Part III Project Report

Investigating Early Human Demography Using Sequenced Human DNA

Candidate 4635I
Supervisor: Dr Jamie Blundell

5,071 words

12 May 2019

INVESTIGATING EARLY HUMAN DEMOGRAPHY USING SEQUENCED HUMAN DNA

Candidate 4635I

ABSTRACT

The dominant model of human demographic history is the recent single-origin model, which holds that modern humans developed in a single region in Africa and subsequently expanded outwards into the rest of the world in a series of migrations. Evidence supporting this theory lies in human mitochondrial DNA and archaeological findings. Here, we develop a model that captures the key features governing the complicated dynamics of selectively neutral mutations. By examining the frequency distribution of mutations in the human population, known as the site frequency spectrum, we apply the model to a rich dataset of sequenced human genetic data to predict features of early human demography. We predict an expansion out of Africa around 144,000 years ago, a value that agrees with the literature. Additionally, when the model is applied to humans of African ethnicity our findings of a more recent population expansion match common knowledge of this population's history. The model is also validated using a computer simulation method that captures the essential dynamical processes described in our model.

CONTENTS

1. INTRODUCTION

Although human-like species are believed to have been globally present millions of years ago, with a recent study placing hominins in southern China 2.1 million years ago,[1] anatomically modern humans are thought to have developed more recently: in the region of a few hundred thousand years ago. Whether this evolution occurred in a single region or not, however, is a question that is somewhat unsettled.

The current prevailing[2] theory of modern human demographic history is the "recent single-origin" (RSO) model, which hypothesises that modern humans developed in eastern Africa and subsequently migrated out through a variety of routes[3] into the Middle East and Asia in a series of waves. This expansion, commonly named "Out of Africa II" to distinguish it from earlier hominin migrations out of the continent, is dated between 100,000 and 300,000 years ago.[4–7] The main alternative theory is the multi-regional evolution (MRE) hypothesis,[8,9] which holds that modern humans evolved globally within a dispersed population that included species such as *H. erectus* and *H. neanderthalensis*.

RSO predicts that all humans today, or at least those outside of Africa, are descended from a single east African population. This therefore allows for genetic tests of the hypothesis. The most recent common ancestors of all current living humans, through their maternal ancestry (known as "mitochondrial Eve") and through their paternal ancestry (known as "Y-chromosomal Adam") are currently dated at around 150,000 years ago[10,11] for Eve and between 200,000 and 300,000 years ago[12,13] for Adam. The first dating of mitochondrial Eve at between 140,000 and 200,000 years ago in 1987[14] was seen by many as the deciding vote in what had been a fierce debate[15,16] between RSO and MRE proponents.

RSO is now the most popular theory[17] and MRE has shifted closer to RSO,[18] allowing a greater role for Africa in the human origin story. Additionally, fossil evidence[19] has lent further credence to RSO. However, there are occasional findings that do cast doubt, such as Templeton's 2002 study[20] that observed modern genetic variants present in Asian populations well before "Out of Africa II". Indeed, some of the original MRE proponents still support their theory today.[21]

A useful measure of the global genetic variation is the *site frequency spectrum* (SFS), which is simply the distribution of mutation frequencies in a given population. In this project, we use a simple model of the dynamics of selectively neutral mutations in humans to link features in the SFS to demographic history. Today, the freely available database gnomAD[22,23] allows access to sequenced genetic data of over a hundred thousand individuals. We make use of this rich dataset to generate an SFS of neutral mutations and then apply our theory to examine its plausibility in view of RSO. Specifically, we look at what demographic expansion time is most likely based on our theory and the gnomAD-derived SFS, comparing it to the believed time of "Out of Africa II". We also apply our model to genetic data from those of African ethnicity, to see whether their ancestral demographic history may have been different.

2. THEORY

Darwinian natural selection is a key mechanism through which evolutionary adaptation occurs: mutations which positively impact the reproductive fitness of an individual will become more prevalent in the population over time. In the vast majority of cases, mutations are binary: that is, for any given site on the genome there are only two variants observed in the population. In a given population, the frequency f of a mutation evolves under a variety of mechanisms, including:

1. *Mutation*: the given site mutates to or from the variant in question.
2. *Positive selection*: mutations which positively impact the reproductive fitness of an individual will become more prevalent through Darwinian natural selection. Advantageous mutations may increase in frequency until they *fix* in the population; this is where the given site no longer exhibits multiple variants and the initial mutation is exhibited in the genomes of the entire population.

3. *Negative selection*: deleterious mutations that arise in the population are eliminated through natural selection.
4. *Genetic drift*: stochastic frequency variation due to variances in the amount of offspring of individuals in the population, independent of any impact of the mutation on the observable characteristics of the individual.

Any change in frequency through genetic drift or mutation is on a much slower timescale than changes through positive or negative selection. Here we consider selectively *neutral* mutations, that is, those whose dynamics are not directly affected by positive or negative selection.

Although the dynamics of mutations is highly complex, it is possible to construct a simple model describing the time evolution of the frequency f of a neutral mutation. We first make the strong simplifying assumption that each site on the genome is independent of the others, allowing us to consider sites individually. Consider a site with current variant `A`, which mutates to or from an alternate variant `B` at a rate μ. Say the frequency of people who carry `A` is f. Then the change in f has two contributions: an increase from sites mutating from `B` to `A` and a decrease from sites mutating from `A` to `B`. Both occur at a rate μ:

$$\begin{aligned}\dot{f}_{\text{mut}} &= \underbrace{(1-f)\mu}_{\texttt{B}\to\texttt{A}} + \underbrace{(-f)\,\mu}_{\texttt{A}\to\texttt{B}} \\ &= (1-2f)\,\mu\end{aligned}$$

where $\dot{f}_{\text{mut}}$ indicates the rate of change of f due to mutation alone. There is a variety of models for the stochastic process of drift, which is essentially a sampling process (see section 3.1). Generally, the changes in f due to drift carry a $\frac{1}{\sqrt{N}}$ dependence, characteristic of sampling noise. As would be expected from the nature of this process, these models are mostly discretised in time in steps of generations. However, continuous models do exist, such as the continuous-time Wright-Fisher model:

$$\dot{f}_{\text{drift}} = \xi\sqrt{\frac{f(1-f)}{N}}$$

where $\dot{f}_{\text{drift}}$ indicates the rate of change of f due to drift alone and ξ is Gaussian white noise. This leads to the following model describing the time evolution of f:

$$\dot{f} = \underbrace{\mu(1-2f)}_{\text{mutation}} + \underbrace{\xi\sqrt{\frac{f(1-f)}{N}}}_{\text{drift}} \tag{2.1}$$

The terms in equation (2.1) can be understood as follows (see Figure 1):

- Mutation: a "deterministic force" that tends to push f towards 0.5 - as there is no advantage to either of the variants, $f = 0.5$ is a natural stability point. Though mutation is inherently a stochastic process, its nature is such that smoothing occurs for a population of any appreciable size and so can be modelled as deterministic. Changes in frequency due to mutation occur on a timescale of $\frac{1}{\mu}$.
- Drift: small-scale, stochastic variation in f. The effect is strongest around $f = 0.5$, but its amplitude is mainly controlled by the population size N: a larger population leads to a dampening of the effects of random variations in offspring number. Changes in frequency due to drift occur on a timescale of N.

2.1. *Theoretical expressions for the site frequency spectrum*

We can form a theoretical expression for the SFS by adding up the contributions of mutations that enter the population at each time. At time t, the distribution of sizes n of mutations that entered the

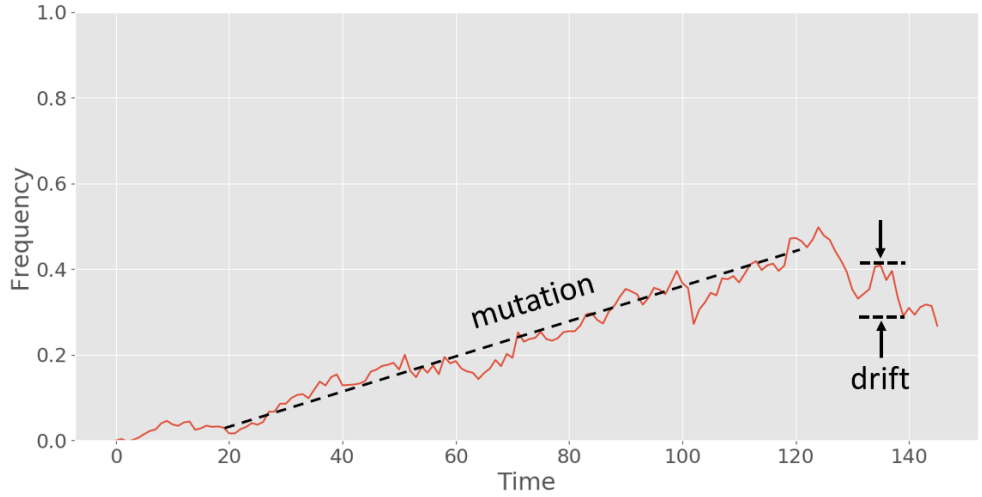


Figure 1: A rough sketch indicating how mutation and drift affect the frequency of a neutral mutation. Note that for clarity in this figure, to make the deterministic nature of mutation more clear, we have chosen a regime that is mutation-dominated, i.e. $N\mu \gg 1$.

population at time T is[24]

$$\rho(n|T) = \frac{\exp\left[s\left(t-T\right)\right]}{\tilde{n}^2} \cdot \exp\left[-\frac{n}{\tilde{n}}\right] \tag{2.2}$$

where s is the *selection coefficient*. $s > 0$ for fitness-enhancing mutations and $s < 0$ for mutations detrimental to fitness. $\tilde{n}$ is given by

$$\tilde{n} = \frac{\exp\left[s\left(t-T\right)\right] - 1}{s} \tag{2.3}$$

For neutral sites, $s \to 0$. Therefore we seek

$$\lim_{s\to 0} \rho(n|T) = \lim_{s\to 0} \left[\underbrace{\frac{\exp\left[s\left(t-T\right)\right]}{\tilde{n}^2}}_{①} \cdot \underbrace{\exp\left[-\frac{n}{\tilde{n}}\right]}_{②} \right] \tag{2.4}$$

Firstly consider the term that is labelled ① in equation (2.4). We have

$$\begin{aligned} ① &= \lim_{s\to 0} \left[\frac{\exp\left[s\left(t-T\right)\right]}{\left(\frac{\exp[s(t-T)]-1}{s}\right)^2} \right] \\ &= \lim_{s\to 0} \left[\frac{s^2 \exp\left[s\left(t-T\right)\right]}{\left(\exp\left[s\left(t-T\right)\right]-1\right)^2} \right] \end{aligned}$$

Performing a Taylor expansion about $s = 0$ to first order in s, we have

$$\begin{aligned}
(1) &\approx \lim_{s\to 0}\left[\frac{s^2\left[1+s(t-T)+\mathcal{O}(s^2)\right]}{\left(s(t-T)+\mathcal{O}(s^2)\right)^2}\right] \\
&\approx \lim_{s\to 0}\left[\frac{s^2+s^3(t-T)}{s^2(t-T)^2}\right] \\
&\approx \lim_{s\to 0}\left[\frac{1}{(t-T)^2}+\frac{s}{(t-T)}\right] \\
&\approx \frac{1}{(t-T)^2}
\end{aligned} \tag{2.5}$$

Now consider the second term, labelled (2) in equation (2.4). We have

$$(2) = \lim_{s\to 0}\left[\exp\left(-\frac{n}{\frac{\exp[s(t-T)]-1}{s}}\right)\right]$$

We again use a Taylor expansion about $s = 0$:

$$\begin{aligned}
(2) &\approx \lim_{s\to 0}\left[\exp\left(-\frac{n}{\frac{s(t-T)}{s}}\right)\right] \\
&\approx \exp\left[-\frac{n}{t-T}\right]
\end{aligned} \tag{2.6}$$

Therefore, substituting equations (2.5) and (2.6) into equation (2.4), we obtain the neutral sites version of equation (2.2):

$$\rho(n|T) = \frac{1}{(t-T)^2}\exp\left[-\frac{n}{t-T}\right] \tag{2.7}$$

We can use this expression to obtain the distribution of sizes of all mutations by integrating over T:

$$\rho(n) = \int_0^t \rho(n|T)\cdot(N\mu)\,\mathrm{d}T \tag{2.8}$$

where $(N\mu)\,\mathrm{d}T$ is the number of mutations entering the population between times T and $T + \mathrm{d}T$. Inserting equation (2.7) we have

$$\rho(n) = \int_0^t \frac{1}{(t-T)^2}\exp\left[-\frac{n}{t-T}\right]\cdot(N\mu)\,\mathrm{d}T \tag{2.9}$$

2.1.1. *Constant population size*

If the population size N is constant then equation (2.9) has an analytical solution.

$$\begin{aligned}
\rho(n) &= \int_0^t \frac{1}{(t-T)^2}\exp\left[-\frac{n}{t-T}\right]\cdot(N\mu)\,\mathrm{d}T \\
&= N\mu\int_0^t \frac{1}{(t-T)^2}\exp\left[-\frac{n}{t-T}\right]\mathrm{d}T
\end{aligned}$$

Making the substitution $x = t - T$ we obtain

$$\begin{aligned}
\rho(n) &= N\mu\int_0^t \frac{1}{x^2}e^{-\frac{n}{x}}\,\mathrm{d}x \\
&= \frac{N\mu}{n}e^{-\frac{n}{t}}
\end{aligned} \tag{2.10}$$

The SFS is a distribution in frequency, so transforming to $f = \frac{n}{N}$ we have

$$\begin{aligned}\rho(f) &= \left|\frac{\mathrm{d}n}{\mathrm{d}f}\right| \rho(n) \\ &= N\rho(n) \\ &= \frac{N\mu}{f} \exp\left[-\frac{f}{\left(\frac{t}{N}\right)}\right]\end{aligned} \tag{2.11}$$

It is instructive to examine $\rho(\ln f)$ as when plotted it is easier to examine lower frequency regions: as we shall see, this is where most of the interesting features we wish to describe with our theory lie. Therefore we calculate $\rho(\ln f)$:

$$\begin{aligned}\rho(\ln f) &= \underbrace{\left|\frac{\mathrm{d}f}{\mathrm{d}(\ln f)}\right|}_{f} \rho(f) \\ &= N\mu \exp\left[-\frac{f}{\left(\frac{t}{N}\right)}\right]\end{aligned} \tag{2.12}$$

We will also plot this on a logarithmic scale, so we calculate

$$\ln\left[\rho(l)\right] = \ln N\mu - \frac{N}{t}e^{l} \tag{2.13}$$

where we define $l = \ln f$ for clarity. We see that this looks like a flat SFS with an exponential cut-off at high frequencies. We understand the dynamics as a constant feeding process with a maximum frequency that mutations can reach.

2.1.2. *Changing population size*

The human population size has clearly not been constant so if we wish to obtain an expression for the SFS that matches observed data then we must use a variable N in equation (2.9). The simple model of population growth we use here is

$$N(t) = \begin{cases} N_0 & t \leq t_b \\ N_0 e^{\frac{t-t_b}{t_g}} & t > t_b \end{cases} \tag{2.14}$$

where N_0 is the initial population size, t_b is the bottleneck time and t_g is the exponential growth lifetime, i.e. the time taken for the population to grow by a factor of e. We use this model as it has a clear connection to RSO: under the RSO hypothesis, a constant-sized population that then expands when migration out of Africa occurs is plausible. It is also a simple and tractable model. It is important to note that while there are seemingly three parameters (N_0, t_b and t_g), t is an additional parameter: the length of time for the entire process is not set.

Equation (2.9) does not have a closed-form solution for the growth model in equation (2.14). Instead we obtain the integral expressions

$$\rho(f) = \frac{\mu N_0}{f}\left[e^{-\frac{Nf}{t}} - e^{-\frac{Nf}{t-t_b}}\right] + \mu N^2 \int_0^{t-t_b} \frac{1}{x^2} e^{-\frac{Nf}{x}} e^{-\frac{x}{t_g}}\,\mathrm{d}x \tag{2.15}$$

$$\rho(\ln f) = \mu N_0\left[e^{-\frac{Nf}{t}} - e^{-\frac{Nf}{t-t_b}}\right] + f\mu N^2 \int_0^{t-t_b} \frac{1}{x^2} e^{-\frac{Nf}{x}} e^{-\frac{x}{t_g}}\,\mathrm{d}x \tag{2.16}$$

These expressions are only valid for times $t > t_b$. As the derivation for these expressions is very similar to that in section 2.1.1, we do not provide it in this section but instead include it in Appendix A for reference. An attempt was also made at producing an approximation to these integrals via the method of steepest descent; this achieved limited success and is detailed in Appendix B.

3. MATERIALS AND METHODS

3.1. *Simulation*

Dynamics of neutral sites as described by the theory in section 2 can be simulated numerically. This is a useful method to check consistency of the theory and also to observe the variation of the SFS over time. We seek a recursive relation that allows us to calculate the frequency f_i in generation i given the frequency f_{i-1} in the previous generation. We do this by calculating the change in frequency at each generation due to both mutation and drift.

For mutation, from equation (2.1) we know that the rate of change of frequency due to mutation is $\mu(1-2f)$. Working in units of time where $\Delta t = 1$ indicates the time between subsequent generations, we obtain an expression for the change in frequency due to mutation between generations:

$$\Delta f_{\text{mut}} = \mu(1 - 2f_{i-1}) \tag{3.1}$$

Now we consider drift. In the Wright-Fisher drift model,[25] the population n_i carrying the mutation in the i^{th} generation is a random sample from a binomial distribution with N trials and a mean equal to the mutation frequency in the previous generation f_{i-1}. Essentially, we flip a biased coin for each person in the new generation to determine if they carry the mutation, the bias of which is determined by the frequency in the preceding generation.

$$n_i \sim \text{Binomial}\left(N, \underbrace{\frac{n_{i-1}}{N}}_{f_{i-1}}\right) \tag{3.2}$$

Therefore the change in frequency due to drift is

$$\begin{aligned}\Delta f_{\text{drift}} &= \frac{n_i}{N} - \frac{n_{i-1}}{N} \\ &= \frac{1}{N}\text{Binomial}\left(N, f_{i-1}\right) - f_{i-1}\end{aligned} \tag{3.3}$$

Combining these contributions from equations (3.1) and (3.3), the recursive relation is therefore

$$\begin{aligned}f_i &= f_{i-1} + \Delta f_{\text{mut}} + \Delta f_{\text{drift}} \\ &= \mu(1 - 2f_{i-1}) + \frac{1}{N}\text{Binomial}\left(N, f_{i-1}\right)\end{aligned} \tag{3.4}$$

If we add in a variable total population size, then we modify equation (3.4) as

$$f_i = \mu(1 - 2f_{i-1}) + \frac{1}{N_i}\text{Binomial}\left(N_i, f_{i-1}\right) \tag{3.5}$$

Note that the binomial distribution has N_i trials but probability f_{i-1} as we have a population N_i, each individual of which is assumed to have a probability f_{i-1} of carrying the mutation.

The simple relation (3.5) was used to simulate 500,000 sites. The histogram of the f_i for all sites gives the SFS at the i^{th} generation. Parallelisation was used to reduce simulation times such that run times for cases examined in this report were around eight hours. Further optimisation may have been possible but was deemed unimportant: the goal for the simulation was to make sure the simple model was understood and that a simulation produced the results consistent with the theoretical integral expressions.

Figure 2 illustrates the development of the SFS over time through such a simulation, with theoretical curves generated from equation (2.13) overlaid to show the agreement with theory. In accordance with

known[26] values for humans, we set $\mu = 2.5\times10^{-8}$. The value of N was set to 4000; this value is justified in section 4.1 as the predicted size of the human population before "Out of Africa II" expansion in our theory. Therefore we have $N\mu \ll 1$ and drift dominates the dynamics; mutation acts as a feeding process that sets the overall number of mutations in the population. The key point to note here is that after a sufficient number of generations, the SFS looks largely flat at high frequencies. The simulation results show good agreement with the theoretical curves.

3.2. *gnomAD*

The **gnom**e **A**ggregation **D**atabase (gnomAD)[22,23] is a freely available resource collating sequencing data from a range of sources. The database is rich, containing 125,748 exome sequences. *Exomes* are sequences of all the genome's *exons*; the name exon is derived from "expressed region" - therefore, the exome can be roughly thought of as the "coding portion" of the genome. To obtain an SFS for neutral sites from gnomAD an understanding of the different types of mutation is required.

Synthesis of proteins from DNA occurs by reading out base nucleotides to form tRNA, a process known as *transcription*, and subsequently *translating* the tRNA into proteins. The DNA sequence is subdivided into small sequences of three base pairs, known as *codons*. Each codon corresponds to an amino acid that makes up the protein chain. Therefore it is in large part the mutation's effect on the codon that determines its functional consequence. As there are four possible base pairs, there are $4^3 = 64$ possible codons, three of which do not correspond to an amino acid, but signal a termination point and are known as *stop codons*. In contrast, there are 20 possible amino acids – therefore there is a roughly threefold degeneracy in the correspondence between codons and amino acids (61:20). Three important mutation types can be classified as follows:

1. *Missense* mutation: a mutation that changes the codon to one that corresponds to a different amino acid.
2. *Synonymous* mutation: a mutation that changes the codon to one that corresponds to the same amino acid.
3. *Nonsense* mutation: a mutation that changes the codon to a stop codon. This results in premature termination of the relevant protein and the product is therefore often non-functional. Conditions such as thalassaemia and cystic fibrosis have been linked with nonsense mutations.[27,28]

In this report we consider synonymous mutations, using their SFS for comparison with our neutral sites theory. It is important to note, however, that synonymous mutations are not entirely selectively neutral. Though different codons may code for the same amino acid, they will produce different tRNA when transcribed, which for example can have effects on the rate of translation. Indeed, for any given amino acid we do not see all corresponding codons at the same frequency, indicating some may be selectively advantageous. However, the synonymous mutation type is the most selectively neutral type available in gnomAD in sufficient numbers for a high quality SFS, so we consider it in our analysis here.

The site frequency spectrum for synonymous mutations derived from gnomAD is shown in Figure 3. We observe sharp oscillations at the lowest frequencies. This is due to the fact that, in the experimental gnomAD data, we count frequencies by dividing the number of people with the given mutation in the population by the population size. Hence the frequency only exists as multiples of 1/(population size). At low frequencies this discreteness therefore becomes more apparent as we are binning after taking the log of the data – therefore data at low frequencies is more "spread out". Hence we observe these oscillations from counting mutations that appear in one person, two people, three people etc.

The main point, however, is that the synonymous SFS differs hugely in shape from the stable SFS in section 3.1. Although we do see a flat tail, similar to that in section 3.1, there is a strong skew towards lower frequencies. Demography could be the reason for this: in a growing population, more individuals can acquire mutations within a single generation; this would increase the number of low-

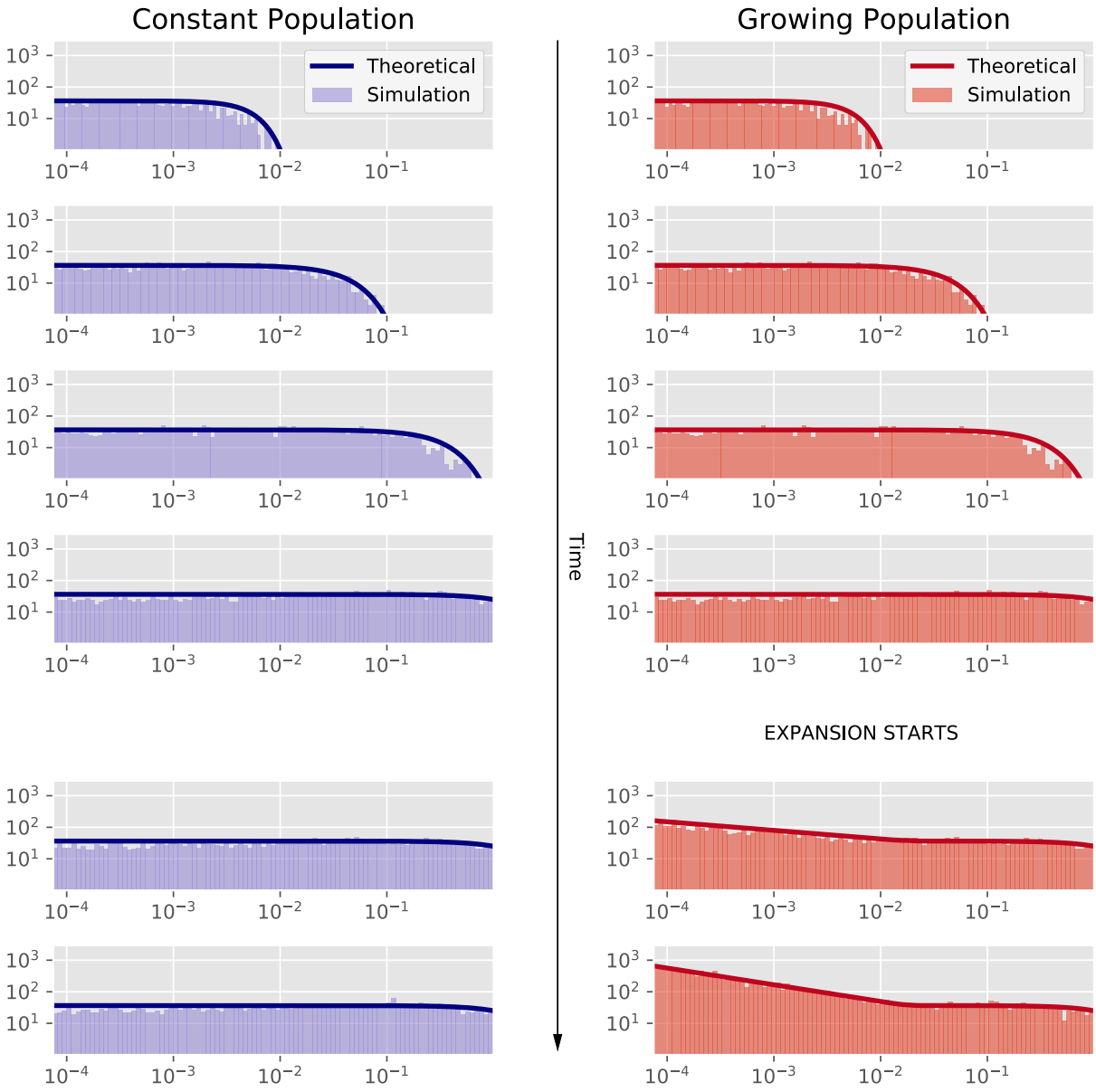


Figure 2: Development of the SFS over time with $N_0 = 4000$, $\mu = 2.5 \times 10^{-8}$. Each subplot measures mutation frequency on the horizontal axis and the number of mutations on the vertical axis; labels are omitted for clarity. The left column indicates results (light blue histograms) for a simulation witha constant population size; good agreement is shown with the dark blue theoretical curves generated by equation (2.13). The right column indicates results (light red histograms) for simulation of a population that follows growth model (2.14), with the lower two subplots showing the SFS for $t > t_b$. There is a clear increase in the count of low-frequency mutations above the prediction for a constant population, in line with the discussion at the end of section 3.2. Good agreement is shown with dark red theoretical curves generated by equation (2.15).

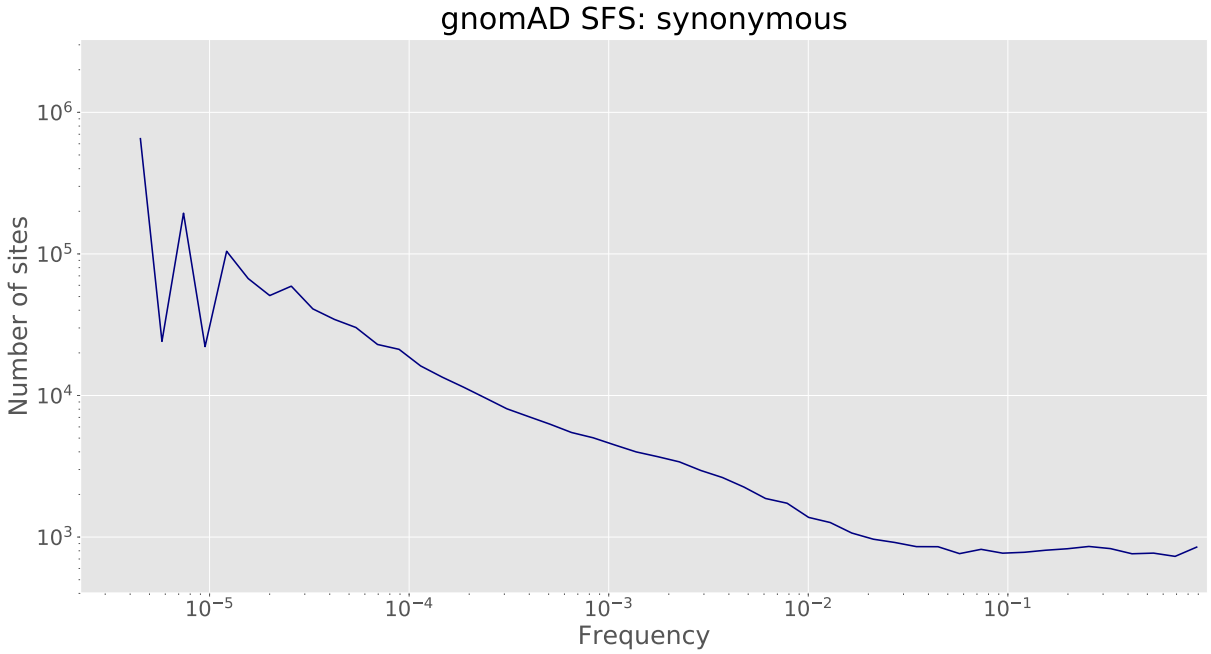


Figure 3: gnomAD-derived site frequency spectrum for synonymous mutations.

frequency variants which then become further diluted as the population continues to grow. This was investigated by performing a simulation as described in section 3.1, and the prediction of an uptick at lower frequencies is exhibited in the results, as shown in Figure 2. It is by examining this uptick that we will be able to constrain the parameter values for the growth model in equation (2.14) and thereby make predictions about historic human demography.

3.3. *Extraction of demographic parameters from the SFS*

As no closed-form solution exists for the theoretical expression of the SFS for a growing population, we must use the integral formula in equation (2.15). Fortunately, there are some heuristic methods we can use to fix growth parameter values relatively easily.

We start by examining the high-frequency tail end of the SFS. There are two features of interest here: the frequency below which the SFS is no longer flat and the tail's flat shape. As we saw in section 3.1, obtaining a flat-tailed SFS requires the dynamics for a constant population to have evolved for a minimum amount of time. This places a minimum bound on the bottleneck time t_b. This parameter also controls the frequency at which the uptick occurs: the higher t_b is, the lower the frequency of the turning point. This variation is shown in Figure 4a.

After t_b has been fixed, we have three remaining growth parameters: t, N_0, t_g. We now want to match our theoretical curve's uptick to that of the SFS. In essence, the qualitative relation is that faster population growth is indicated by a steeper uptick in the SFS - this variation is shown in Figure 4b. We can apply the constraint that $N(t) = 7 \times 10^9$ to obtain two free parameters. A two-parameter search in our chosen free parameters N_0 and t_g can then be used to finally match this region to the observed site frequency spectrum.

4. RESULTS

4.1. *Predicted demography for the total population*

The method described in section 3.3 was used to find the growth parameters for a theoretical curve based on equation (2.15) that best matched the gnomAD synonymous SFS. A simple one-parameter search for t_b was performed manually, producing an optimal value of $t_b = 1.93 \times 10^5$. Assuming a generation is 25 years long, this is equivalent to $t_b = 4.83$ million years. This is a not entirely unrealistic value, as the most recent common ancestor between humans, bonobos and chimpanzees, our closest living relatives, is thought to have existed between four and seven million years ago.[29]

After t_b was fixed, a two-parameter search was performed in N_0 and t_g as detailed in Figure 5. The resulting optimal values were $N_0 = 3982$ and $t_g = 401$. Using equation (2.14) we can find $t - t_b$ as

$$t - t_b = t_g \ln\left(\frac{N}{N_0}\right)$$

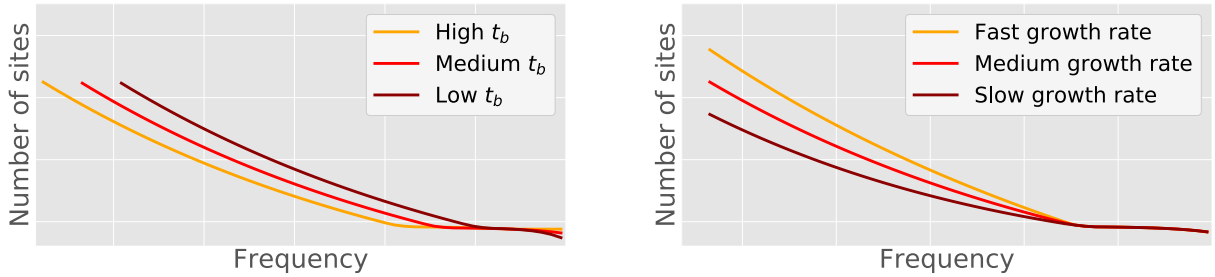
Using $N = 7 \times 10^9$, we obtain $t - t_b = 5766$, i.e. the prediction is that exponential population growth started 5,766 generations ago. Again assuming a generation is 25 years long, this corresponds to around 144 millennia. This lies within the range of currently believed times for "Out of Africa II".[4–7]

The value of N_0 predicted here seems lower than what we might expect the population size to be at the time of expansion. There are multiple factors that may be responsible for this. Firstly, as discussed in section 3.2, synonymous mutations are not completely selectively neutral. If there is any selectivity, we would expect a higher proportion of mutations to exist at lower frequencies, raising the gradient of the uptick. This would inflate the required pace of growth anticipated in a fully neutral theory, leading to a lower t_g and, given that the end population size is fixed, possibly a lower N_0 to allow for faster growth (though it is important to note that this is not a cut-and-dried relationship as the total length of the dynamics is not fixed). Furthermore, rather than "Out of Africa II" being a single mass migration, multiple dispersals[4] are thought to have occurred spanning hundreds of thousands of years. Our simple population growth model cannot capture this. Thirdly, our model assumes a fairly uniform amount of offspring from individual to individual. In reality, we know that this was certainly not the case: Genghis Khan, for example, is thought to have fathered hundreds of children and was famously estimated[30] in 2003 to be a direct ancestor of around 1 in 200 individuals today. This would allow some mutations to reach fair higher frequencies, skewing our prediction of N_0. Finally, it is key to note that N_0 is not quite the total population size at the time of expansion. Instead, it gives a measure of the population size of the *ancestors* of today's population. Clearly, not all lines of descent from the population hundreds of thousands of years ago will have continued unbroken into today; therefore we might in fact *expect* N_0 to be low. Unfortunately, it is difficult to quantify this expected discrepancy.

4.2. *Predicted demography for population of African ethnicity*

gnomAD labels data by the self-declared ethnic heritage of the individual. An interesting population to look at as a test for our theory would clearly be people of African ethnicity. These individuals' ancestors would likely not have partaken in the "Out of Africa II" migration and so this should be reflected in the site frequency spectrum of their mutations. gnomAD data primarily comes from surveys taken in Europe or North America, and so the number of exomes for people of African/African American ethnicity is only 8,128, compared to the 125,748 in total. However, this is sufficient to produce a sufficiently high quality SFS, as shown in Figure 6.

We immediately see that the shape of the SFS is very different to that of the total population. There is no flat high-frequency tail and the SFS appears to increase in gradient at frequencies below $\approx 10^{-3}$. If we try to perform the same procedure in section 4.1, we obtain $t_b = 0$ and parameters suggesting a population growing at a constant exponential rate for 3.54 million years from $N_0 = 6214$ to our



(a) Example site frequency spectra generated using equation (2.16). All have the same parameters except for t_b. If t_b is low then the tail is not flat; there is a drop-off at high frequencies. As t_b is increased, the tail becomes flatter and the frequency of the uptick becomes lower.

(b) Example site frequency spectra generated using equation (2.16). All have the same parameters except for N_0 and t_g. The faster the growth rate, the higher the gradient of the uptick. Curves are normalised so that the tails overlap.

Figure 4: Relation between shape of SFS and growth parameters.

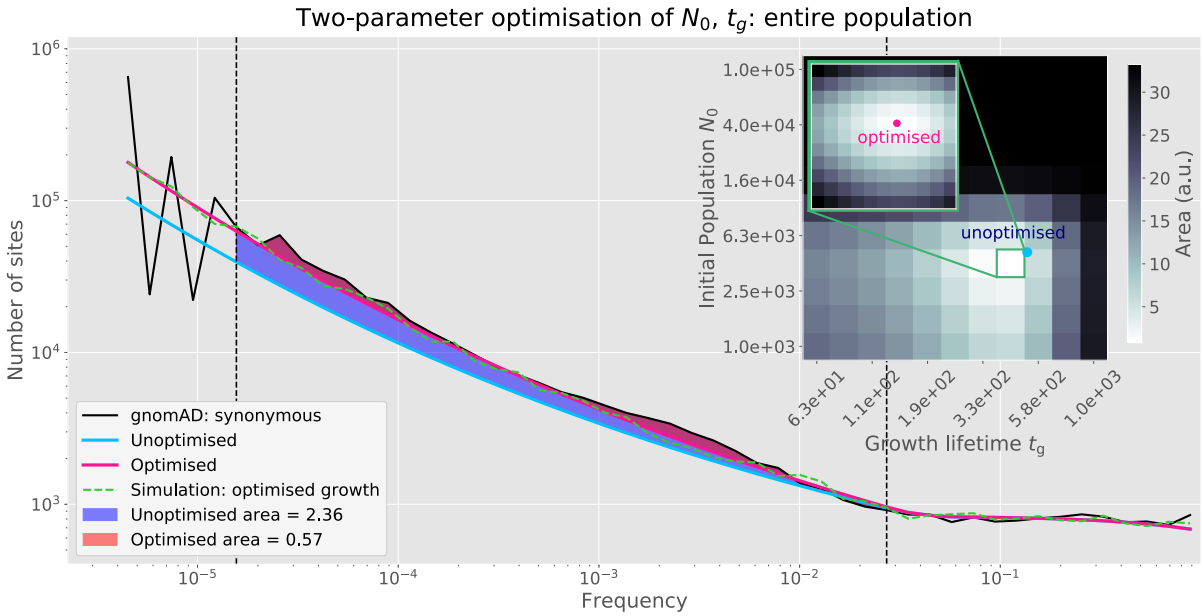


Figure 5: Two-parameter optimisation of N_0, t_g. The initial values were chosen as $N_0 = 5000$ and $t_g = 500$, producing the light blue curve seen using equation (2.16). The inset shows the "grid" of logarithmically spaced parameter values used for the search: for each grid block, the values were used to generate a curve with equation (2.16). The similarity metric was the area between the curves in the region of interest, indicated by the black, vertical, dashed lines. The blocks are shaded by their area: lighter blocks indicate a produced curve with higher similarity according to the area metric. After the first search was complete, the search was repeated at a finer-grained level in the parameter space indicated by the block highlighted in green; this search is shown in the smaller inset. The optimised parameters are indicated by the pink circle in the smaller inset and these were used to produce the pink curve with equation (2.16). This has good agreement with the gnomAD synonymous SFS and has parameters $N_0 = 3982$ and $t_g = 401$, predicting that population expansion started around 144 millennia ago. The parameter values were simulated in the process described in section 3.1, producing the result shown as the green dashed line. It has close agreement with the theoretical curve.

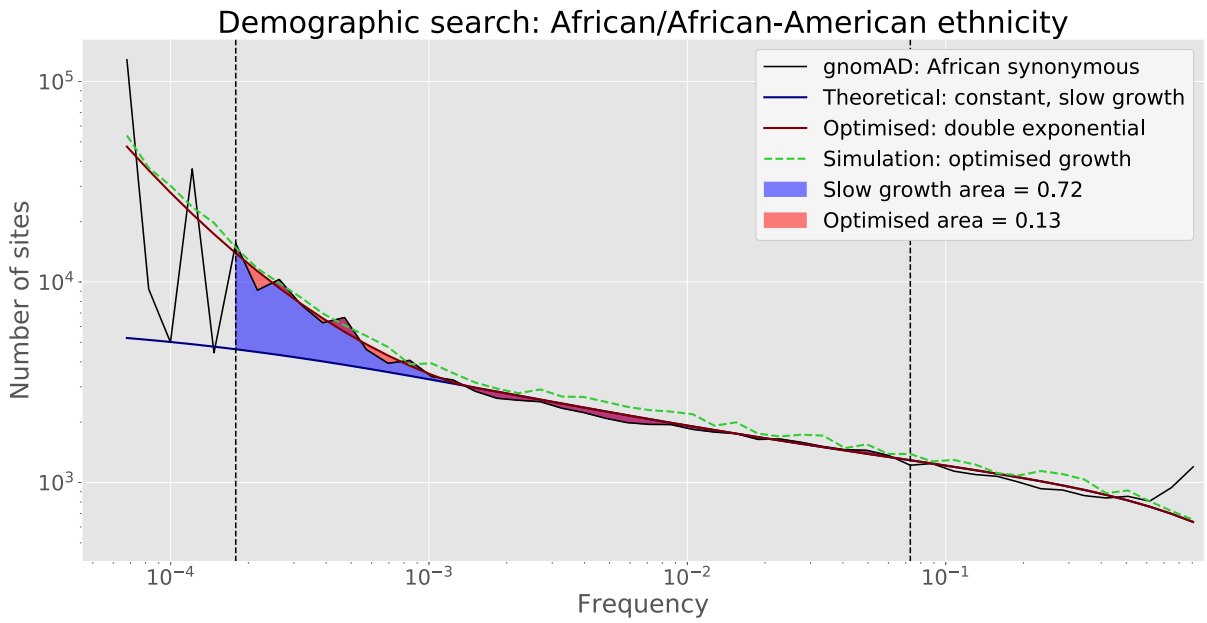


Figure 6: Fitting a demographic model to the gnomAD-derived SFS (black) for people of African/African-American ethnicity. The initial attempt to fit using the same method as in section 4.1 produced the blue curve, which has growth parameters indicating constant exponential growth from 3.54 million years ago until today. A subsequent second optimisation attempt was made by introducing a more recent bottleneck time and a subsequent higher rate of exponential growth. The relevant parameters for this were found again using the same method as in section 4.1. The resulting red curve shows good agreement with the gnomAD-derived SFS. Parameter values indicate that faster growth began 1150 years ago. These parameter values were simulated in the process described in section 3.1, producing the result shown as the green dotted line. It has close agreement with the theoretical curve.

fixed endpoint of $N = 1.5 \times 10^9$ today. As the produced blue curve in Figure 6 clearly fails to match the observed SFS, we repeated the procedure for a proposed later bottleneck in order to produce the higher gradient seen for $f \lesssim 10^{-3}$. The result indicated faster exponential growth starting only 46 generations, or 1150 years, ago.

To understand whether this value is realistic, we need to examine the history of the individuals sequenced for this SFS. The gnomAD population of African/African-American individuals is largely comprised of individuals living in Europe or the USA today. These individuals' ancestors likely left Africa around the time of the slave trade, 300-500 years ago: more recent than the predicted bottleneck. However, there are individuals in the study who are still living in Africa, whose ancestral population would have likely undergone quite different demographic changes to the other. If this population experienced slower population expansion than the African-American population, this would skew the apparent bottleneck time to be less recent.

Another distinguishing feature of this SFS is the larger uptick at very high frequency. This could indicate stronger positive natural selection, which may be plausible as the African population faces very different selection pressures to the European/North American population. For example, the sickle cell mutation that protects from malaria is highly common in individuals living in Sub-Saharan Africa, although this is a missense mutation rather than a synonymous one.[31]

An issue which is harder to explain is that the model does not predict that the population size was ever constant. It is hard to justify why this SFS does not have a flat high-frequency tail. A possible answer could be that the population that left Africa in "Out of Africa II" is different to the ancestral

population of Africans/African-Americans today. This would contravene the prediction of RSO that *all* humans are descended from a single population as recently as 300,000 years ago. This is not a convincing explanation for the discrepancy and so further investigation is merited.

4.3. *Evidence of strongly negative selection*

The site frequency spectra for synonymous and missense mutations derived from gnomAD are shown in Figure 7 (inset), normalised so that the tails overlap - this makes the shapes easier to compare. We see that the missense mutations have a higher gradient uptick at low frequencies. This is to be expected: if the functional changes due to missense mutations are deleterious, the mutations are subject to negative selection and so remain "stuck" at low f. Additionally, observe the larger uptick at very high f: this is due to mutations that increase fitness and hence reach high f through positive selection. Interestingly, the uptick gradient at low frequencies is similar for both types of mutation until about $f = 3 \times 10^{-3}$, indicating that perhaps negative selection is only playing a role at frequencies below this.

Due to the degeneracy between codons and amino acids described in section 3.2, we can expect that roughly a third of point mutations will be synonymous and roughly two thirds will be missense mutations. Hence in the gnomAD data we might expect that at low frequencies, where selection plays less of a role, the count of missense mutations will be double that of the synonymous mutations. In reality, we observe that the ratio of missense:synonymous is $\approx$1.8 rather than 2, i.e. there are fewer missense mutations than expected. This indicates the possibility that missense mutations exist that are so strongly selected against, their frequencies in the population are below the detectability threshold, i.e. frequencies below 1/(sample size):

$$\int_{f_0}^{f_{\min}} \rho(f)\,\mathrm{d}f \sim \frac{1}{\text{sample size}}$$

A simple way to investigate this is to measure the missense:synonymous ratio at low frequencies with varying sample size. If the sample size is higher, then the lowest observable frequency is lower and hence we exclude fewer deleterious mutations. Therefore we expect the ratio to increase, as we expect more missense mutations to be deleterious.

This was investigated by taking random samples of varying sizes from the total gnomAD population, forming the SFS and finding the ratio of missense to synonymous mutations at low frequency. Figure 7 shows that the ratio increases with sample size, indicating that strongly negative selection may indeed be present.

5. CONCLUSIONS

In conclusion, we have used a rich dataset of sequenced human exomes in combination with a simplistic model of neutral mutation dynamics to elucidate a model of human demography that seems realistic, given genetic and archaeological evidence. We constructed the model by considering the highly complex system governing the evolution of the frequency of a mutation and simplifying this to three key features: mutation, drift and demography. By investigating the values of the quantities that parametrise these processes, we were able to extract quantitative predictions for historic human demography. For the overall population, we predicted an expansion time that lies within the range of accepted values, and our predictions for the African/African-American population exhibit key differences that tally with common knowledge of this population's history. These predictions were also validated using a method of computer simulation that clearly represents the core components of the dynamics described in the model. Further validation could come from examining the site frequency spectra of East Asian and South Asian populations, comparing their demography to European populations.

A clear direction for further investigation is to account for selectivity. The application of our model to the gnomAD data was limited by the fact that the synonymous mutations used for the SFS were not

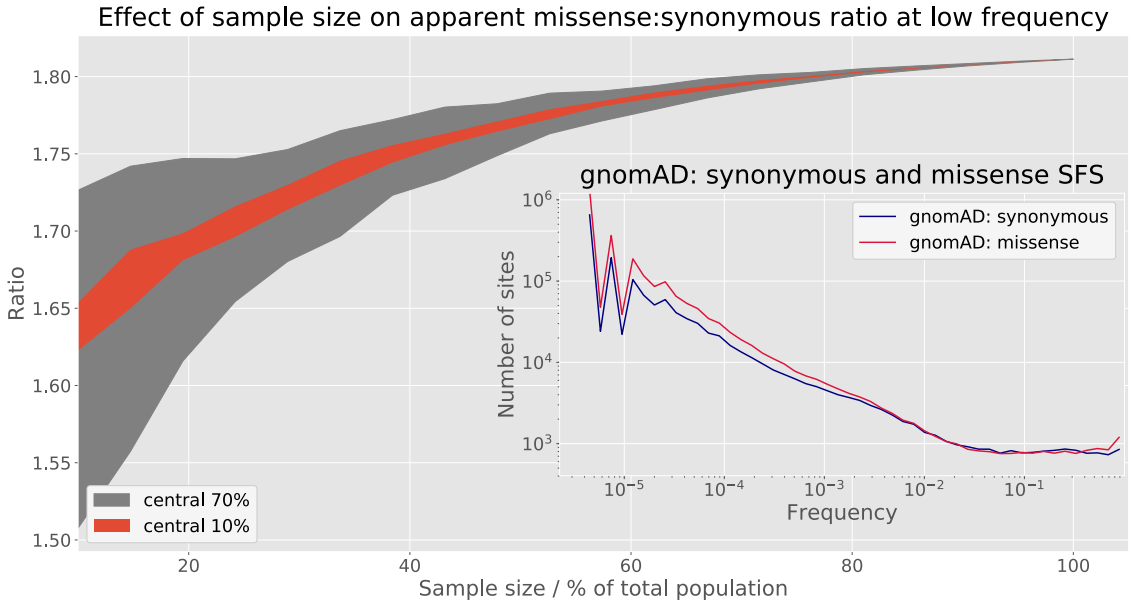


Figure 7: The ratio of the number of missense mutations to the number of synonymous mutations in the lowest frequency bin of the SFS with varying sample size taken from the total population. Samples were taken randomly and so the ratio varied for each sample size. The grey region indicates the central 70% of all ratios and the orange region indicates the central 10%. The ratio is seen to increase with increasing sample size, indicating the presence of strongly negative selection. Inset: gnomAD-derived site frequency spectra for missense (red) and synonymous (blue) mutations.

strictly neutral. Equation (2.2) would be the starting point: by incorporating a realistic distribution of selection coefficients s, a new expression for the SFS could be found. Another possible point to address is our assumption of independent sites. It is known that this is not the case: indeed, sites that are close to each other on the genome travel almost asexually. There is rich information available by considering the interaction between sites, an example of which is *genetic hitchhiking*, a mechanism affecting the frequency of a neutral site that was not included in our model. Put simply, genetic hitchhiking is when a selectively (dis)advantageous mutation appears on a site close to the neutral site in question. If the mutation is selectively advantageous it will quickly sweep towards fixation, "pulling" the neutral mutation along with it as recombination is unlikely to split the sites. The neutral mutation's frequency in the population will increase, and do so quickly because natural selection acts faster than mutation or drift. Hitchhiking can lead to large, rapid changes in a mutation's frequency and so may be responsible for the high-frequency uptick in the African/African-American SFS, for example.

References

[1] Zhu, Z. *et al.* Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago. *Nature* **559**(7715), 608–612 (2018).

[2] Stringer, C. Modern human origins: Progress and prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences* **357**(1420), 563–579 (2002).

[3] Liu, H. *et al.* A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *The American Journal of Human Genetics* **79**(2), 230–237 (2006).

[4] Bae, C.J. *et al.* On the origin of modern humans: Asian perspectives. *Science* **358**(6368) (2017).

[5] Armitage, S.J. *et al.* The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science* **331**(6016), 453 LP – 456 (2011).

[6] Balter, M. Was North Africa the Launch Pad for Modern Human Migrations? *Science* **331**(6013), 20 LP – 23 (2011).

[7] Quintana-Murci, L. *et al.* Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics* **23**(4), 437–441 (1999).

[8] Thorne, A.G. *et al.* The Multiregional Evolution of Humans. *Scientific American* **266**(4), 76–83 (1992).

[9] Wolpoff, M.H. *et al.* Multiregional Evolution: A World-Wide Source for Modern Human Populations BT - Origins of Anatomically Modern Humans. pp. 175–199. Springer US, Boston, MA (1994).

[10] Poznik, G.D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**(6145), 562–565 (2013).

[11] Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* **23**(7), 553–559 (2013).

[12] Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research* **25**(4), 459–466 (2015).

[13] Cruciani, F. *et al.* A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa. *The American Journal of Human Genetics* **88**(6), 814–818 (2011).

[14] Gitschier, J. All About Mitochondrial Eve: An Interview with Rebecca Cann. *PLoS Genetics* **6**(5), e1000,959 (2010).

[15] Frayer, D.W. *et al.* Theories of Modern Human Origins: The Paleontological Test. *American Anthropologist* **95**(1), 14–50 (1993).

[16] Stringer, C. *et al.* Methods, Misreading, and Bias. *American Anthropologist* **96**(2), 416–424 (1994).

[17] Stringer, C. Why we are not all multiregionalists now. *Trends in Ecology & Evolution* **29**(5), 248–251 (2014).

[18] Stringer, C. Modern human origins-distinguishing the models. *African Archaeological Review* **18**(2), 67–75 (2001).

[19] Stringer, C. Human evolution: Out of Ethiopia. *Nature* **423**(6941), 693–695 (2003).

[20] Templeton, A.R. Out of Africa again and again. *Nature* **416**(6876), 45–51 (2002).

[21] Wolpoff, M.H. Multiregional evolution (2018).

[22] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**(7616), 285–291 (2016).

[23] Karczewski, K.J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* p. 531210 (2019).

[24] Levy, S.F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181 (2015).

[25] Tran, T.D. *et al.* An introduction to the mathematical structure of the Wright-Fisher model of population genetics. *Theory in Biosciences* **132**(2), 73–82 (2013).

[26] Nachman, M.W. *et al.* Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1), 297–304 (2000).

[27] Chang, J.C. *et al.* Beta 0 thalassemia, a nonsense mutation in man. *Proceedings of the National Academy of Sciences* **76**(6), 2886 LP – 2889 (1979).

[28] Tsui, L.C. The spectrum of cystic fibrosis mutations. *Trends in Genetics* **8**(11), 392–398 (1992).

[29] Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527 (2012).

[30] Zerjal, T. *et al.* The Genetic Legacy of the Mongols. *The American Journal of Human Genetics* **72**(3), 717–721 (2003).

[31] Rees, D.C. *et al.* Sickle-cell disease. *The Lancet* **376**(9757), 2018–2031 (2010).

[32] Butler, R.W. *Saddlepoint Approximations with Applications.* Cambridge University Press, Cambridge (2007).

APPENDICES

Appendix A

Derivation of integral expression (2.15) for SFS

We start with equation (2.9):

$$\rho(n) = \int_0^t \frac{1}{(t-T)^2} \exp\left[-\frac{n}{t-T}\right] \cdot (N\mu)\, \mathrm{d}T \tag{2.9, repeated}$$

Inserting our population growth model (2.14) we obtain

$$\begin{aligned}\rho(n) \;=\; & \underbrace{\mu N_0 \int_{T=0}^{t_b} \frac{1}{(t-T)^2} \exp\left[-\frac{n}{t-T}\right] \mathrm{d}T}_{①} \\ & + \underbrace{\mu N_0 \int_{T=t_b}^{t} \frac{1}{(t-T)^2} \exp\left[-\frac{n}{t-T}\right] \exp\left[\frac{T-t_b}{t_g}\right] \mathrm{d}T}_{②}\end{aligned} \tag{5.1}$$

First consider integral ①. We change variables to $x = t - T$:

$$\begin{aligned}① &= \mu N_0 \int_{x=t-t_b}^{t} \frac{1}{x^2} \exp\left[-\frac{n}{x}\right] \mathrm{d}T \\ &= \frac{\mu N_0}{n}\left[e^{-\frac{n}{t}} - e^{-\frac{n}{t-t_b}}\right]\end{aligned} \tag{5.2}$$

Now consider integral ②. We again change variables to $x = t - T$:

$$\begin{aligned}② &= \mu N_0 \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{n}{x}} e^{\frac{(t-t_b)-x}{t_g}} \mathrm{d}x \\ &= \mu \underbrace{N_0 e^{\frac{t-t_b}{t_g}}}_{N} \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{n}{x}} e^{-\frac{x}{t_g}} \mathrm{d}x \\ &= \mu N \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{n}{x}} e^{-\frac{x}{t_g}} \mathrm{d}x\end{aligned} \tag{5.3}$$

Therefore, substituting equations (5.2) and (5.3) into equation (5.1) we obtain

$$\rho(n) = \frac{\mu N_0}{n}\left[e^{-\frac{n}{t}} - e^{-\frac{n}{t-t_b}}\right] + \mu N_0 e^{\frac{t-t_b}{t_g}} \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{n}{x}} e^{-\frac{x}{t_g}} \mathrm{d}x$$

We now restrict ourselves to times $t > t_b$, which is the regime of validity for any observed SFS in today's human population. Therefore we have $n = Nf$ with $N = N_0 \exp\left(\frac{t-t_b}{t_g}\right)$ and $\rho(f) = N\rho(n)$ and so

$$\begin{aligned}\rho(f) &= N\frac{\mu N_0}{Nf}\left[e^{-\frac{Nf}{t}} - e^{-\frac{Nf}{t-t_b}}\right] + \mu N^2 \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{fN}{x}} e^{-\frac{x}{t_g}} \mathrm{d}x \\ &= \frac{\mu N_0}{f}\left[e^{-\frac{n}{t}} - e^{-\frac{n}{t-t_b}}\right] + \mu N^2 \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{fN}{x}} e^{-\frac{x}{t_g}} \mathrm{d}x\end{aligned} \tag{2.15, repeated}$$

Finally, we find $\rho(\ln f) = f \cdot \rho(f)$:

$$\rho(\ln f) = \mu N_0\left[e^{-\frac{n}{t}} - e^{-\frac{n}{t-t_b}}\right] + f\mu N^2 \int_{x=0}^{t-t_b} \frac{1}{x^2} e^{-\frac{fN}{x}} e^{-\frac{x}{t_g}} \mathrm{d}x \tag{2.16, repeated}$$

Appendix B

Attempt at approximation to integral formula (2.15)

Here we follow Butler (2007).[32] The *method of steepest descent*, or *saddlepoint approximation*, is a method for approximating integrals of the form

$$I = \int g(x) e^{cx} \, \mathrm{d}x$$

This can be written as

$$I = \int \exp\left(-h(c,x)\right) \, \mathrm{d}x$$

where $h(c,x) = -cx - \ln\left(g(x)\right)$. We now take c constant and expand $h(c,x)$ in x about $x = x_0$:

$$h(c,x) = h\left(c, x_0\right) + \left.\frac{\partial h}{\partial x}\right|_{x=x_0} (x - x_0) + \frac{1}{2} \left.\frac{\partial^2 h}{\partial x^2}\right|_{x=x_0} (x - x_0)^2 + \cdots$$

We can calculate the derivatives as

$$\begin{aligned} \left.\frac{\partial h}{\partial x}\right|_{x=x_0} &= -c - \left.\frac{\partial \left(\ln g(x)\right)}{\partial x}\right|_{x=x_0} \\ \left.\frac{\partial^2 h}{\partial x^2}\right|_{x=x_0} &= - \left.\frac{\partial^2 \left(\ln g(x)\right)}{\partial x^2}\right|_{x=x_0} > 0 \end{aligned}$$

where we assume the last equality, as it is required for this approximation.

Now, as the name of the approximation implies, we focus on x_t, the solution to $h'\left(c, x_t\right) = 0$. We assume that this gives a minimum for $h(c,x)$, and therefore maximises I. The crux of this approximation is now apparent: we assume that the integral is dominated by contributions close to x_t and therefore expand around this point. Expanding to second order, we obtain

$$\begin{aligned} I &\approx \int \exp\left[-h\left(c, x_t\right) + \frac{1}{2} \left.\frac{\partial^2 \left(\ln g(x)\right)}{\partial x^2}\right|_{x=x_0} (x - x_t)^2\right] \mathrm{d}x \\ &= e^{-h(c,x_t)} \int \exp\left[\frac{1}{2} \left.\frac{\partial^2 \left(\ln g(x)\right)}{\partial x^2}\right|_{x=x_0} (x - x_t)^2\right] \mathrm{d}x \end{aligned}$$

We recognise this as a Gaussian integral with solution

$$I \approx e^{-h(c,x_t)} \sqrt{\frac{2\pi}{\left.\frac{\partial^2 \left(\ln g(x)\right)}{\partial x^2}\right|_{x=x_0}}} \tag{5.4}$$

which is a form of the saddlepoint approximation. In our case, from equation (5.3), the integral we wish to approximate is

$$I = \int_0^{t-t_b} \underbrace{\frac{1}{x^2} e^{-\frac{Nf}{x}}}_{g(x)} \underbrace{e^{-\frac{x}{t_g}}}_{c=-\frac{1}{t_g}} \mathrm{d}x$$

The issue here is that the stationary point of $h(c,x) = \frac{x}{t_g} + 2\ln x + \frac{Nf}{x}$ lies outside the integration range for parameter values used here. x_t is defined by

$$\left.\frac{\mathrm{d}h}{\mathrm{d}x}\right|_{x=x_t} = \frac{2x_t - Nf}{x_t^2} + \frac{1}{t_g} = 0$$

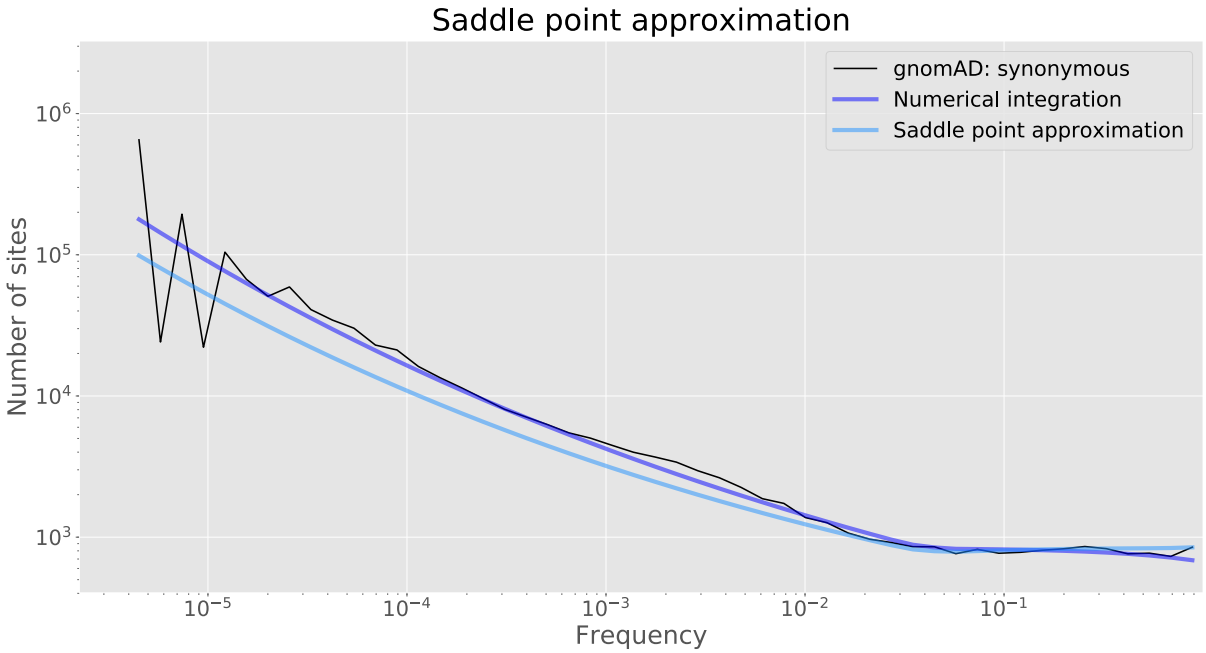


Figure 8: Saddle point approximation. The curve generated using numerical integration of equation (2.15) is shown in dark blue. The saddlepoint approximation to this is shown in light blue. Note that in keeping with the rest of the report, we normalise so that the curves overlap at high frequency. The saddlepoint approximation fails to capture key features of the gnomAD SFS (black), such as the gradient of the uptick at low frequencies and the shape of the high-frequency tail.

which has positive solution

$$x_t = t_g \left(\sqrt{\frac{Nf}{t_g} + 1} - 1 \right)$$

For parameter values $N = 7 \times 10^9$, $t_g = 401$ and $t - t_b = 5766$ that are determined in our analysis in section 4.1, we see that x_t is only inside the integration range for $f \lesssim 10^{-5}$. Therefore outside this range, the accuracy of the approximation decreases as f increases. The result is shown in Figure 8. The main failure, however, is that this approximation fails to give an expression for the integral that makes clear the influences of the parameter values on the SFS' shape. Therefore we are forced to use heuristic arguments and optimisation methods as described in section 3.3. There is no benefit to using this approximation over direct numerical integration of the expressions (2.15) and (2.16).