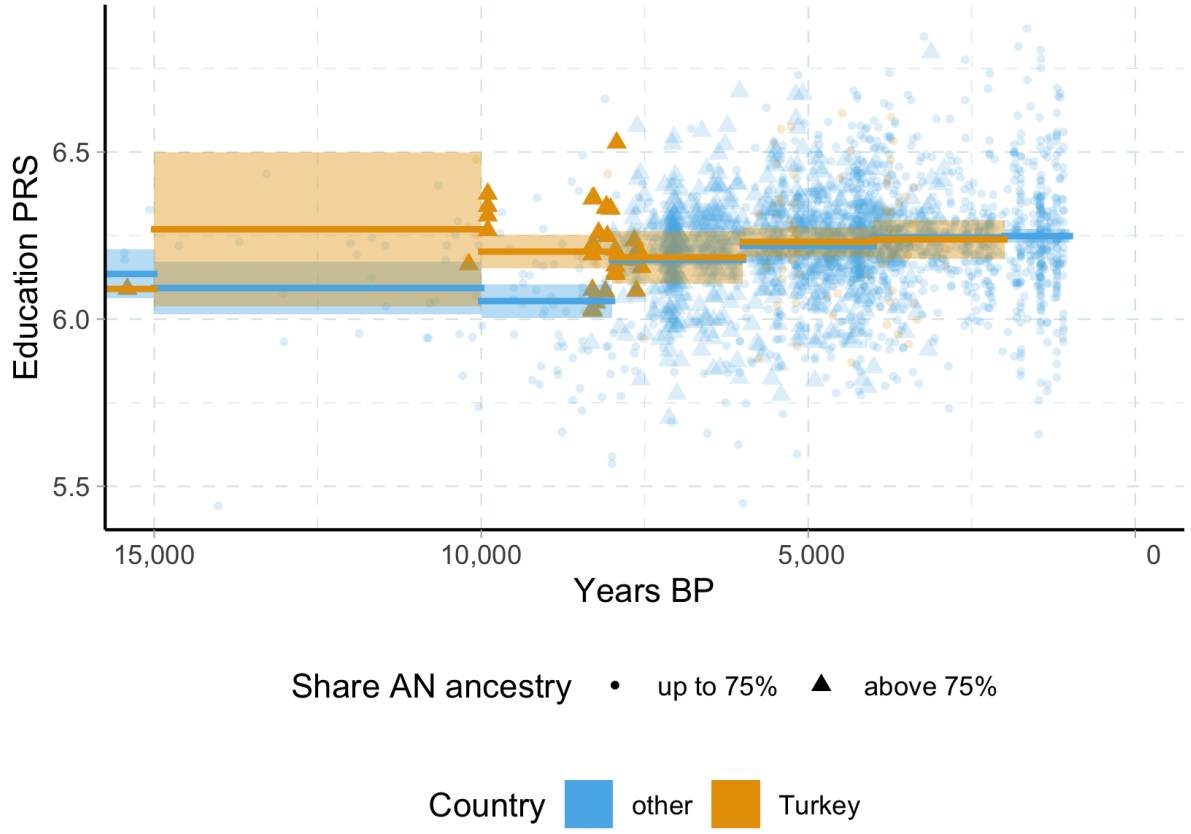


Figure A-13: Education PRS over time by ancestry and archaeological country

Note: the figure plots the polygenic score (PGS) of ancient individuals in the analysis sample (points) and average PGS scores in each period (solid lines). The individual PGS score are differentiated by country of archaeological site and by share of AN ancestry estimated using ADMIXTURE. The shaded areas around the solid lines correspond to 95% confidence intervals based on t-distribution.

A-7 Migration

The Figure A-13 shows that average PGS of ancient individuals found outside of Turkey rapidly "caught up" to the AN-level. The only time period when they were distinct is between 10K-8K. By the end of this period, average PGS of individuals in Turkey is indistinguishable from average PGS of individuals in other countries. This is consistent with the fast displacement of WHG by AN in Figure A-4.