Selection and the Roy Model

Fatima Jandarova, Aldo Rustichini

Winter School 2022 at the Delhi School of Economics,
Lecture 2

Question addressed here

1. We test the hypothesis that in the recent past ($< 14K$ years Before Present (BP)) there has been a selective force operating in the direction of an increase of the frequency of Educational Attainment (EA) enhancing alleles, following the transition from foraging to agriculture in the Neolithic Agricultural Transition (NAT).
2. The general hypothesis we consider is that technical change induced genetic change through a change in allele frequencies, and as a consequence a change in economic activity and institutions
3. This is a special instance of a general relation we hypothesize between technology, genetics and institutions, each level determining the next.

Neolithic Agricultural Transition

NAT

1. NAT was the wide-scale transition (starting 11,600 years BPE) of many human cultures during the Neolithic period from a lifestyle of hunting and gathering to one of agriculture and settlement (foragers to farmers).
2. Preceded by earlier transformations of foraging (Natufians), and associated with a climate change (warming).
3. It involved a widespread process of *genetic transformation* of plants and animals (domestication). It is reasonable to conjecture that a similar process occurred for human population.
4. Originated in the Fertile Crescent and spread to the rest of Europe at the speed of 1 km per year

Neolithic Demographic Transition

NDT

1. The NAT made an increasingly large population possible, inducing a Neolithic Demographic Transition (NDT)
2. NDT is identifiable by the increase in the proportion of immature skeletons in cemeteries ($p_{(5,15)}$, Bouquet-Appel Masset, 1977, Bouquet-Appel 2002)
3. This increase signals a sudden increase in fertility, in turn induced by change in the energy balance of mothers (both in intake –change in diet in the direction of carbohydrates– and outtake –reduction of mobility).
4. Followed by an increase in mortality induced by sedentism, crowded living conditions. (Time sequence opposite to Contemporary DT)

NAT and NDT as Institutional Change

1. A potential puzzle is derived from *claims* that (i) productivity among foragers was not lower than among farmers, and (ii) health conditions of foragers were not worse, or even better than among farmers.
2. Common interpretation of these claims (Bowles 2011, Bowles & Choi 2013, Robson, 2011, Rowthorne & Seabright, 2010): the transition from foraging to farming was induced by (or was associated with) an institutional change.
3. The key transformation was the emergence of private property: planted seeds and domesticated animals had to be protected from theft.
4. Model (RS) Nash equilibrium in a game where two groups choose between foraging and farming. The game is a Prisoner's dilemma (foraging = cooperate, farming = defect): if a group chooses farming, it develops a military force that can be used for defending but also for robbing; thus the other has to adopt it as well.

Mullah Nasreddin's Lost Ring

Mullah had lost his ring in the living room. He searched for it for a while, but since he could not find it, he went out into the yard and began to look there. His wife, who saw what he was doing, asked:
"Mullah, you lost your ring in the living room, why are you looking for it in the yard?"
Mullah stroked his beard and said:
"The room is too dark and I can't see very well. I came out to the courtyard to look for my ring because there is much more light out here"

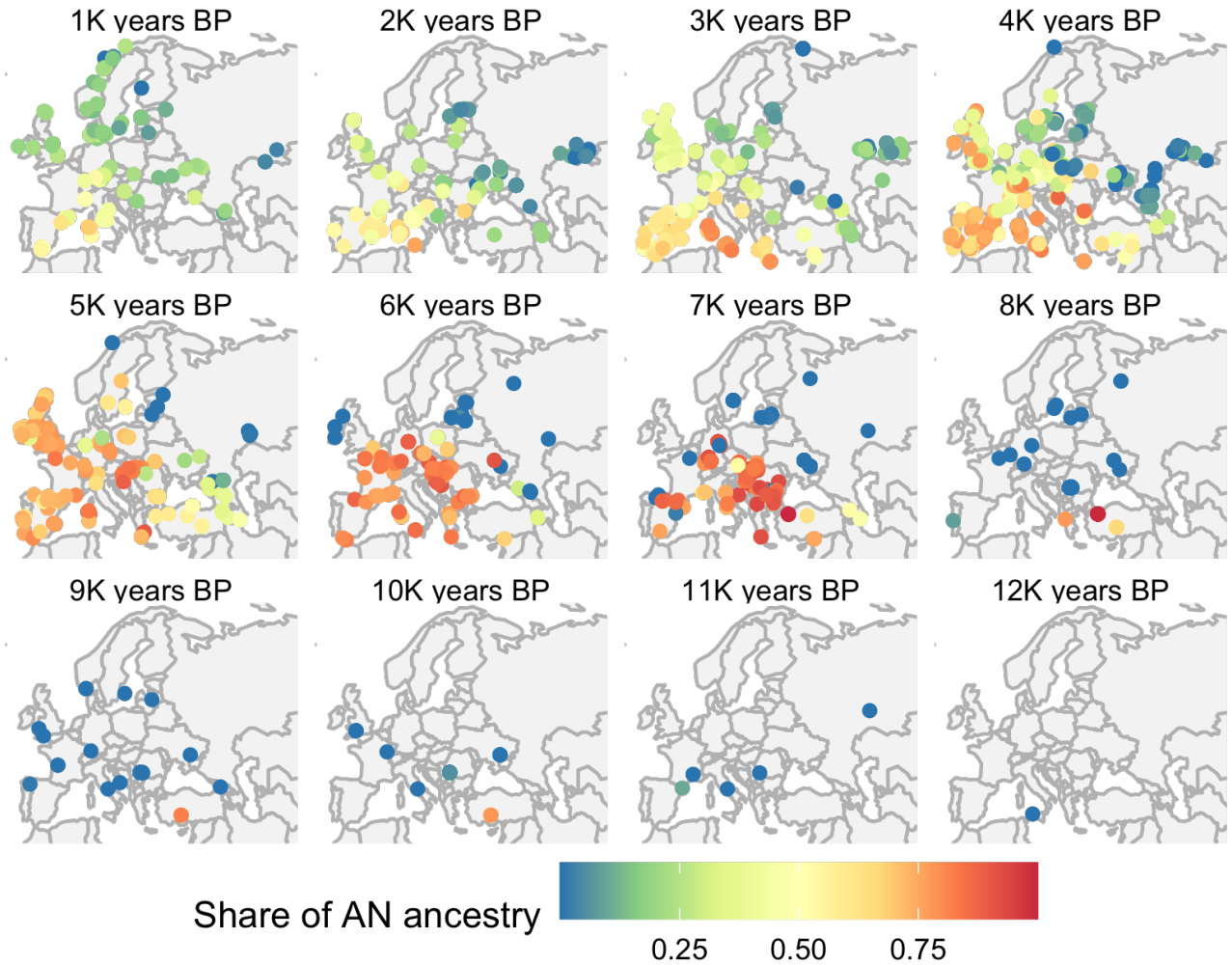
The Discreet Charm of the Hunter-Gatherer

The evidence about early Neolithic living standards perhaps adds substance to the eternal appeal that myths of the noble savage have had throughout human history, since such myths have seemed to suggest, counter-intuitively, that economic development since the time of the alleged fall has been both inevitable and regrettable.
[Rowthorne & Seabright, 2010]

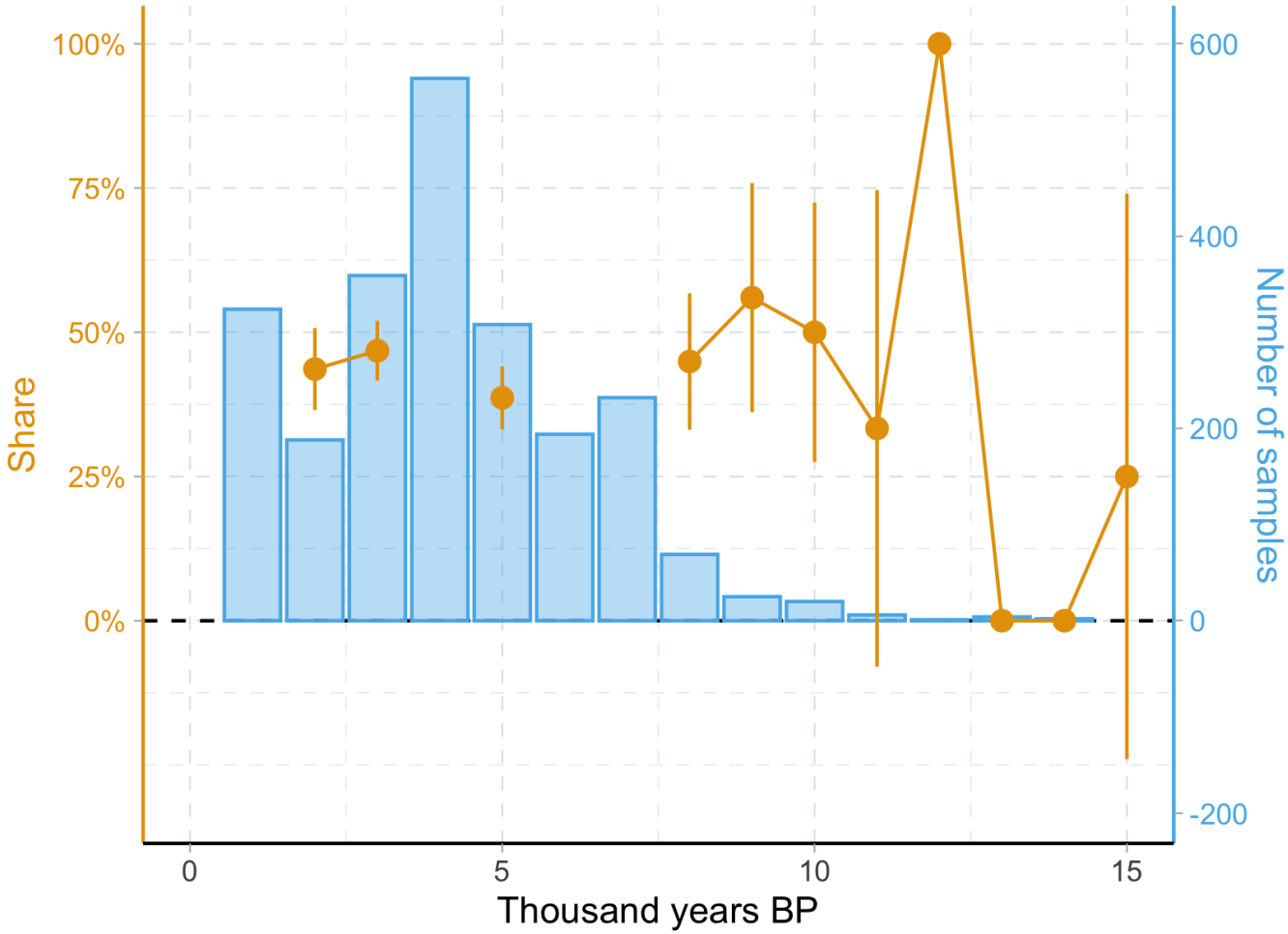
Alternative Interpretation and Modeling of the Two Transitions

1. A well documented climate change induced a change in the productivity of two technologies, foraging and farming
2. This change of climate in turn induces a change in the allocation of individuals with different skills to different occupations
3. The different productivity induces a different reproductive fitness, and selective pressure, hence a change in the distribution of genotype
4. Creation of larger concentrated communities, cities, and new institutions

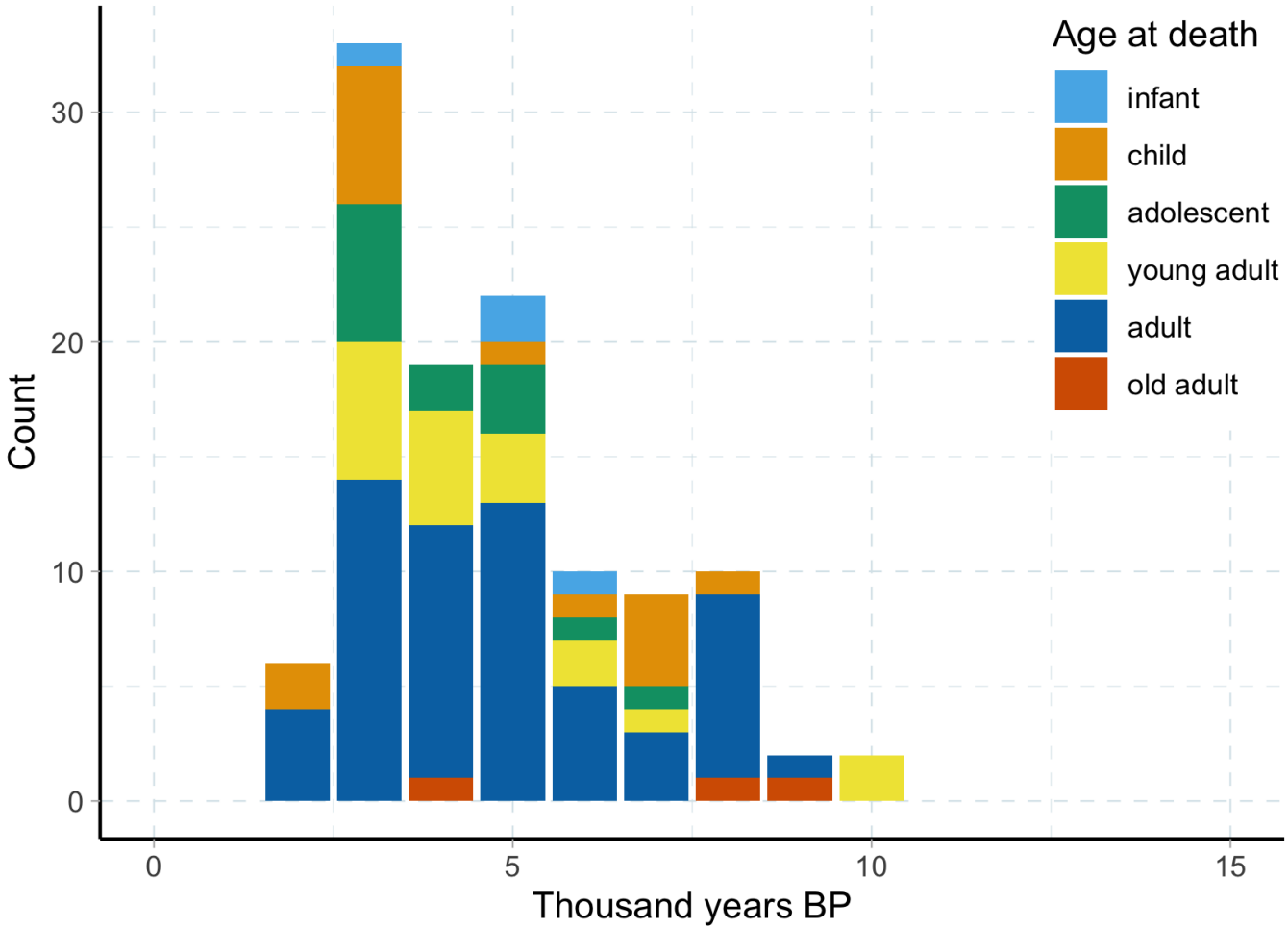
Geographic Distribution of *aDNA* sample, Anatolian Farmers



Sex of skeletal remain



Age of skeletal remain



Climate Change

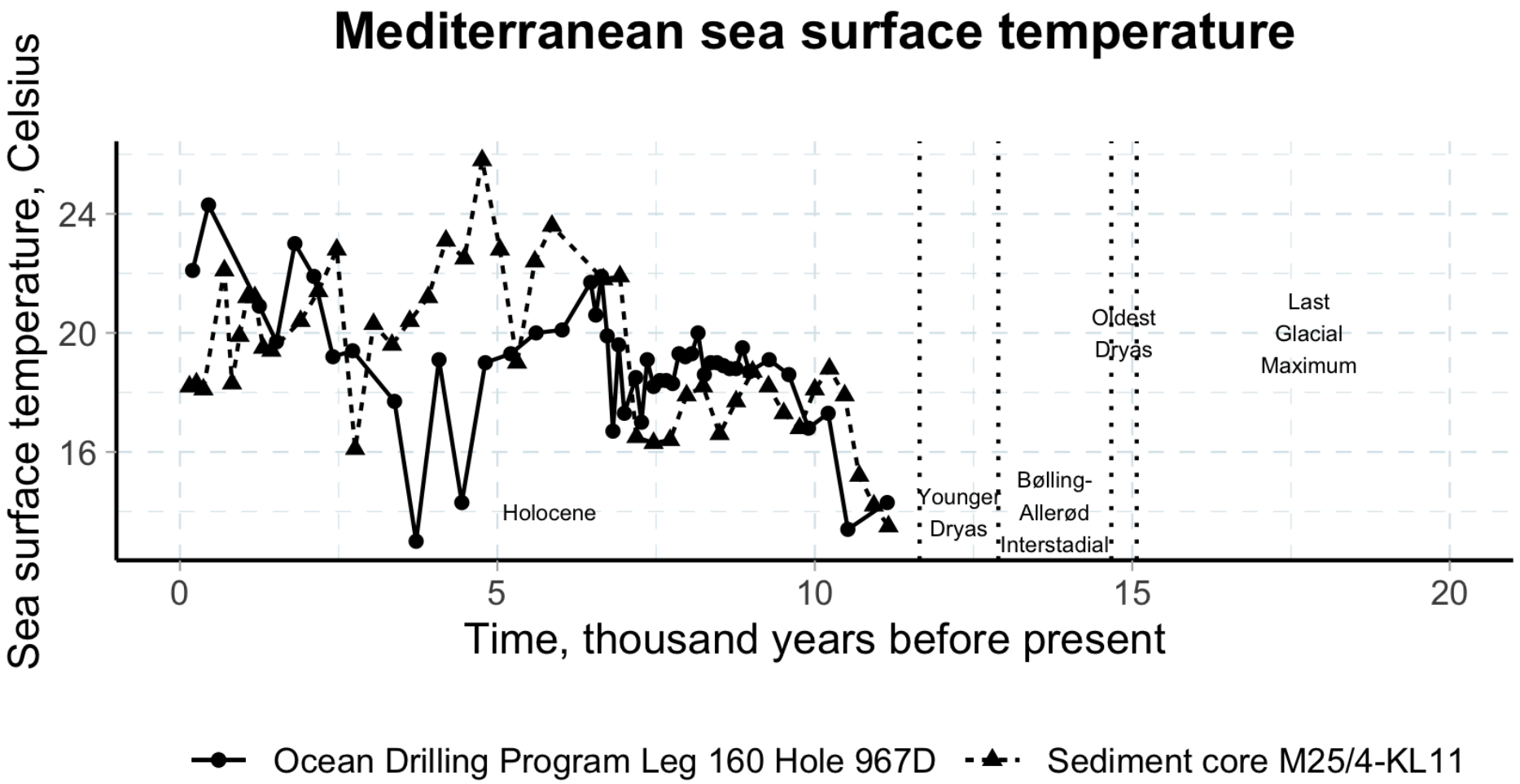
Dryas octopetala

Climate Change at the end of Pleistocene

Summary of facts

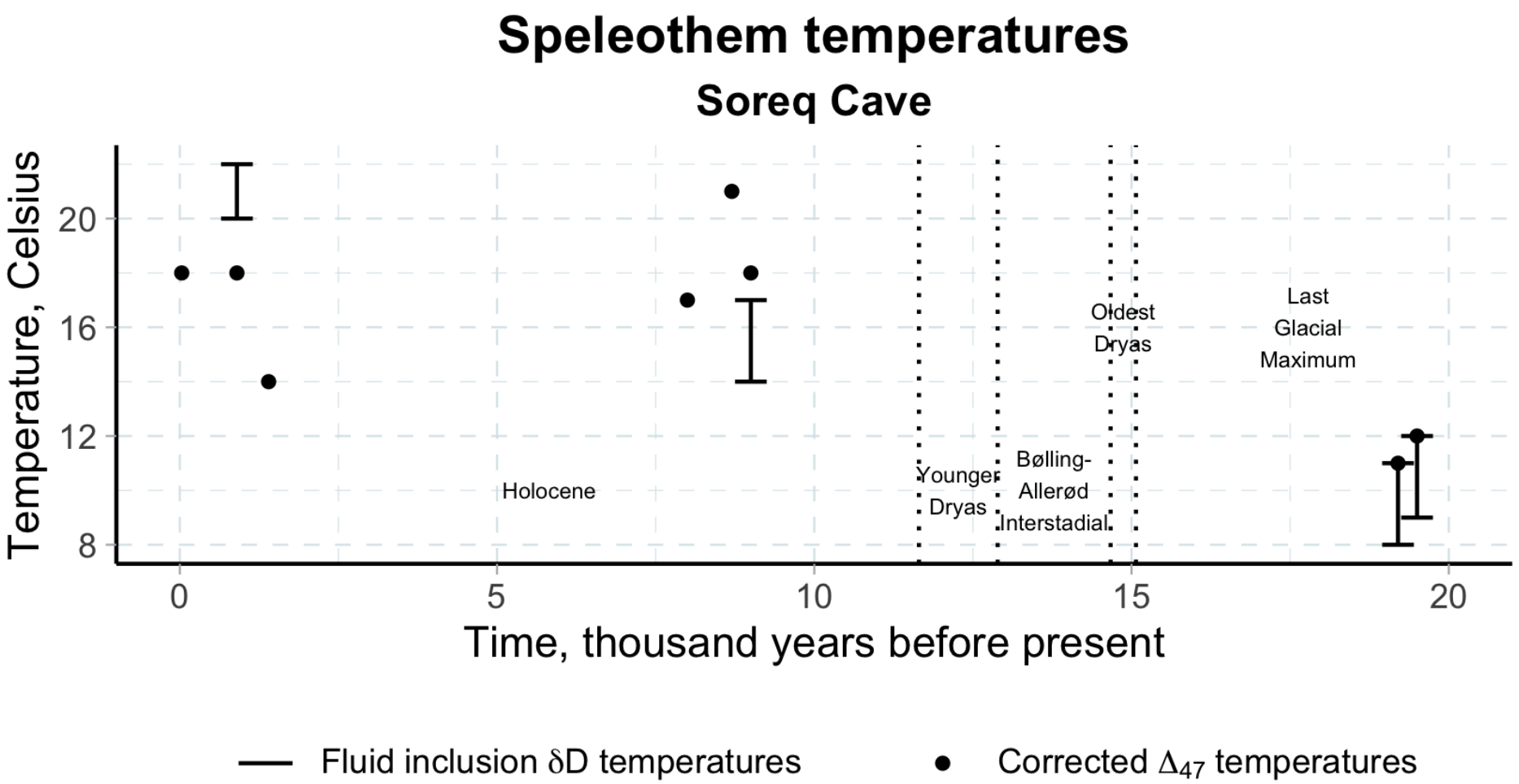
1. Temperature increase in the Bølling Allerød (BA) period (14,690 to 12,890 Years Before Present (YBP)), followed by a cooling in the Younger Dryas (YD) (12,890 to 11,600 YBP), followed by a sustained increase in temperature, creating in the Fertile Crescent condition favorable to agriculture
2. Early adoption of advanced *storage* techniques, perhaps early adoption of *farming* techniques (Natufian Civilization, BA and YD, 15K to 11.5 K years YBP)
3. Rapid spread of the Anatolian Farmers (AN) population after the YD.

Climate History



Source: Emeis et al. (2000)

Local Temperature (Soreq cave)



Source: Affek et al. (2008)

Isotope data: Orland et al, Climate Deterioration...", *Quaternary Research*, (2009).

Estimation Problem

1. We take the frequency of alleles in the *WHG* population as initial condition
2. We take the current frequency of alleles as final condition
3. Provide a selection model depending on a parameter that measures direction and strength of selection
4. Measure the distance between predicted final frequency and current frequency, and test the hypothesis that the strength of selection parameter is positive and significantly so.
5. Initial phenotype value given parameters is 5.8, final is 6.3

A model of selection

1. Wright-Fisher (WF) model on a finite population, with mutation and selection;
2. The unit of selection is the genotype; however, WF in the original form is not suitable to study the distribution of allele frequency because of sexual reproduction. (think of fitness of a single allele with fitness of heterozygote strictly larger than both homozygote);
3. Model random mating, so the allele population is in HW equilibrium in every period;
4. Selection operating through a fitness function.
5. Setup as in Rustichini et al, Educational attainment and Intergenerational Mobility, *Journal of Political Economy*, 2023

Haplotypes and Genotypes Models

A simple illustration

1. Two loci, $\{A, a\} \times \{B, b\}$. A, B count as 1; a, b as 0
2. Father: $(A, B; A, B)$, genotype of father $(2, 2)$,
3. Mother number 1: $(a, B; A, b)$
4. Mother number 2: $(a, b; A, B)$
5. Genotype of both mothers $(1, 1)$
6. The map from genotype of parents to to distribution of genotypes of offspring is not well defined

Genotypes and allele frequencies

1. K is the number of loci
2. M is the number of individuals;
3. $N = 2M$ total number of alleles;
4. $\mathbb{H} \equiv \{0,1\}^K \times \{0,1\}^K$ the set of haplotypes;
5. $\mathbb{G} \equiv \{0,1,2\}^K$ the set of genotypes;
6. $p \in [0,1]^K$ the allele frequency vector;
7. $\pi \in \Delta_M(\mathbb{H})$ frequency of haplotypes;
8. If $S : \mathbb{H} \to \mathbb{G}$, $S(h_l, h_r) = h_l + h_r$, $S_*(\pi) \in \Delta_M(\mathbb{G})$

Phenotype and Fitness

Let

1. Z is the phenotype set (real numbers);
2. $z : \mathbb{G} \to Z$ the phenotype map, linear

$$z(g) = \sum_k \alpha_k g(k);$$

3. A fitness function $f : Z \to \mathbb{R}$, such as

$$f(z) = \omega_d z \text{ (directional)}$$

or

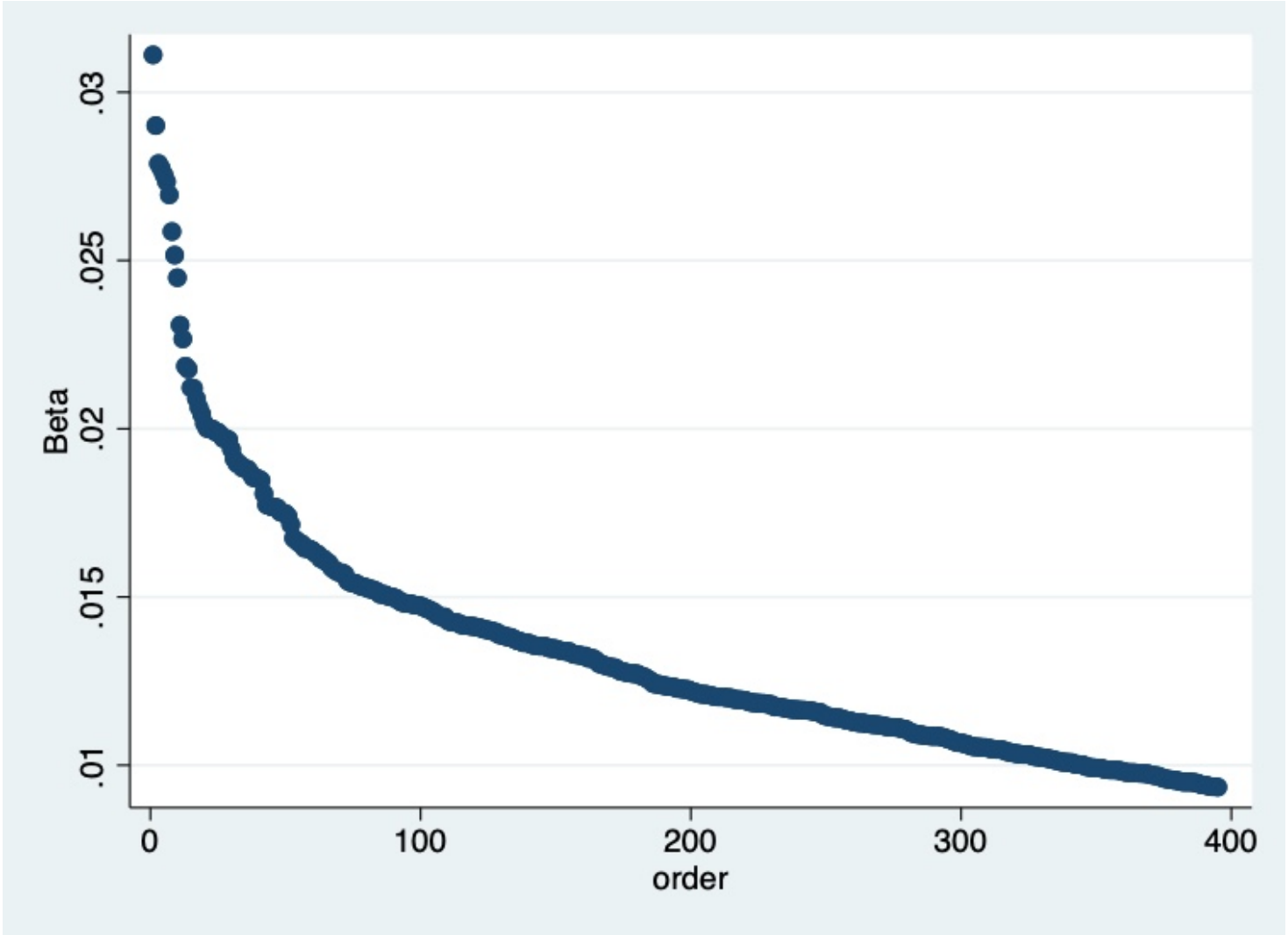
$$f(z) = -\omega_s (z - \hat{z})^2 \text{(stabilizing)}$$

or something more complex (see later);

4. $F(z) = e^{f(z)}$ selection factor

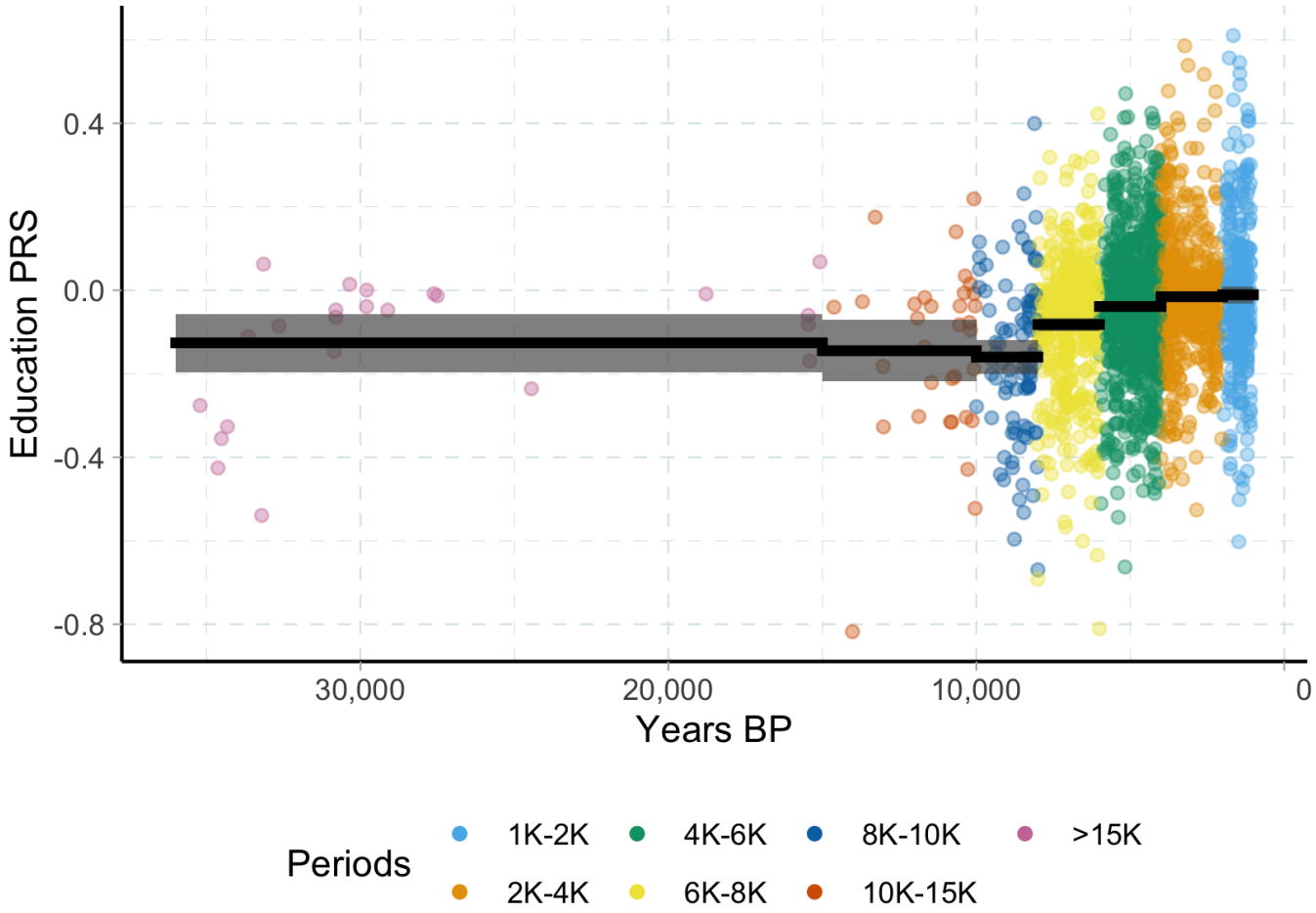
Phenotype and Genotype

Estimated coefficients of top SNP's in EA3



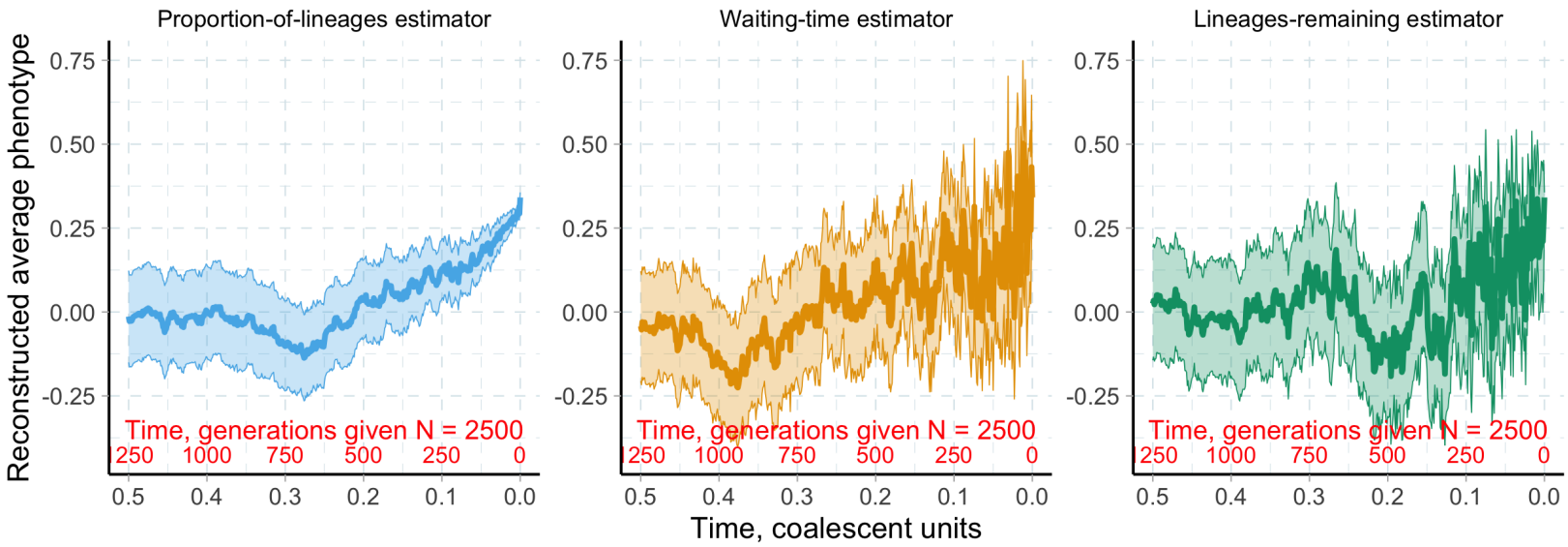
The Data: Overview

Ancient DNA



The Data: Overview

Contemporary DNA



Process within a generation at t

1. (**Initial Condition**) $p(t) \in [0,1]^K$ initial allele frequency;
2. (**Mutation**) For the given p, $q \equiv p + \frac{n-m}{N}$ where
$$m \sim \mathbf{bn}(pN, \mu), n \sim \mathbf{bn}((1-p)N, \mu);$$
3. (**Hardy-Weinberg**) For every k, $hp(k) \in \Delta(\{0,1,2\})$ is the HW probability;
4. (**No LD**) For every hp, $\otimes hp \in \Delta(G)$ is the product of the marginals:
$$\otimes hp(g) = \Pi_{k=1}^{K} hp(k, g(k));$$
5. (**Selection**) New distribution $x \in \Delta(G)$:
$$x(g) = \frac{F(z(g)) \otimes hq(g)}{E_{\otimes hq} F(z(\cdot))};$$
6. (**Wright-Fisher**) New population $(P(g, t+1) : g \in G)$:
$$P(\cdot, t+1) \sim \mathbf{mn}(M, x);$$
7. (**New Frequency**) For all k:
$$p(k, t+1) = \sum_{h \in G_{-k}} (0.5P((h, A_k, a_k), t+1) + P((h, A_k, A_k), t+1))$$

producing $p(t+1)$; start over.

Steady states with large population

1. The process from $p(t)$ to $p(t+1)$ describes a stochastic process, with randomness depending on K (cardinality of relevant SNP's), M (population size), and μ (mutation rate); call this map Ψ;
2. Considering first the large population limit ($M \to +\infty$), we get:

$$\Psi(p,k) - p(k) =$$

$$\frac{q(k)(1-q(k))}{E_{\otimes hq}F(z(\cdot))}\left(E_{q(k)}\left(E_{\otimes_{-k}hq}F(z(\cdot|A_k,\cdot) - E_{\otimes_{-k}hq}F(z(\cdot|a_k,\cdot))\right)\right);$$

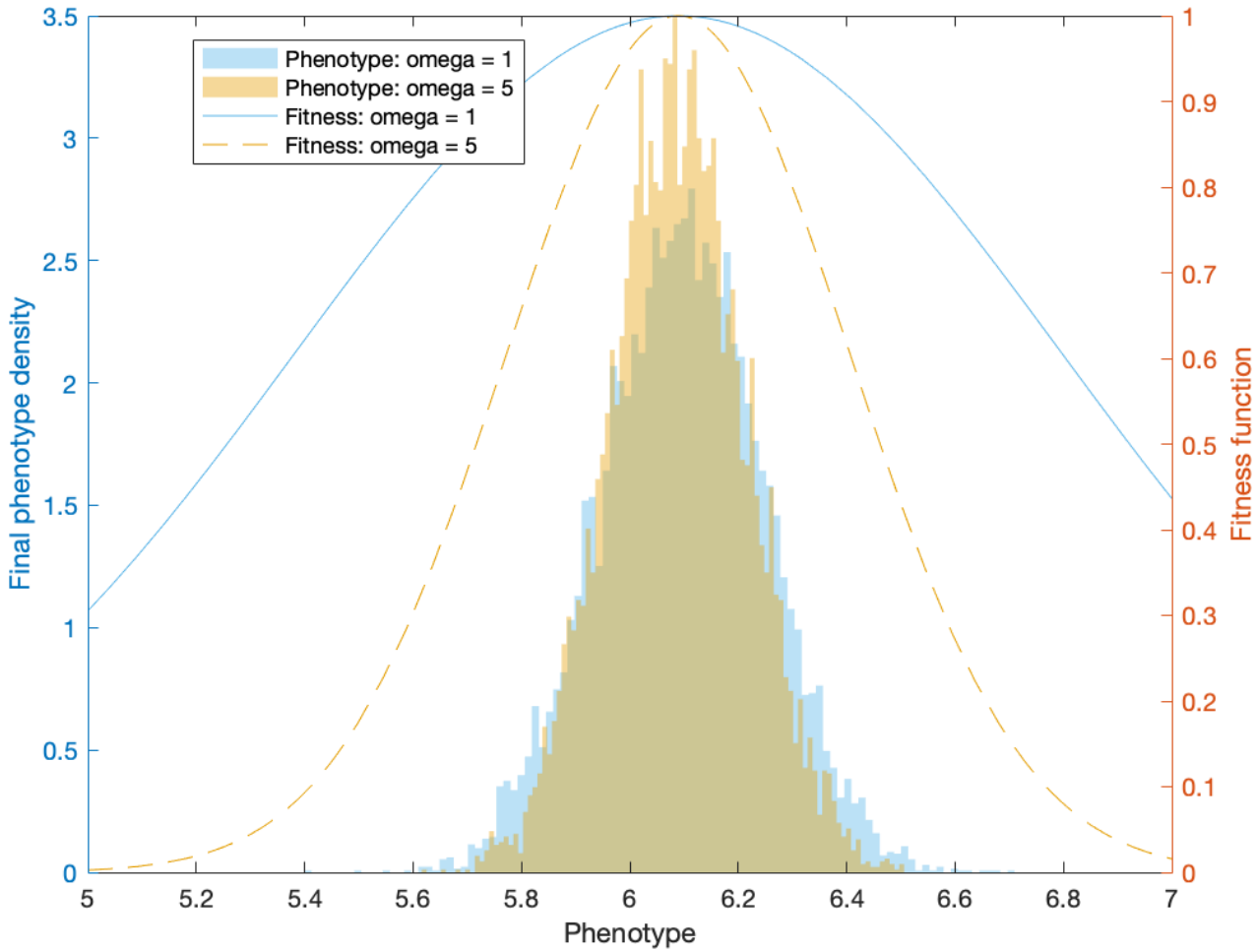
3. Using the fact that each $\beta(k)$ is small, and setting $\mu = 0$, so $q = p$:

$$\Psi(p,k) - p(k) \simeq$$

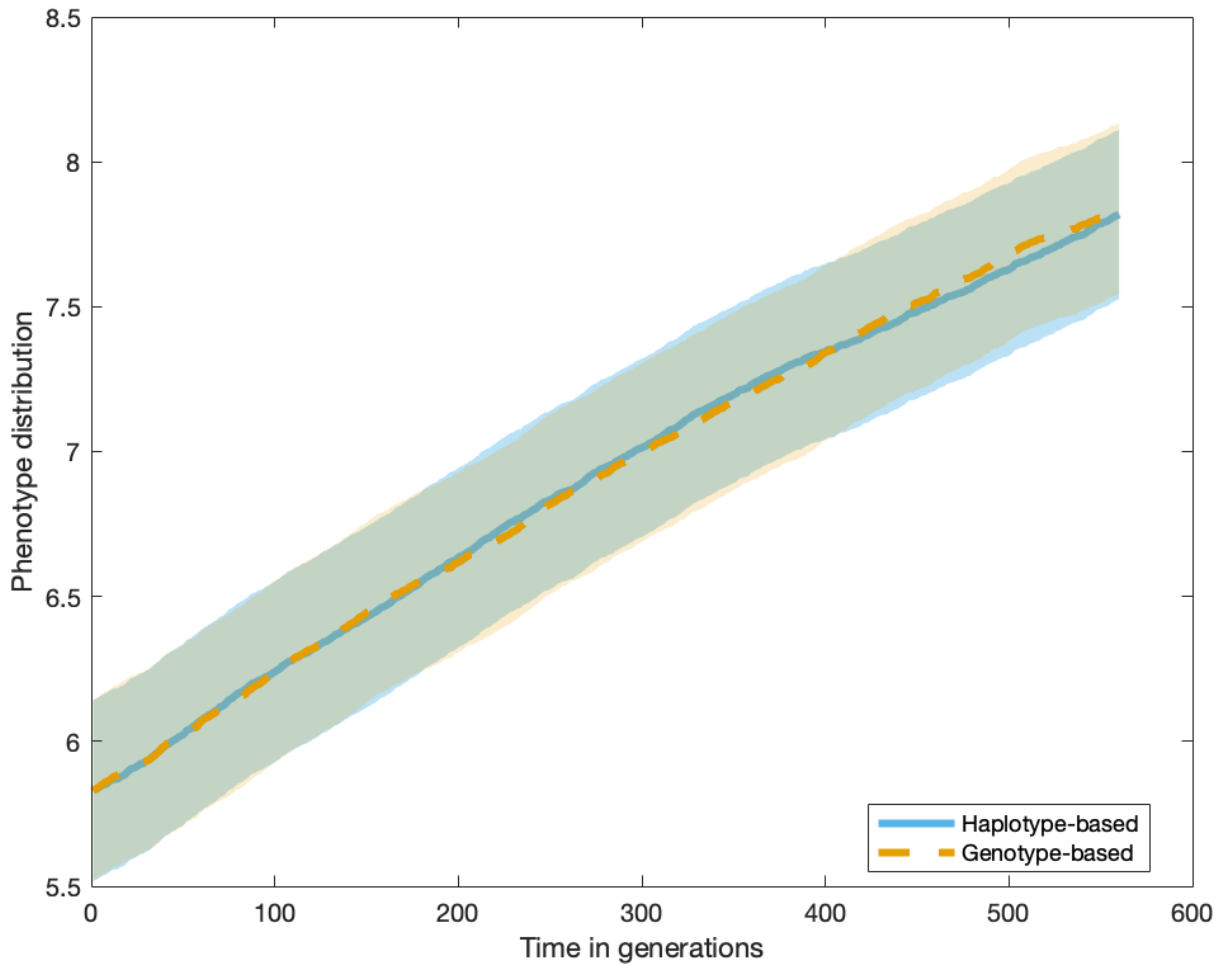
$$\beta(k)p(k)(1-p(k))\frac{E_{\otimes hp}F'(z(\cdot))}{E_{\otimes hp}F(z(\cdot))}.$$

The shape of the Invariant Measure

Variance of Fitness Function



Comparing Haplotype and Genotype Models

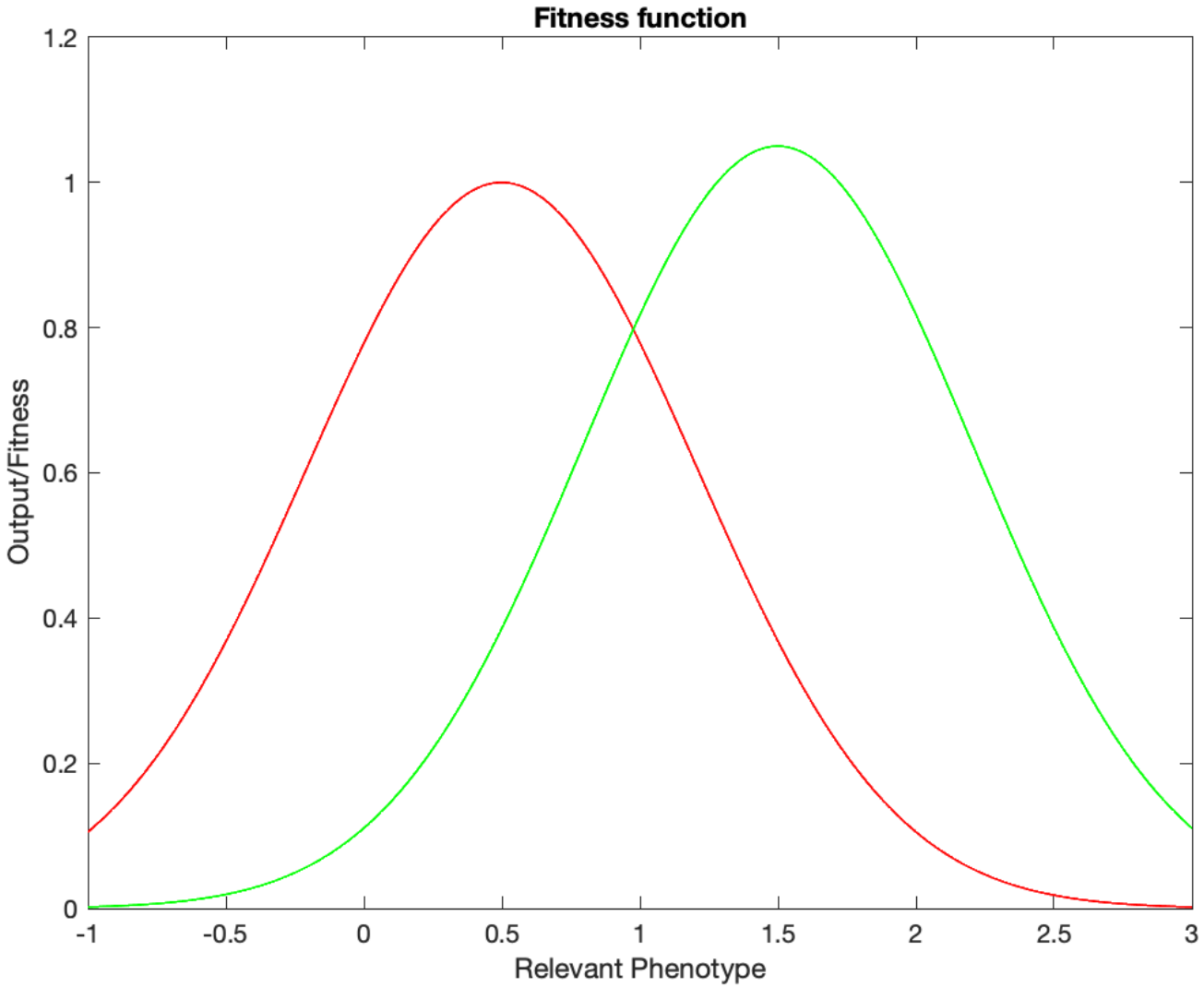


Technology, fitness and selection

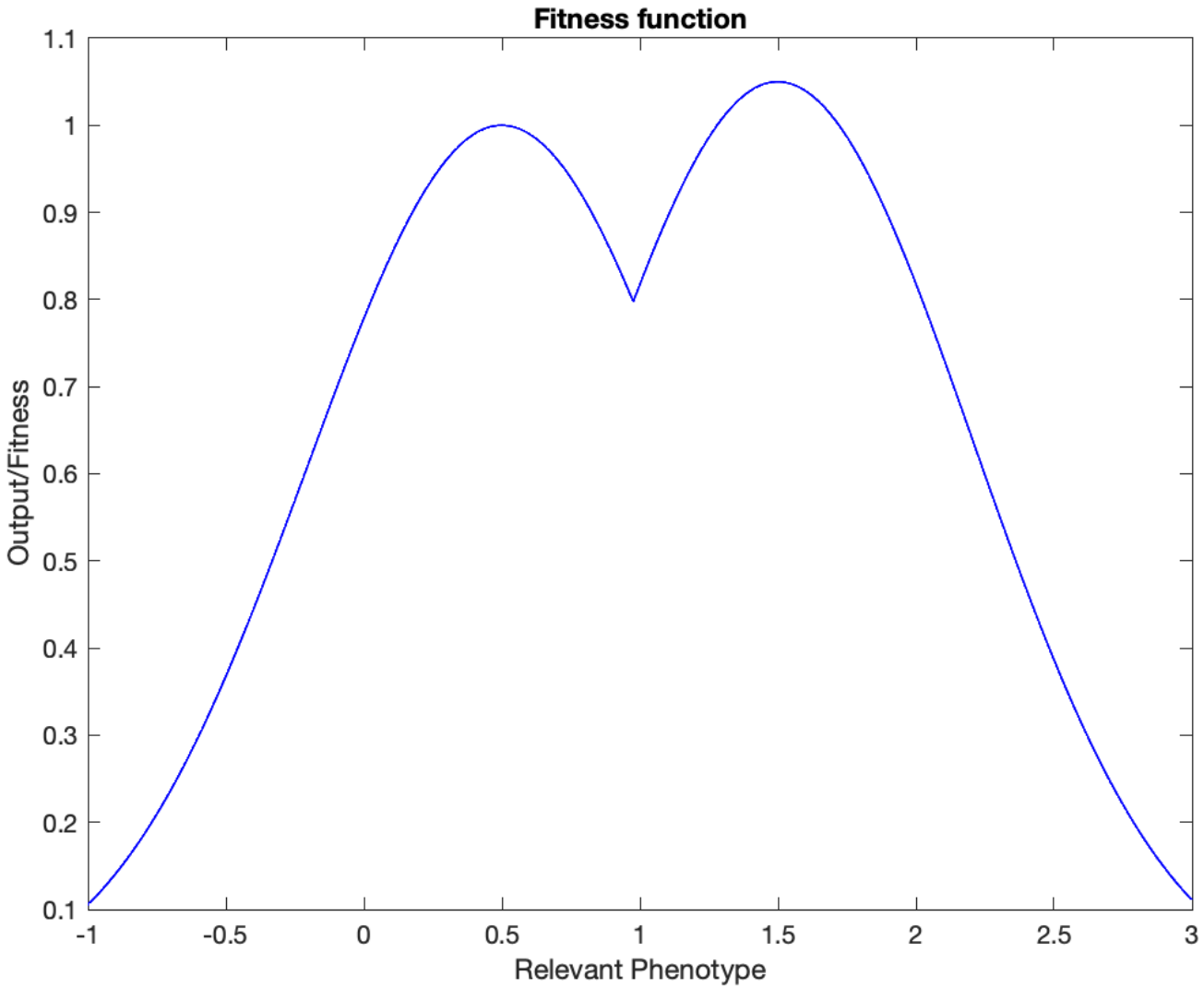
1. Same model of selection we have already introduced
2. Fitness of g as composition of
   - phenotype-map z from g to Z;
   - output-map from phenotype to output;
   - fitness-map from output to fitness
3. The two technologies and corresponding fitness functions **coexist**. Individuals chooses the best technology given their specific z;
4. Population growth leads to selection within a region and migration (differently from standard WF)

New Fitness Function Hypothesis

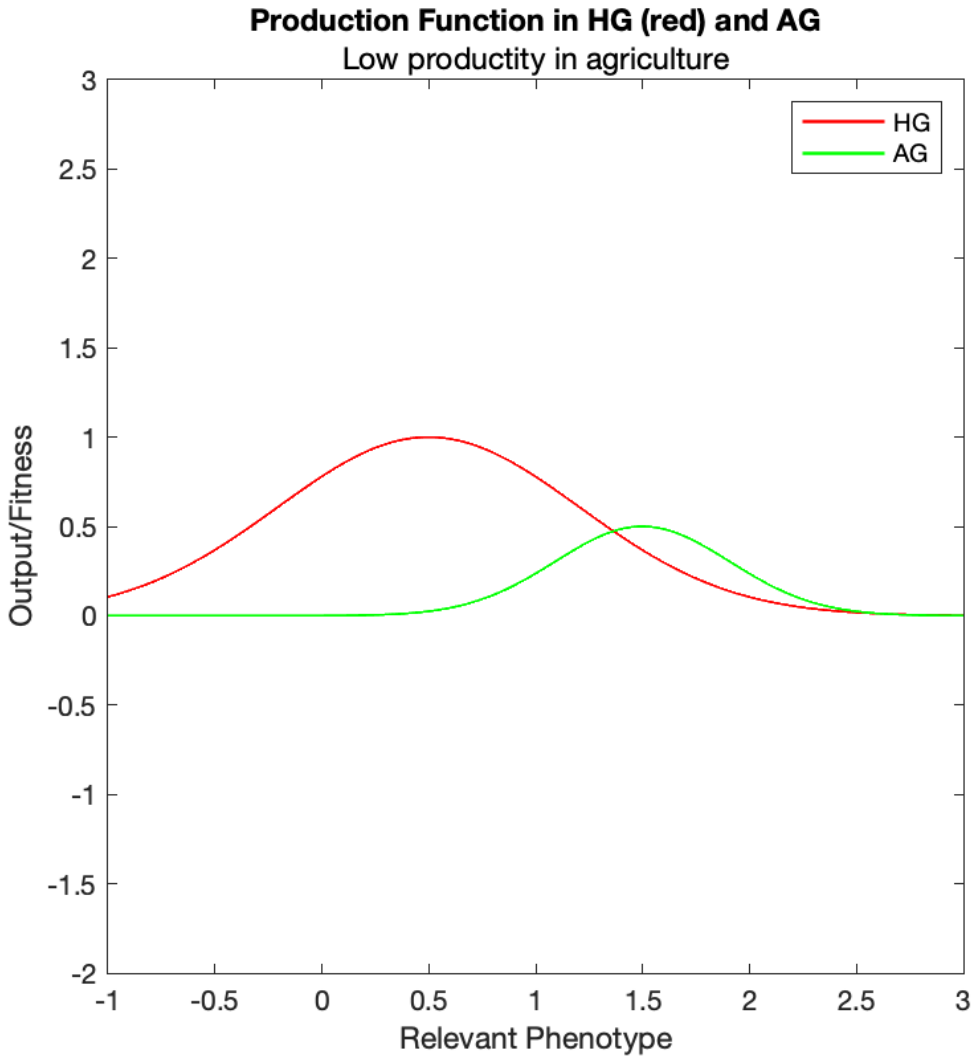
Choice between two technologies, Gausssian Fitness



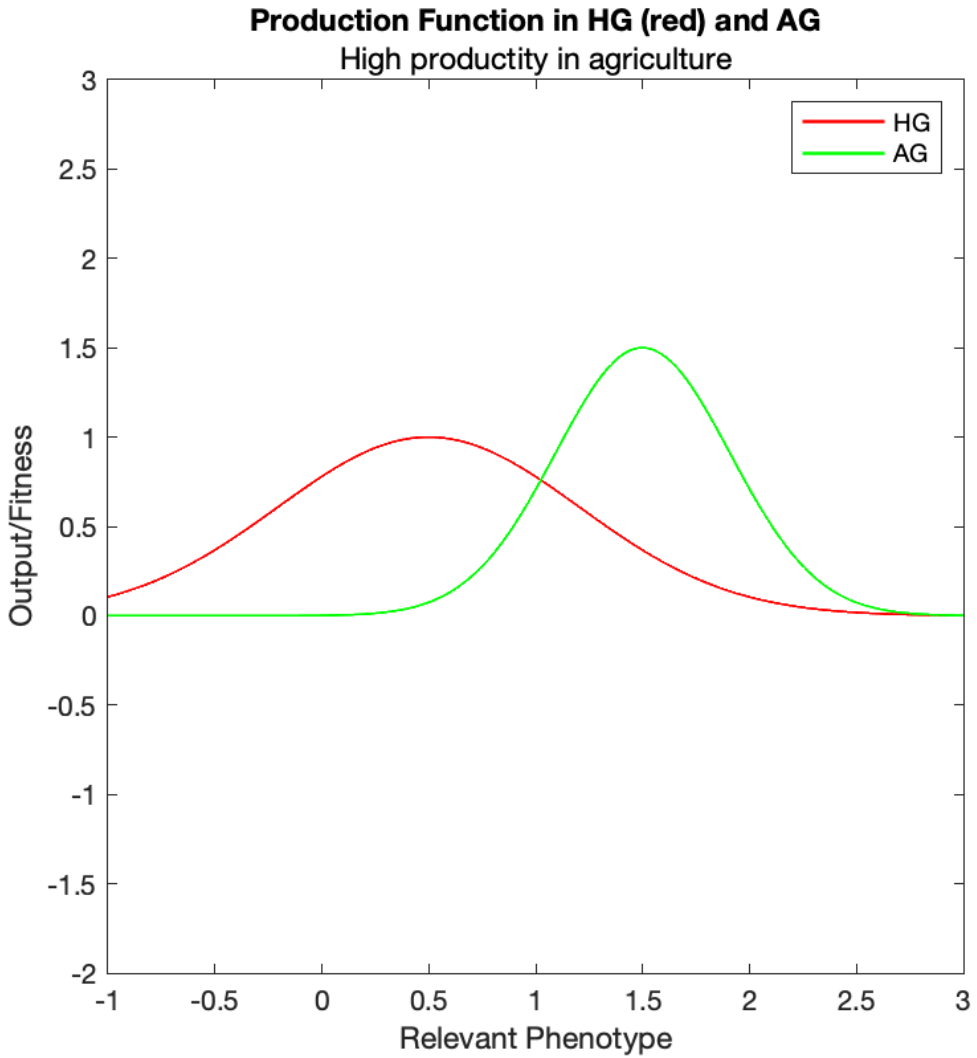
New Fitness Function Hypothesis, with Individually Rational Choice



Hunter-Gatherer (HG) technology and Agriculture (AG)

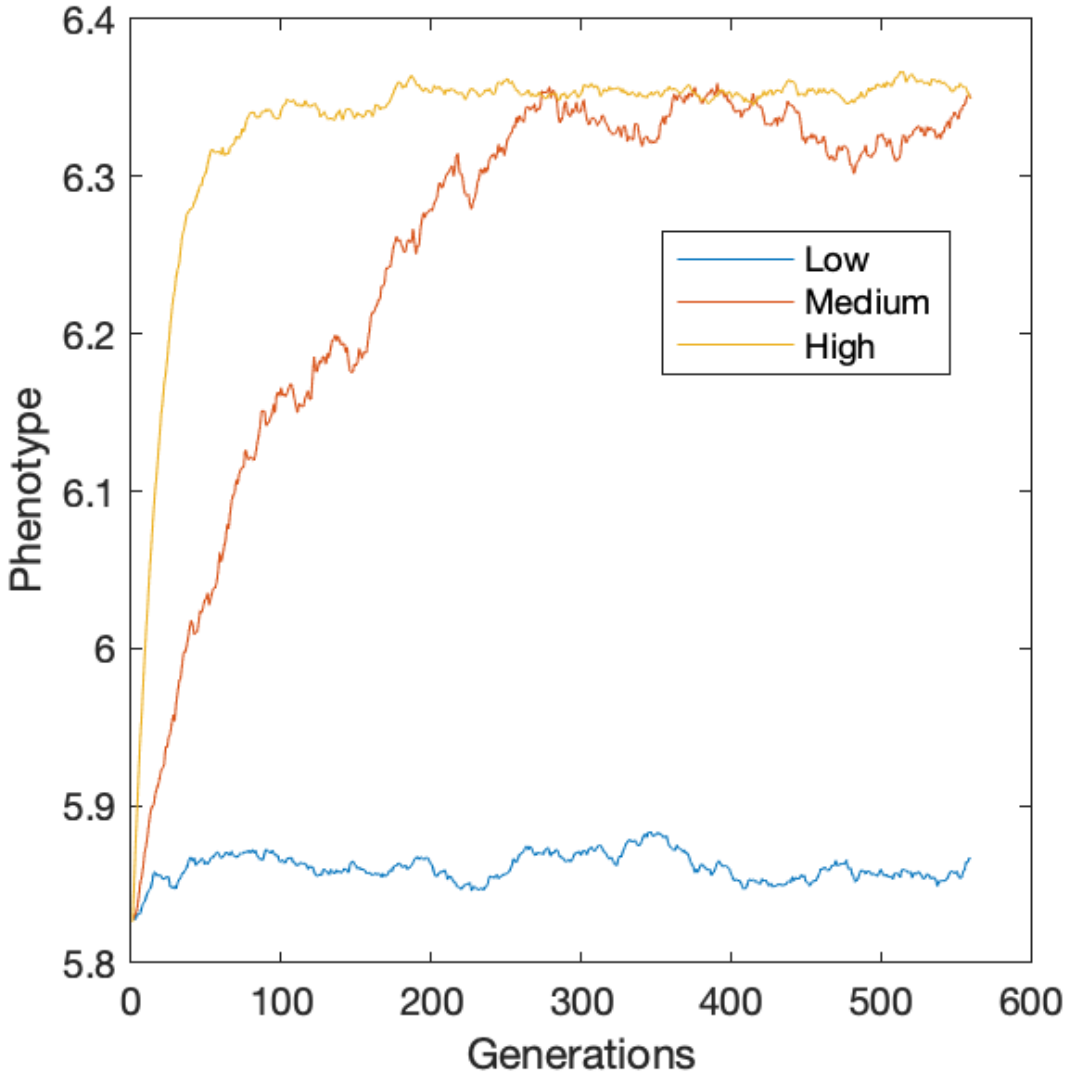


Hunter-Gatherer (HG) technology and Agriculture (AG)



Phenotype Paths Different Scenarios

Low, Medium and High Effect of Productivity in Agriculture



Test of the Hypothesis

Continuous Time Approximation

1. The full model can be approximated with simple continuous time model
2. We first present the idea in a simplified setup with set of genoptypes with two (2!) elements, then extend to our setup
3. We refer to this as simple model

Very Simple Case

Continuous Time Approximation

1. A population of M individuals that can be of type in $\{0, 1\}$ (this set replaces $\mathbb{G}$).
2. At time $i \in \{0, 1, \ldots, I\}$ the frequency of 1 is $p(i, M)$
3. The next population is also of size M. Each individual in the new population is chosen to be of type 1 with probability $p(i, M)$, but changed a little (selection).
4. Let the function (think of γ as $\gamma \equiv \omega\beta(k)$)

$$F(x, \gamma, M) \equiv \frac{e^{\frac{\gamma}{M}} x}{1 - x + e^{\frac{\gamma}{M}} x}$$

5. Let $\mathbf{bn}(q, M)$ denote the binomial with probability q of success and M draws. The next population has frequency

$$p(i + 1, M) = \frac{1}{M} \left(\mathbf{bn}(F(p(i, M), \gamma, M), M)\right)$$

Continuous Time Approximation

Very Simple Case

1. Let $dp(i, M) \equiv p(i+1, M) - p(i, M)$, so that:
2. $E(dp(i, M)) = \gamma \left(p(i, M)(1 - p(i, M))\right) \frac{1}{M}$
3. $Var(dp(i, M)) = \left(p(i, M)(1 - p(i, M))\right) \frac{1}{M}$
4. Let $\Delta t \equiv \frac{1}{M}$, define the continuous time extension on $[0, I\Delta t]$

$$x(t, M) \equiv p(\left\lfloor \frac{t}{\Delta t} \right\rfloor)$$

5. We get, if $Z \sim N(0, 1)$

$$dx(t, M) = \gamma \left(x(t, M)(1 - x(t, M))\right) \frac{1}{M} + \sqrt{x(t, M)(1 - x(t, M))} \frac{Z}{\sqrt{M}}, t \in [0, I]$$

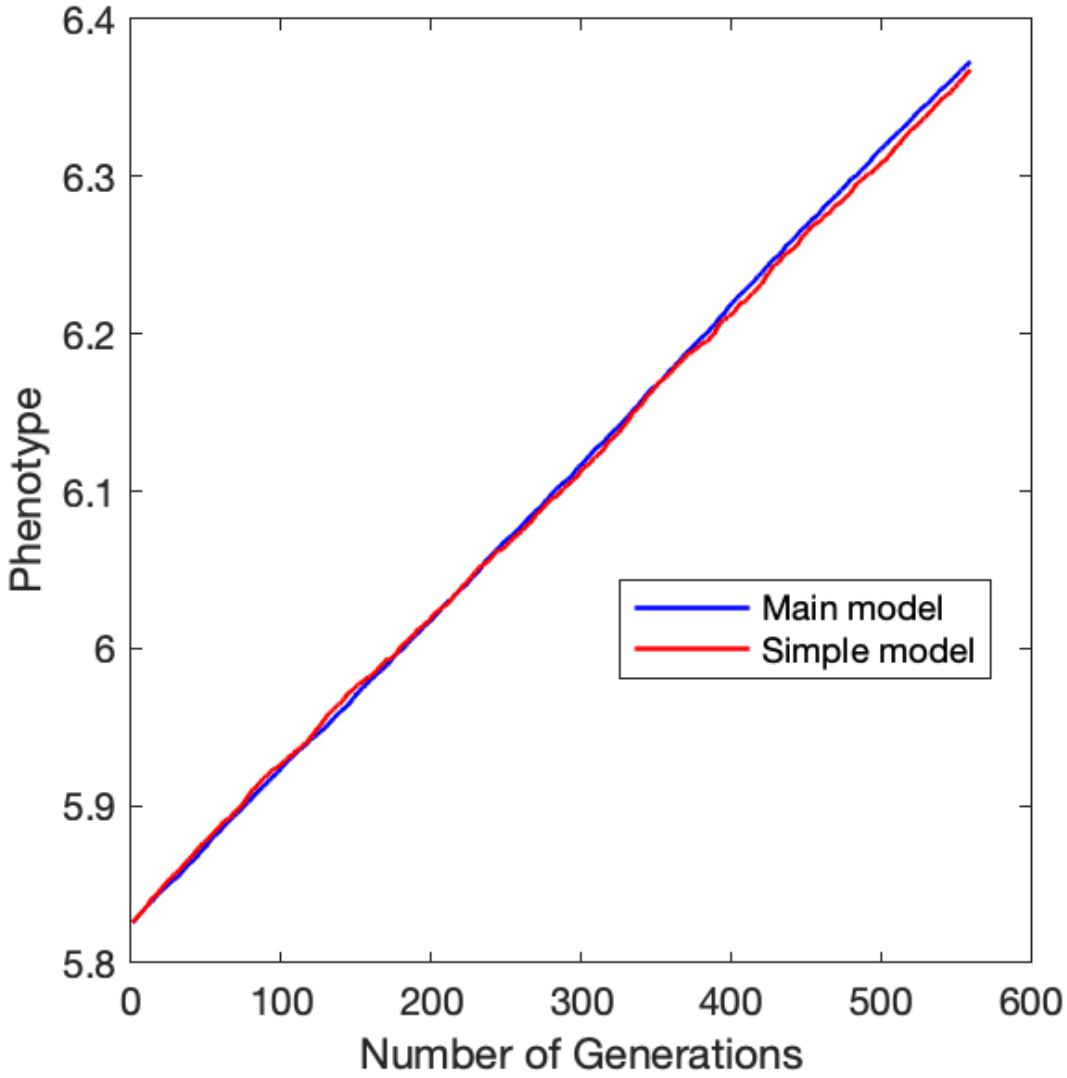
6. Approximating, with $W(t)$ standard Brownian motion at t:

$$d\xi(t) = \gamma \left(\xi(t)(1 - \xi(t))\right) dt + \sqrt{\xi(t)(1 - \xi(t))} dW(t),$$

for $t \in \left[0, \frac{I}{M}\right]$

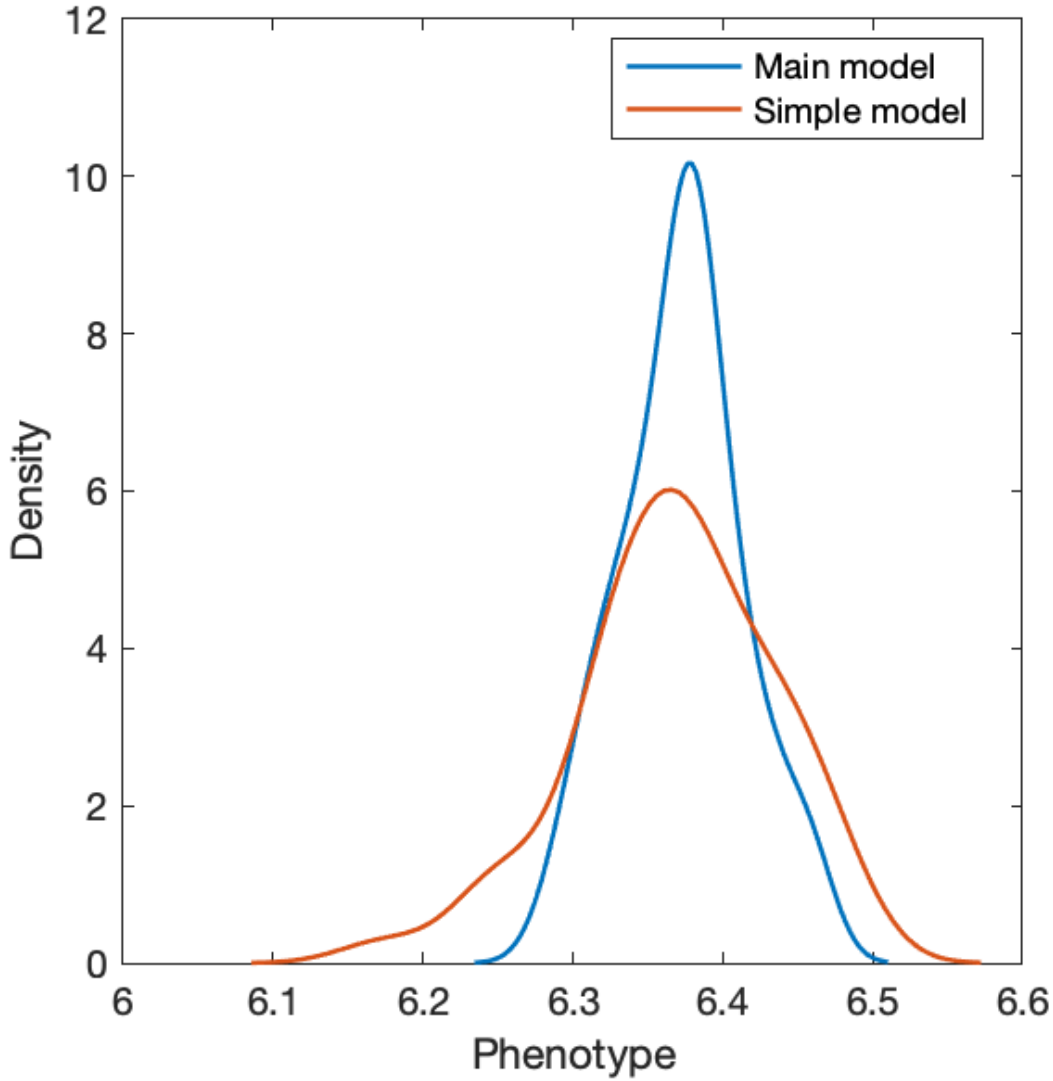
Comparison Full Model and Cts Time Approximation

Phenotype Path over 560 (=14,000/25) Generations



Comparison Full Model and Cts Time Approximation

Kernel Density of Phenotype in Final Generation



Reduction to a non-linear regression

1. Considering that the solution of :

$$d\xi(t) = \gamma\left(\xi(t)(1-\xi(t))\right)dt, \xi(0) \text{ given initial condition}$$

is

$$\xi(t) = \frac{e^{\gamma t}\xi(0)}{1-\xi(0)+e^{\gamma t}\xi(0)}$$

2. Adding a normal random variable, zero mean and variance $T\xi(T)(1-\xi(T))$ we can use Non-linear Least Squares to estimate the parameters (ω, T)

Estimate of Continuous Time Model

1. Parameters are (ω, T), ω strength of selection in the linear fitness model and T time horizon.
2. Use Non-linear Least Squares for parameter estimation

Maximum Likelihood Estimation of Fitness Parameter

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

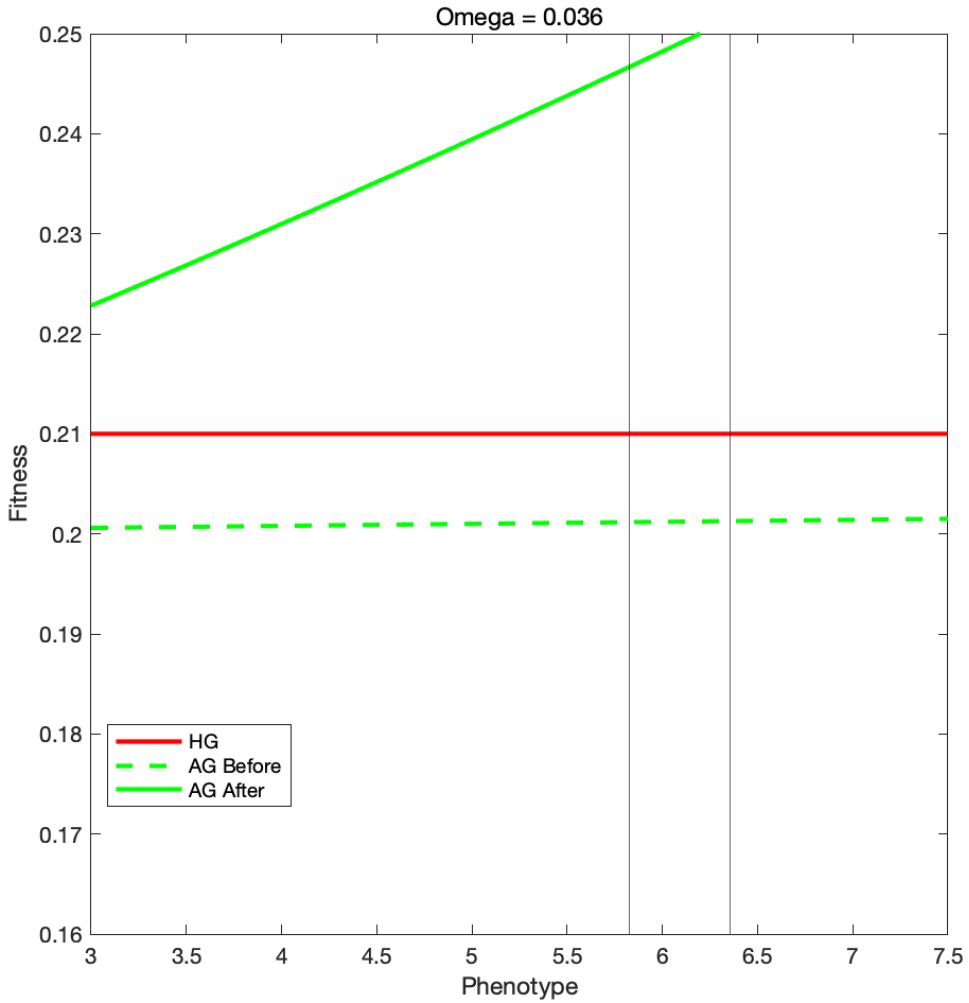
Estimation using the Full Model

1. We take the initial frequency, fix time horizon to 14K years/25 years, 560 generations
2. Estimate of the strength of fitness parameter, $\hat{\omega}$ provided by the continuous time approximation model as initial estimate
3. Run the full model with varying strength of fitness parameter ω, compute for each run the distance between the *allele frequencies predicted by full haplotype model*, with
   - given strength of fitness parameter,
   - standard parameters such as mutation rate, recombination locations and rates
   - random mating

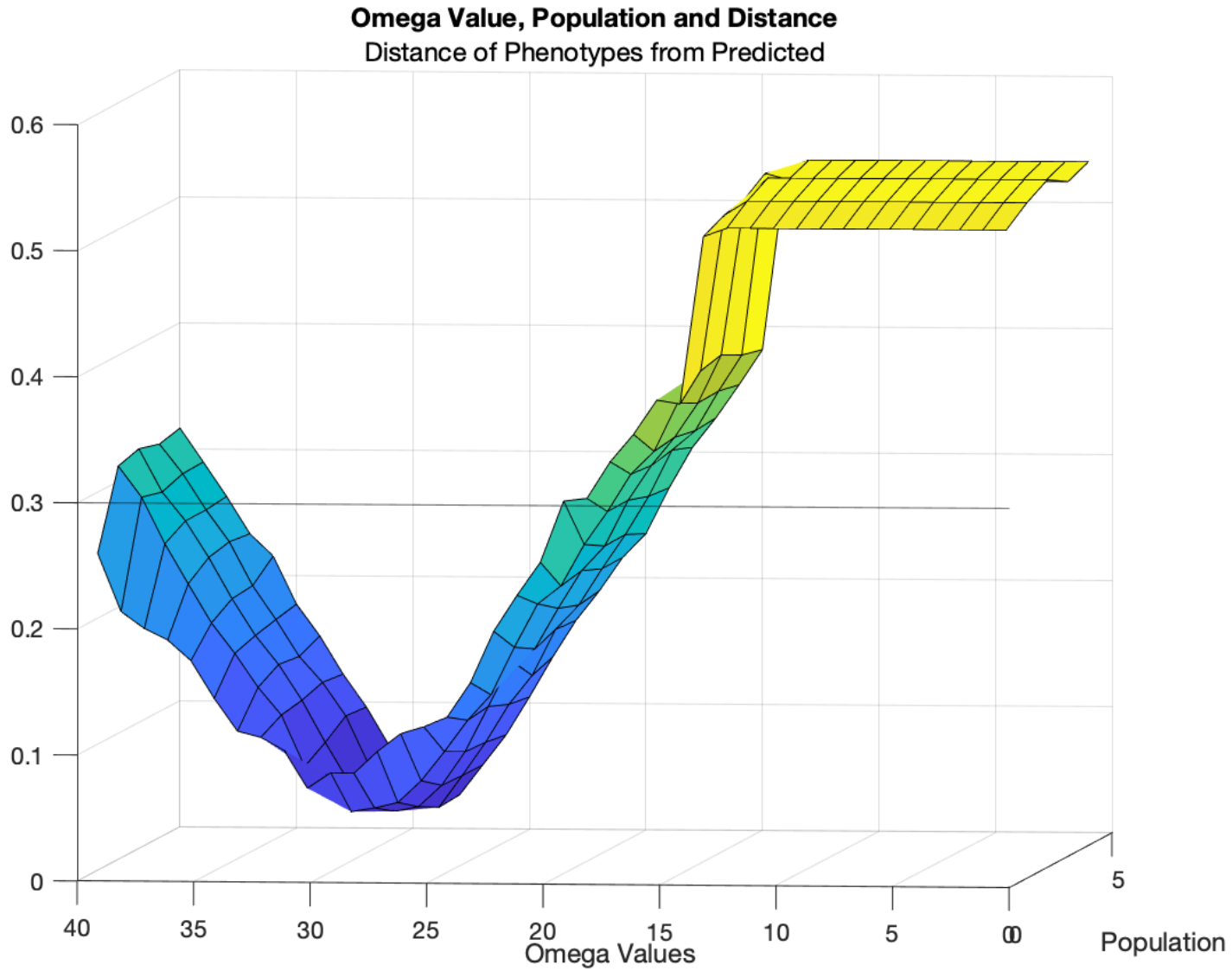
   and *current allele frequencies*
4. Minimize the distance between predicted and observed, use bootstrap to estimate standard errors
5. Do this for different population sizes, to analyze the effect of population size

Locally Linear Model

Two Technologies: *HG*, *AG* before, *AG* after.

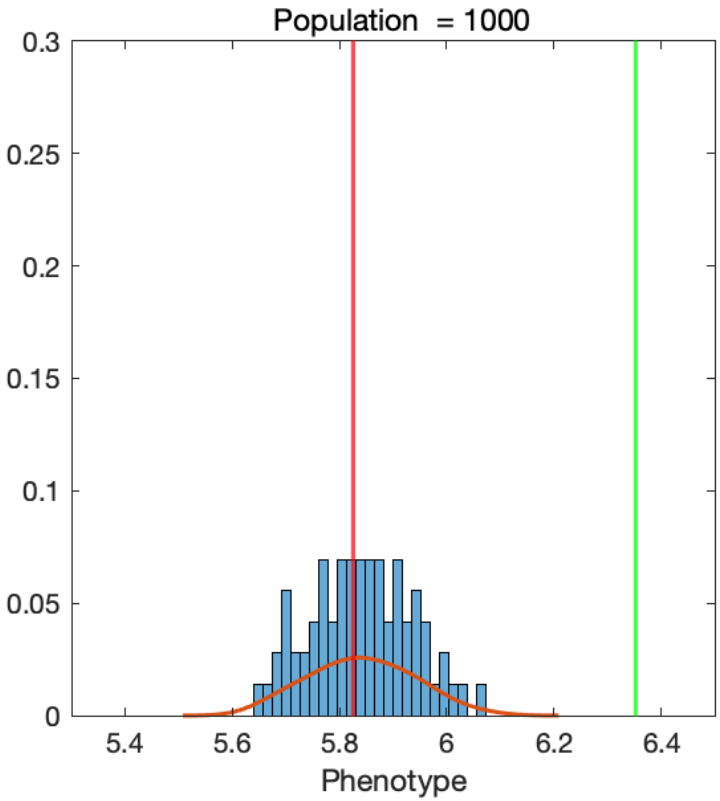


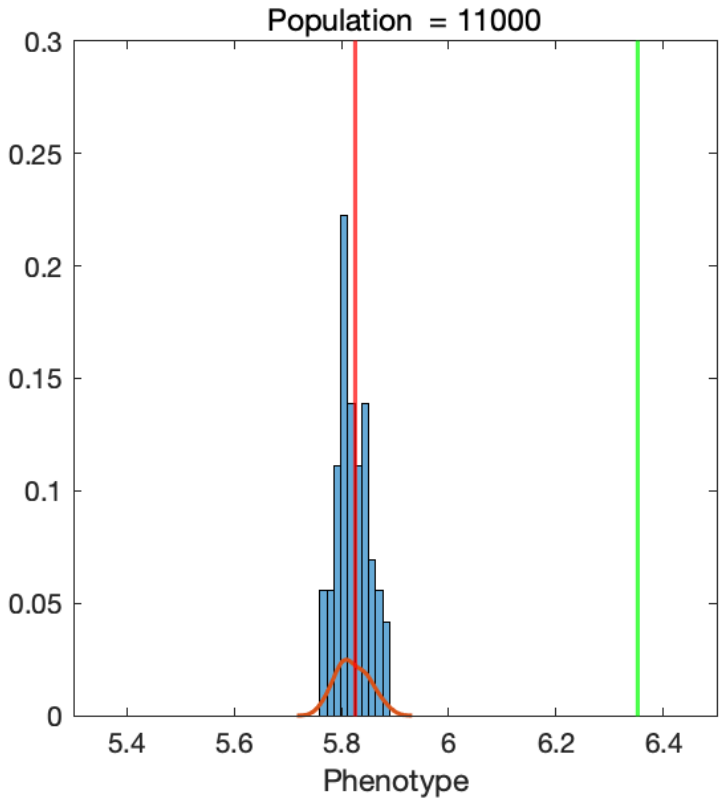
Distance from Final Phenotype



Null Hypothesis: $\omega = 0$

Small and Large Population Size

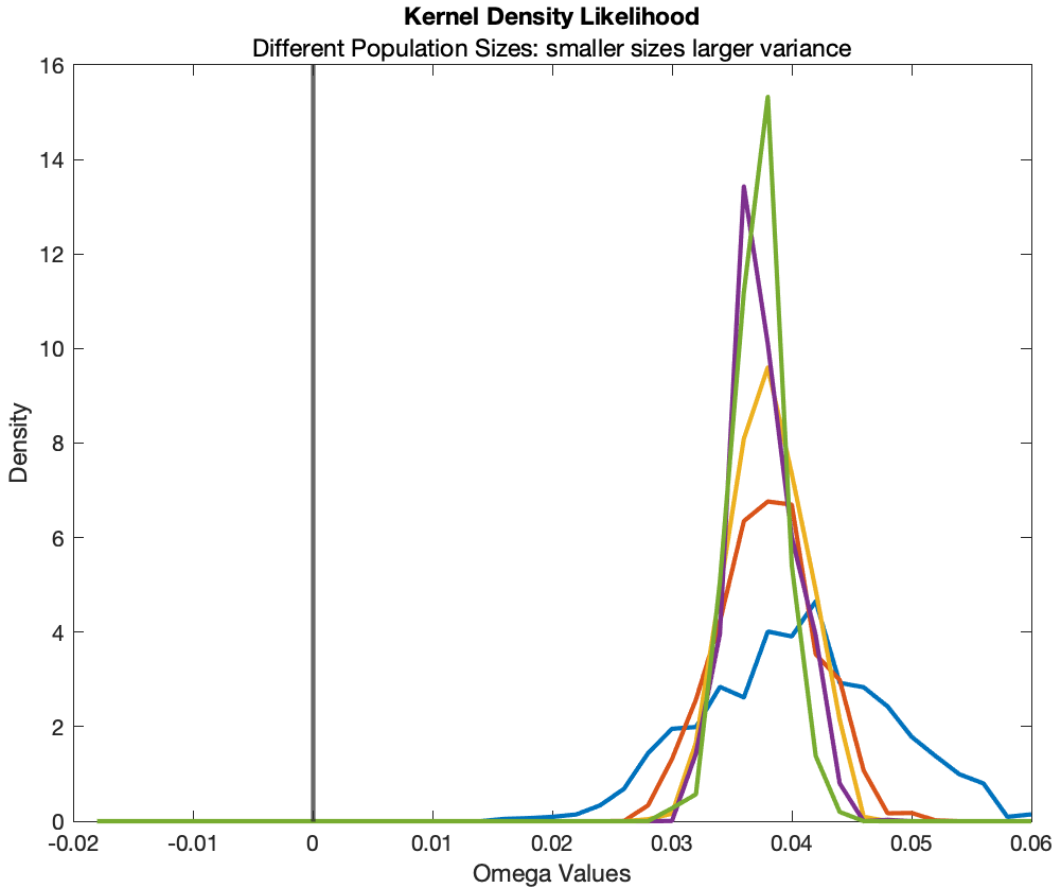




Likelihood Function Estimations

1. We derive the Likelihood Function (LF) for a given vector of parameters
2. The table reports estimation results for the parameters in the Wright-Fisher diffusion process (Simple Model)
3. Sample selection on the basis of the response variable: if an allele is fixated it is not observed (truncated regression)
4. The estimations under normal approximation use conventional algorithms and report conventional standard errors in parentheses. The estimations under numerical approximation use the grid search algorithm and report bootstrapped standard errors (10,000 repetitions) in parentheses.

Likelihood Function Estimations



Conclusion

Look for the ring in the right place.

1. There has been a selective pressure in the direction of EA enhancing alleles in the period following 14K YBP;
2. Time evolution of the distribution of genotype was produced in part by the change in productivity of two technologies due to climate change;
3. Speed of selection is in scale with observed phenomena
4. Proof in this case of a general relation between technology, genetics and institutions: each level determines the next.